

UNIVERSIDADE FEDERAL DE PELOTAS

Programa de Pós-Graduação em Biotecnologia



Tese

**Caracterização genômica de genes de síntese de amido e elementos repetidos  
e suas aplicações no melhoramento de plantas**

**Karine Elise Janner de Freitas**

Pelotas, 2021

**Karine Elise Janner de Freitas**

**Caracterização genômica de genes de síntese de amido e elementos repetidos  
e suas aplicações no melhoramento de plantas**

Tese apresentada ao Programa de Pós-Graduação em Biotecnologia da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutora em Ciências (área do Conhecimento: melhoramento vegetal).

**Orientador:** Antonio Costa de Oliveira

**Coorientadores:** Filipe de Carvalho Victoria  
Carlos Busanelo  
Camila Pegoraro

Pelotas, 2021

Universidade Federal de Pelotas / Sistema de Bibliotecas  
Catalogação na Publicação

F863c Freitas, Karine Elise Janner de

Caracterização genômica de genes de síntese de amido e elementos repetidos e suas aplicações no melhoramento de plantas / Karine Elise Janner de Freitas ; Antonio Costa de Oliveira, Filipe de Carvalho Victoria, orientadores ; Carlos Busanello, Camila Pegoraro, coorientadores. — Pelotas, 2021.

123 f. : il.

Tese (Doutorado) — Programa de Pós-Graduação em Biotecnologia, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2021.

1. SSRs. 2. mtDNA. 3. cpDNA. 4. SSRGs. 5. CREs. 6. Arroz. I. Oliveira, Antonio Costa de, orient. II. Victoria, Filipe de Carvalho, orient. III. Busanello, Carlos, coorient. IV. Pegoraro, Camila, coorient. V. Título.

CDD : 584.93015

## **BANCA EXAMINADORA**

Prof. Dr. Frederico Schmitt Kremer (UFPel, CDTec)

Prof. Dra. Vivian Ebeling Viana (UFPEL, DF)

Prof. Dr. Luis Willian Pacheco Arge (UFRJ, CCS)

Prof. Dr. Antônio Costa de Oliveira (Orientador, UFPel, CGF)



Dedico ao meu filho, Mathias André.  
Para que esses pequenos grandes olhos possam enxergar longe!

## **Agradecimentos**

À Deus, que antes de tudo, é a maior força que acompanha, me guia e me protege todos os dias da minha vida! Obrigada por me permitir tanto!

Ao meu esposo Márcio, meu exemplo de humildade e amor. Obrigada pela paciência que tiveste mesmo nos momentos que deveriam ser teus. Tenho certeza de que eu não poderia ter escolhido um pai mais amoroso e dedicado para o nosso filho Mathias, nosso pequeno príncipe, presente de Deus durante o doutorado, que nos surpreende todos os dias. Vocês são tudo para mim! Meu alicerce, minha felicidade, meu porto-seguro. Eu amo vocês!

A minha mãe Siegrid e meu pai Nilton, que mesmo diante das mais diversas dificuldades da vida nunca mediram esforços para que eu e meus irmãos estudássemos e tivéssemos uma boa formação. Aos meus queridos irmãos, Márcio e Eduardo, que são meus exemplos de garra e vitória. À essa família amada, meu muito obrigada por tudo que fizeram por mim e o que significam na minha vida. Eu amo vocês!

Ao professor Antonio de Oliveira, meu orientador. Obrigado por todo conhecimento ensinado, pela paciência e compreensão em todos os momentos. O senhor é um exemplo para todos os seus alunos, primeiramente por ser um grande mestre, mas principalmente pelo grande ser humano que és!

Aos meus coorientadores, professores Carlos Busanelo e Camila Pegoraro. Muito obrigada por todo conhecimento ensinado, pela paciência, pela amizade e todas as valiosas contribuições que fizeram para nossos artigos. Ao professor Filipe Victoria, o grande incentivador para que eu seguisse os passos na pós-graduação, obrigada por ensinar tanto e ser esse exemplo de conhecimento.

A todos os professores do CGF, pelos conhecimentos ensinados que contribuem para a excelência desse grupo.

Ao amigo e professor Railson Schreinert dos Santos, que além de deter grande conhecimento, é um ser humano iluminado. Muito obrigada pelas contribuições em cada capítulo dessa tese, toda tua ajuda e conselhos, Railson! Também a pesquisadora Vivian Ebeling Viana pela amizade e grandes contribuições nos capítulos dessa tese.

A todos os colegas e amigos do CGF. Em especial a amiga Jennifer Lopes e sua família querida. Obrigada por essa sincera amizade que nós construímos e por

todos os mates e risadas compartilhados. As amigas(os) Ana Frank, Cintia Garcia, Stefania Garcia, Victoria Oliveira, Tatieli Silveira, Rebeca Cataneo, Luis Chairez e, todos os demais que fizeram parte dessa caminhada, meu muito obrigada por todos os momentos de alegria, companheirismo e descontração. Estou certa de que nossa amizade continuará muito além do período da pós-graduação.

À CAPES pelo apoio financeiro fornecido.

Ao programa de pós-graduação em Biotecnologia. Em especial a Daiane Nunes, pela sua eficiência e competência. Ao CLC e ao Programa de Idiomas Sem Fronteiras que permitiu a realização do curso de Alemão, o qual estimei tanto fazer. A querida professora Larissa, por todo carinho e ensinamentos repassados.

À Universidade Federal de Pelotas, que foi minha casa durante esses 4 anos de muito estudo.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001”

Muito obrigada!

“A verdadeira viagem de descobrimento não consiste em procurar novas paisagens,  
mas em ter novos olhos”.  
(Marcel Proust)

## Resumo

De Freitas, Karine Elise Janner. **Caracterização genômica de genes de síntese de amido e elementos repetidos e suas aplicações no melhoramento genético vegetal.** 2021. 123f. Tese (Doutorado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

A disponibilidade de sequências de genomas vegetais nos bancos de dados públicos tem permitido a elaboração de estudos que trazem importantes conhecimentos sobre a evolução das espécies e que auxiliam os programas de melhoramento genético visando a tolerância a estresses ambientais que, não somente diminuem a produção, mas também afetam a qualidade de grão. Ainda assim há muito a se entender, principalmente no que tange à evolução de sequências repetitivas em genomas acessórios e em genes relacionados com qualidade do grão em arroz sob estresses abióticos. Aqui 4 estudos são apresentados, trazendo novas informações nas referidas áreas. O primeiro deles trata da abundância, distribuição e evolução dos microsatélites (SSRs) em genomas mitocondriais de plantas e algas. Os resultados oferecem importantes *insights* sobre a evolução dos SSRs nos grupos de plantas. Em um dos outros estudos são oferecidos 26 iniciadores para amplificação de SSRs nos genomas plastidiais de Aveia que possibilitam a distinção de espécies e populações do gênero *Avena*. Um estudo sobre a qualidade de grão do arroz foi elaborado também, caracterizando 19 importantes genes que atuam na síntese e modificação do amido em 11 espécies de *Oryza*. Nesse estudo, sugere-se que as deleções/mutações de aminoácidos em sítios ativos resultam em variações que podem afetar negativamente etapas da biossíntese do amido no endosperma, o que pode ser bastante útil em programas de melhoramento de genótipos de arroz visando qualidade de grão a partir de espécies selvagens. E por último, foi proposta uma estratégia, através da identificação de um SNP, para ser testada e utilizada em programas de melhoramento para tolerância aos principais estresses que afetam a qualidade do grão em arroz.

**Palavras-chave:** SSRs, mtDNA, cpDNA, SSRGs, CREs, arroz.

## **Abstract**

De Freitas, Karine Elise Janner. **Genomic characterization of starch synthesis genes and repeated elements and their applications in plant genetic improvement.** 2021. 123f. Tese (Doutorado) - Programa de Pós-Graduação em Biotecnologia. Universidade Federal de Pelotas, Pelotas.

The availability of plant genome sequences in public databases has allowed the development of studies that bring important understanding regarding the evolution of species. This has helped genetic improvement programs aiming at tolerance to environmental stresses that not only reduce yield, but also affect grain quality. Still, there is a lot to understand, especially regarding the evolution of repetitive sequences in accessory genomes and in genes related to grain quality in plants under abiotic stresses. Here, five studies are presented, bringing new information in those areas. The first one deals with the abundance, distribution, and evolution of microsatellites (SSRs) in plant and algae mitochondrial genomes. The results provide important insights into the evolution of SSRs in plant groups. In the other study, 26 primers are offered for the amplification of SSRs in the plastid genomes of Oat that allow the distinction of species and populations of the *Avena* genus. Study for rice grain quality was conducted, characterizing 19 important genes that act in the synthesis and modification of starch in 11 species of *Oryza*. In this study, it is suggested that amino acid deletions/mutations in active sites in wild species result in variations that can negatively affect stages of starch biosynthesis in the endosperm, which can be very useful in rice breeding programs aiming at grain quality a from wild species. Finally, a strategy was proposed, through the identification of a SNP, to be tested and used in breeding programs for tolerance to the main stresses that affect rice grain quality.

**Keywords:** SSRs, mtDNA, cpDNA, SSRGs, CREs, rice.

## **Lista de Figuras**

- Figura 1.** Diagrama esquemático dos eventos de domesticação no arroz cultivado. 17
- Figura 2.** Diagrama representativo de um típico promotor eucariótico 23



## Lista de Abreviaturas

ABA: Absciscic Acid  
AC: Amylose Content  
AGPL: ADP-glucose Pyrophosphorylase Large Subunit  
AGPS: ADP-glucose Pyrophosphorylase Small Subunit  
ALK: Starch Synthase III  
CpDNA: Chloroplast genome  
CREs: *Cis* Regulatory Elements  
DBE: Debranching Enzyme  
DPE1: Disproportionating Enzyme 1  
ET: Ethylene  
GT: Gelatinization Temperature  
GA: Giberellin  
GBSSII: Granule-Bound Starch Synthase II  
IAA: Indole-3-Acetic Acid  
ISA: Isoamylase  
JA: Jasmonate  
MEME: Multiple Em for Motif Elicitation  
MtDNA: Mitochondrial genome  
PUL: Pullulanase  
RAP-DB: The Rice Annotation Project Database  
SBE: Starch Branching Enzyme  
SA: Salicylic Acid  
SEBF: Silencing Element Binding Factor  
SNP: Single Nucleotide Polymorphism  
SS: Starch Synthase  
SSRs: Simple Sequence Repeats  
SSRGs: Starch Synthesis-Related Genes  
TFs: Transcription Factors  
TFBSs: Transcription Factor Binding Sites  
WAXY: Granule-Bound Starch Synthase 1

## SUMÁRIO

1 INTRODUÇÃO GERAL	14
2 REVISÃO BIBLIOGRÁFICA	16
2.1 A cultura do arroz e a qualidade do grão	16
2.2 Propriedades de qualidade em <i>Oryza sativa</i> L. e as enzimas envolvidas na síntese de amido em arroz	18
2.3 Estresses abióticos/bióticos afetam a qualidade do grão de arroz	20
2.4 O que se sabe sobre as características de qualidade nas espécies selvagens?	21
2.5 Elementos de regulação cis (CREs) podem fornecer novas possibilidades para a engenharia genética	22
2.8 Microsatélites ou SSRs	24
2.9 <i>Avena sativa</i> L.	27
3 HIPÓTESE E OBJETIVOS	28
3.1 Hipótese	28
3.2 Objetivo Geral	28
3.3 Objetivos Específicos	29
4 CAPÍTULOS	30
4.1 Artigo 1. An empirical analysis of mtSSRs: Could microsatellites distribution patterns explain the evolution of mitogenomes in plants?	30
4.2 Artigo 2. <i>Starch Synthesis-Related Genes</i> (SSRG) Evolution in the Genus <i>Oryza</i>	55
4.3 Artigo 3. <i>In silico</i> analysis of <i>Oryza</i> genomes: Occurrence of cis-regulatory elements of starch-related genes provide new breeding possibilities for increased grain quality in rice under stress	82
4.4 Artigo 4. Mapping and analysis of plastid Simple Sequence Repeats in genus <i>Avena</i>	111

5 CONCLUSÃO GERAL	118
6 REFERÊNCIAS	119

## 1 INTRODUÇÃO GERAL

As condições ambientais adversas, como seca, salinidade e temperaturas extremas, encontradas pela planta durante o seu ciclo de vida, impõem limitações severas ao crescimento, reprodução vegetal, produção e qualidade de grão (Gull et al., 2019).

Nesse sentido, uma característica importante a ser estudada, e que vem sendo foco dos estudos nos centros de pesquisa em melhoramento de arroz irrigado é a qualidade de grãos, sendo o amido um dos principais componentes do grão que fornece proteínas, vitaminas e minerais, além de possuir baixo teor de lipídios para o consumo. Além disso, as características como teor de amilose (AC) e temperatura de gelatinização (GT), que tem grandes efeitos na qualidade do cozimento (QC) e no consumo, são controlados pelas propriedades físico-químicas do amido no endosperma do grão (Walter et al., 2008). Por ser amplamente consumido em todo o mundo, o arroz tem diferentes formas de preparação, tornando o conceito de qualidade algo diferente em cada país. Considerando o aumento da demanda por este grão para os mercados de exportação internacionais, os aspectos que envolvem a qualidade estão sendo cada vez mais visados para estudo.

Em relação a base genética da qualidade do grão, a ação coordenada de enzimas tem relação direta com a síntese de amido, sendo os grupos AGPase (ADP - glicose pirofosforilase), SS (amido sintase), SBE (enzima de ramificação do amido) e DBE (enzima de desramificação do amido), os que têm maior impacto neste processo. Sabendo disso, torna-se importante o conhecimento estrutural e funcional dos genes envolvidos nas rotas de síntese e degradação do amido em arroz, o que facilitaria a modificação de tais processos.

Além disso, a investigação relacionando o comportamento dos genes de amido (SSRGs) frente a diferentes estresses que podem acometer a planta, bem como os elementos presentes no promotor desses genes é importante no sentido de trazer informações relevantes sobre aspectos que podem ser utilizados em programas de melhoramento genético. Assim, os elementos de ação cis (CRES) presentes nos promotores de genes são a fechadura para ativar ou reprimir genes em resposta a sinalização devido a fatores externos desafiadores, como estresses bióticos e abióticos, o que também altera as propriedades físico-químicas do amido no endosperma do

grão de arroz que podem refletir na aceitação pelo consumidor (Wittkopp & Kalay, 2012). Nesse sentido, é necessário destacar os motivos de DNA provavelmente relacionados à modificação da expressão de genes envolvidos nas vias relacionadas ao amido a fim de auxiliar no entendimento da regulação dos SSRGs em plantas sob estresse. Além disso, é importante detalhar e discutir a cascata de sinalização que envolve esses genes identificando assim, possíveis usos para o melhoramento relacionado a qualidade.

Outros motivos, neste caso repetidos, são os microssatélites ou SSRs (Do inglês, *Simple Sequence Repeats*). Essas sequências de 1 a 6 nucleotídeos repetidos, estão presentes em genomas nucleares, plastidiais e mitocondriais de plantas e animais. Discute-se muito a função destes genes os quais podem ser utilizados como uma ferramenta com diversas aplicações na biologia e biotecnologia (Devey et al., 2009). Nos genomas acessórios, estudos comparativos conduzidos a fim de caracterizar a distribuição e possíveis funções dos SSRs são realizados (George et al., 2015; Rajendrakumar et al., 2005) e contribuem para o entendimento da evolução desses elementos e principalmente demonstram a diversidade de motivos que temos nos genomas. No entanto, o impacto da presença e variação dessas sequências ao longo da evolução, principalmente nos pequenos genomas acessórios, como o mitocondrial (mtDNA), merece atenção. O genoma mitocondrial é conhecido por codificar proteínas associadas ao metabolismo energético e, pouco se sabe sobre as mutações devido ao deslizamento da polimerase como mecanismo causal de polimorfismo.

Por outro lado, o alto nível de transferibilidade de marcadores SSRs e a grande quantidade e dispersão nos genomas em genótipos de gramíneas são os principais atributos, que tornam os SSRs uma potente ferramenta a ser utilizada em programas de melhoramento (Oliveira et al., 2006). Na aveia (*Avena sativa* L.), gramínea de inverno, os SSRs são importantes na distinção de diferentes haplótipos através do uso de iniciadores SSRs, não somente nucleares, mas também plastidiais. Já que, há uma necessidade crescente de estudar diversas variedades de aveia em termos de suas características de importância agrônômica e de buscar novos genótipos que possam servir de base para o desenvolvimento de novas variedades com alta produtividade e resistência a doenças (Gagkaeva et al., 2018). Por isso, faz-se necessário a caracterização dos SSRs nos genomas plastidiais e o desenvolvimento de iniciadores para amplificação das regiões que os contém.

## 2 REVISÃO BIBLIOGRÁFICA

### 2.1 A cultura do arroz e a qualidade do grão

O arroz (*Oryza sativa* L.) é um dos alimentos mais importantes para a nutrição humana, sendo o principal alimento para mais da metade da população mundial, que o utiliza como fonte de energia diariamente. Além disso, desempenha um importante papel tanto no âmbito social e econômico quanto cultural, caracterizado como o segundo cereal mais cultivado e produzido no mundo, após o milho (Vanlalsanga e Yengkhom, 2019). O arroz é uma excelente fonte de energia, devido à alta concentração de amido, fornecendo também proteínas, vitaminas e minerais, e possui baixo teor de lipídios. Nos países em desenvolvimento, o arroz é um dos principais alimentos da dieta, sendo responsável por fornecer, em média, 715 kcal per capita por dia, 27% dos carboidratos, 20% das proteínas e 3% dos lipídios da alimentação. No Brasil, o consumo per capita é de 108 g por dia, fornecendo 14% dos carboidratos, 10% das proteínas e 0,8% dos lipídios da dieta (Walter et al., 2008). Portanto, devido à importância do arroz na dieta de grande parte da população, sua qualidade nutricional afeta diretamente a saúde humana.

Por outro lado, o gênero *Oryza* é composto por 23 espécies (Stein et al., 2018), e destas, somente duas são cultivadas, a *O. sativa*, também chamada de arroz asiático por ter sido domesticado há 10.000 anos na Ásia e, a *Oryza glaberrima* Steud., também chamado de arroz africano, além de 21 parentes silvestres. Essas espécies apresentam 11 tipos diferentes de genoma (AA, BB, CC, BBCC, CCDD, EE, FF, GG, KKLL, HHJJ, HHKK) e têm uma distribuição pan-tropical, crescendo em uma ampla gama de ambientes (Atwell et al., 2014). Uma vez que *O. sativa* foi domesticada a partir de um número limitado de genótipos de *O. rufipogon* (seu parente selvagem mais próximo) estima-se que apenas 10-20% da diversidade genética encontrada em espécies selvagens está presente no germoplasma do arroz tradicionalmente cultivado, sendo o *O. sativa* dividido em somente duas subespécies, indica e japônica, as quais apresentam propriedades físico-químicas do grão bastante diferentes, influenciando diretamente características de cozimento (Feng et al., 2017).

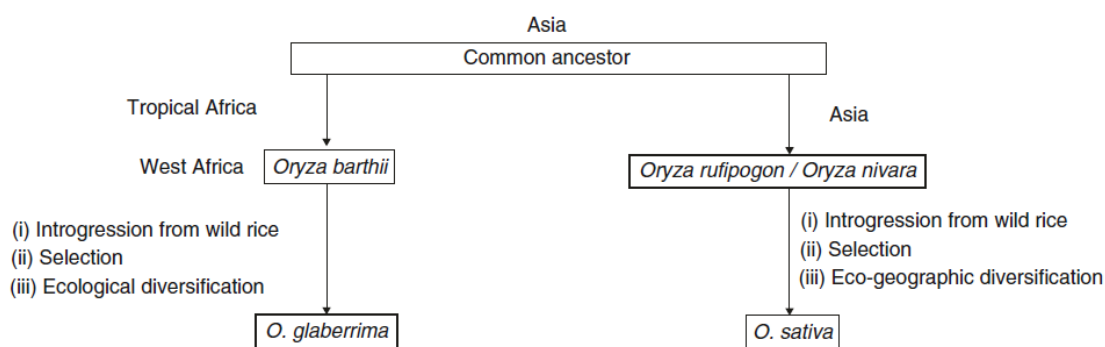


Figura 1. Diagrama esquemático dos eventos de domesticação no arroz cultivado. (Hasan & Henry, 2020).

Especificamente, a qualidade do arroz é uma característica subjetiva e sujeita aos padrões estabelecidos nos diferentes países, que por sua vez, são afetados pelos padrões culturais e pela sua forma de utilização na alimentação. A preferência do consumidor por esse cereal, geralmente, está associada a aspectos econômicos, tradicionais, variando de país para país e até mesmo de região para região dentro de um mesmo país (PEREIRA et al., 2007). De forma geral, pode-se dizer que o termo qualidade é aplicado largamente, para categorizar o comportamento do amido contido no endosperma do grão e as suas dimensões (Castro et al., 1999).

As espécies selvagens de *Oryza* e outras espécies de gramíneas geralmente demonstram características agrônômicas pobres, entre essas características, temos o menor rendimento e qualidade inferior de grãos. Curiosamente, essas espécies contêm muitos genes benéficos de interesse para características relacionadas ao rendimento e à qualidade. No entanto, um grande número desses genes ainda permanece inexplorado devido à dificuldade de transferência em arroz cultivado, e devido a poucos estudos relacionados a essas características de qualidade para genes nas espécies selvagens (Hasan & Henry, 2020). Sendo, portanto, as espécies selvagens, uma excelente fonte para exploração de novos alelos.

Em vista disso, os programas de pesquisa e melhoramento genético de arroz irrigado conduzidos no Brasil e no mundo tem adotado estratégias metodológicas que agrupem biotecnologia e melhoramento convencional, visando desenvolver, além de genótipos comerciais de alta produtividade e estabilidade (adaptadas aos sistemas de cultivo; resistentes estresses bióticos; tolerantes a estresses abióticos), também genótipos com qualidade de grãos que atendam a preferência, inicialmente do

mercado interno e posteriormente, do mercado externo (EMBRAPA CPACT, 2012). Lembrando que, um arroz com alta qualidade, traz maiores retornos aos produtores devido à alta demanda de consumo.

A qualidade de grãos é uma característica complexa, definida por diferentes variáveis e controlada por vários genes em interação com o ambiente. O melhoramento convencional para qualidade de grãos geralmente enfrenta grandes obstáculos. Esta dificuldade advém do estágio de desenvolvimento da planta em que é possível selecionar para qualidade de grãos. Ao contrário de outras características da planta, que podem ser selecionadas antes do florescimento, favorecendo o ganho de seleção por ciclo, a seleção para qualidade de grãos só pode ser realizada após a maturação, colheita e avaliação dos grãos. Um dos obstáculos refere-se à avaliação de qualidade, pois esta necessita de quantidade razoável de grãos. Sendo que nas fases iniciais de um programa de melhoramento a quantidade de grãos obtida por linhagem é limitada. Em geral, avaliações de qualidade de grãos só ocorrem na fase final do programa, quando maiores quantidades são disponíveis.

Portanto, utilizar métodos que facilitem e/ou aumentem a eficiência de seleção para qualidade de grãos, contornando estes obstáculos, é de extrema importância para os programas de melhoramento genético. Por isso, torna-se fundamental e necessário ter um maior conhecimento da composição química, fisiologia, bioquímica e do controle genético do carácter para alcançar este objetivo.

## **2.2 Propriedades de qualidade em *Oryza sativa* L. e as enzimas envolvidas na síntese de amido em arroz**

O amido representa a maior parte do carboidrato estocado no grão de arroz, uma vez que armazenam este nutriente para suprir as necessidades energéticas durante a germinação. Uma vez que o teor de amido influencia diretamente no valor calórico do alimento, a quantificação de seus teores poderá ser utilizada como indicativo indireto de valor nutricional. A quantidade de amido no grão de arroz pode variar entre diferentes genótipos devido a fatores genéticos, do ambiente e das interações genótipo versus ambiente. Além disso, o processamento também influencia o percentual de amido. Todos estes fatores têm impacto na composição do amido e



influenciam profundamente as propriedades físico-químicas do arroz causando variação na proporção de amilose e amilopectina.

O amido no endosperma do arroz é formado por amilose e amilopectina. Sendo o teor de amilose no amido do endosperma uma das características importantes, correlacionada com as propriedades texturais, como maciez, coesão, cor, brilho e volume de expansão (Pandey et al., 2012). E as diferenças varietais na estrutura da amilopectina existem predominantemente devido à variação do comprimento da cadeia de ligações  $\alpha$ -1,4 e  $\alpha$ -1,6 que também desempenham um papel crítico na determinação das propriedades físico-químicas do amido no endosperma. Além da qualidade de cozimento (CQ), o conteúdo de amilose (AC) e temperatura de gelatinização (GT) são as principais medidas para avaliar a qualidade do grãos de arroz. Sendo que AC determina a firmeza e a natureza pegajosa do arroz cozido, enquanto o arroz com alto GT requer temperatura mais alta, mais água e tempo para cozinhar do que aqueles genótipos com GT baixo ou intermediário. Como o GT está diretamente correlacionado ao tempo necessário para cozinhar o arroz, portanto, arroz com GT intermediário são preferidos sobre aqueles com GT alto ou baixo. Estas duas propriedades têm maior efeito na qualidade do grãos de arroz cozido e, portanto, desempenham um papel importante em influenciar a preferência do consumidor. Em vários estudos, ambos AC e GT foram encontrados altamente associados com alimentação e propriedades culinárias do arroz (Pandey et al., 2012).

As variações da estrutura da amilopectina e amilose surgem devido à expressão diferencial de várias isoformas das enzimas que sintetizam o amido. Quatro classes de enzimas catalisam as reações de síntese do amido e seus genes são chamados de SSRGs: AGPase, SS, SBE e DBE. As enzimas ADP – glicose pirofosforilase (AGPase) catalisam a produção do substrato comum (ADP-glicose), para a síntese da amilose e da amilopectina, mesmo sendo estas sintetizadas por vias diferentes. As enzimas ramificadoras do amido (SBE) atuam na ramificação das cadeias de glicose, enquanto enzimas desramificadoras do amido (DBE) atuam na linearização das cadeias de glicose. O grupo das amido sintases (SS), dividem-se em amido sintase ligada ao grânulo do amido (GBSS – amido sintase insolúvel), controla a síntese de amilose no endosperma do arroz, enquanto a amido sintase solúvel (SS) juntamente com as SBE e DBE atuam no controle da síntese da amilopectina. Algumas destas enzimas possuem múltiplas isoformas, codificadas por diferentes

locos da família gênica localizados em cromossomos distintos, sendo que cada loco desempenha um papel distinto na biossíntese do amido. Além disso, todas estas isoformas das enzimas atuam juntas coordenadamente e formam uma rede de regulação para controlar a síntese de amido no endosperma de arroz que afetam tanto a produtividade quanto a qualidade de grãos (Fasahat et al., 2014; Pandey et al., 2014; Hirose & Terao, 2004; Yu et al., 2011).

Estudos envolvendo a evolução e estrutura dos genes e proteínas SSRGs existem, porém não consideram todos os 19 genes, e muito menos investigam e discutem a variação das estruturas entre as espécies selvagens de *Oryza* (Tian et al., 2009; Ohdan et al., 2005; Yu et al., 2011). Além disso, sabe-se que a qualidade é controlada por fatores decisivos, como já apontado, e, reside nos próprios genomas. Por isso, atualmente, temos disponível no banco de dados Ensembl Plants, 11 genomas de espécies de *Oryza* completamente sequenciados, sendo então fonte para diversos estudos. Assim, é necessário que pesquisas de base sejam desenvolvidas no sentido de fornecer insights para programas de melhoramento genético visando qualidade de grão, porém deve-se preconizar, que de forma previsível, possa ser possível alterar componentes de qualidade. Especificamente, a melhoria na qualidade do amido no endosperma de arroz, especialmente para aumentar o nível de amilose, é extremamente importante para que nos programas de melhoramento, sejam selecionadas novas cultivares de qualidade, alto rendimento industrial e aceitação dos padrões culinários de consumo, principalmente no Brasil e também no mundo.

### **2.3 Estresses abióticos/bióticos afetam a qualidade do grão de arroz**

Como apontado anteriormente, enquanto os produtores e os processadores exigem estabilidade de rendimento de grãos e uniformidade de produto, o desafio do momento é manter a qualidade do grão frente a padrões climáticos variáveis. Ser capaz de prever a natureza físico-química geral do amido como resultado do status de crescimento é um passo em direção à agricultura "precisa" necessária para o século 21<sup>a</sup> (Beckles & Thitisaksakul, 2014).

Nesse sentido, fatores decisivos, como o genoma, são a chave para o controle da qualidade no grão de arroz. Porém, o ambiente externo pode afetar, na maioria das vezes, de forma negativa a qualidade do grão. Fatores bióticos como pragas e

doenças, e fatores abióticos, como temperatura, água, salinidade, fertilidade, entre outros podem alterar as cadeias de formação da amilose e amilopectina levando a severa perda de qualidade (Beckles & Thitisaksakul, 2014). Além disso, nos próximos anos as mudanças ambientais se tornarão cada vez mais imprevisíveis (Battisti & Naylor, 2009). Os produtores, em âmbito mundial, já enfrentam mudanças nas épocas de plantio, temperaturas noturnas mais altas do que a média, diminuição da disponibilidade de água e deterioração da qualidade do solo, que pode alterar o acúmulo de amido e as características físicas, tornando a qualidade do grão e da farinha menos desejável para alguns usos posteriormente pretendidos (Hatfield et al., 2011). Antecipar essas mudanças pode ajudar cada grupo ao longo da cadeia de abastecimento a planejar e direcionar os mercados adequados para esses “produtos estressados”. Porém a melhor forma de evitar perdas na lavoura ainda é, e sempre vai ser, o uso de genótipos tolerantes oriundos em programas de melhoramento genético.

As cultivares de arroz precisam ser melhoradas para que suportem o aumento na capacidade de adaptação a ambientes novos e estressantes, bem como ciclo curto juntamente com maior produção, qualidade nutricional superior e resistência a tensões ambientais. Cultivares aprimoradas são necessárias para combater essas ameaças, permitindo que haja rendimento/capacidade para alimentar a crescente população global. Espécies de arroz selvagem tornam-se uma enorme fonte de genes valiosos que podem resistir a muitas tensões. Numerosos genes, especialmente para estresses bióticos e abióticos de arroz selvagem foram incorporados ao arroz cultivado para produzir novas variedades de elite. Essas variedades mostram resistência a fungo de planta marrom, vírus do tungro, crestamento bacteriano e tolerância à salinidade, seca e sulfato ácido entre outros. Desta forma, a segurança alimentar pode ser alcançada através da busca de variação natural em espécies de arroz, identificando novos alelos e genes necessários para o desenvolvimento de variedades de arroz (Ricachenevsky et al., 2016).

## **2.4 O que se sabe sobre as características de qualidade nas espécies selvagens?**

Até o momento, não se sabe muito sobre a composição química e nutricional dos grãos nas espécies selvagens e fatores genéticos que podem ser úteis para o

melhoramento da qualidade de grãos, embora esta abordagem possa ter um grande potencial. Porém, os poucos estudos disponíveis apontam que as características que influenciam a aparência do grão, valor nutricional, qualidade funcional (por exemplo, características de cozimento) e desempenho de processamento podem ser provenientes de espécies selvagens. Os genes úteis disponíveis incluem muitos controles de tamanho e forma de grão, cor de grão e propriedades de amido relacionados aos genes de síntese e outras características de qualidade, sendo que os avanços na tecnologia genômica estão tornando essa diversidade genética mais facilmente disponível das espécies selvagens para uso no melhoramento genético do arroz cultivado, além de apenas aumentar a produtividade (Hasan et al., 2020).

Os efeitos ou resultado dos principais estresses abióticos, desidratação, frio e salinidade, sob a qualidade de grão pode ser observada, principalmente por reduzir o acúmulo de amido em até 40%, levando a mudanças na composição do amido, estrutura e funcionalidade. Estudos tem mostrado isso em arroz e trigo (Gunaratne et al., 2011 e Liu et al., 2010). Outras espécies como a cevada, são mais resistentes (Brooks et al., 1982).

Nesse sentido, pesquisas envolvendo o estudo dos *loci* que codificam amido sintetases e proteínas de armazenamento de sementes, bem como seus reguladores de ação cis e trans, é algo de substancial relevância. Sendo, esses dois últimos, os elementos que afetam altamente a atividade desses genes em diferentes condições.

## **2.5 Elementos de regulação cis (CREs) podem fornecer novas possibilidades para a engenharia genética**

Através de modificações nos padrões de expressões gênicas, as células conseguem se adaptar a todo momento sob diferentes condições ambientais. É possível observar que nem todos os genes são expressos no mesmo momento (LAMBERT, 2018). Para que um organismo se desenvolva e sobreviva é necessário que ocorra o controle da expressão de genes (CASTRILLO et al., 2011). A introdução de novos genes via transgenia ou a alteração de alelos via edição gênica são ferramentas biotecnológicas que podem auxiliar no controle da expressão de características de interesse agrônomo (ZHANG, et al., 2005). A expressão gênica só é possível devido a presença de proteínas, que são conhecidas como fatores de

transcrição (FT) que fazem ligação a elementos de ação cis (CREs), presentes nos promotores dos genes, para ativar ou reprimir um gene (TAN et al., 2020). E essa regulação ocorre tanto em condições favoráveis a planta, quanto em condições adversas.

Os CREs, como potenciadores ou promotores, controlam o desenvolvimento e a fisiologia pela regulação da expressão gênica. Mutações na sequência dos CREs que afetam a função desses elementos contribuem para a diversidade fenotípica dentro e entre as espécies, o que torna esses elementos alvo de muitos estudos utilizando a engenharia genética. Com muitos estudos de caso mostrando atividade regulatória divergente na evolução fenotípica, que puderam ser provados através de abordagens que incluem análise funcional detalhada de CREs individuais e comparação de mecanismos de regulação gênica entre espécies usando as ferramentas genômicas recentes (Wittkopp & Kalay, 2012).

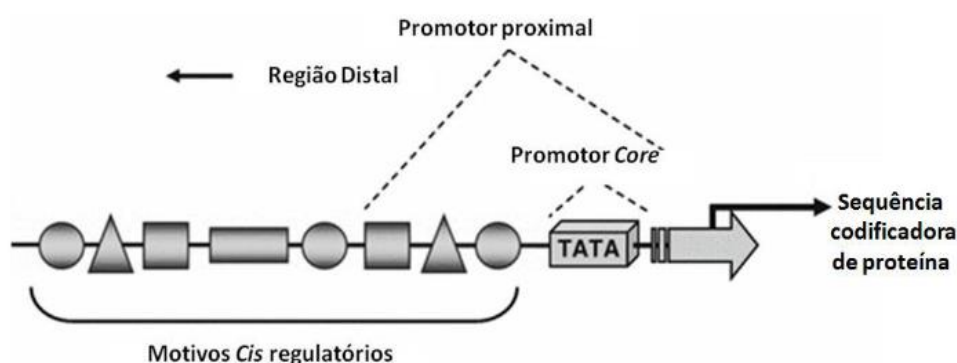


Figura 2. Diagrama representativo de um típico promotor eucariótico. Esquema adaptado de VENTER & BOTHA (2010).

Nesse contexto, os avanços nas pesquisas e na precisão de seus resultados em perfis de expressão do transcriptoma tem levado à identificação de várias combinações de atuação (*cross-talk*) dos elementos cis, nas regiões promotoras de genes induzidos por estresses como a desidratação, temperaturas extremas, e condições do solo, entre muitos outros. E envolvidos com respostas hormonais. Existe a região core que é definida como a porção mínima necessária para direcionar a transcrição. Nesta porção estão localizados elementos regulatórios importantes como o TATA box, com a sequência consenso TATAAA, e o iniciador (Inr) à montante do

ponto de início da transcrição. Ainda podemos ter o elemento promotor a jusante (DPE, do inglês, *downstream promoter* elemento) (VENTER & BOTHA, 2004).

Os sítios de ligação compreendem a menor parte dos nucleotídeos dentro de uma região promotora, a qual é geralmente intercalada por regiões maiores sem sítios de ligação. Os espaçamentos entre os sítios de ligação de fatores de transcrição podem apresentar grandes variações de tamanho, dependendo das interações proteicas que ocorrem durante a ligação aos elementos cis (WRAY et al., 2003). O reconhecimento de CREs essenciais para as plantas na tolerância a estresses abióticos, bem como interessantes alvos para a edição já foram relatados (Swinen et al., 2019).

Nos promotores dos genes da qualidade do arroz, nenhum estudo abordando as vias de transdução de sinal e o incremento dos CREs nas espécies selvagens foi realizado. Os estudos mais recentes ainda estão tentando compreender a complexa rede de regulação que ocorre na formação do amido, avaliando a expressão temporal dos genes envolvidos (Pandey et al., 2012; Yu et al., 2011), aspectos genômicos estruturais (De Freitas et al., 2021; Batra et al., 2017; Georgelis et al., 2008), entre outros estudos, inclusive utilizando a engenharia genética no nocaute ou mutação dos genes de síntese de amido para obtenção de mutantes mais assertivos e eficientes na produção da proporção amilose/amilopectina (Sun et al., 2010; Shufen et al., 2019; Xu et al., 2020). Apesar do grande esforço na compreensão da via de síntese do amido, ainda é necessário compreender e explorar as divergências na evolução fenotípica através do estudo dos CREs nos promotores SSRGs nas espécies selvagens de *Oryza*, para mais tarde, permitir que se altere efetivamente o padrão de expressão desses genes de maneiras específicas, criando fenótipos tolerantes as pressões ambientais sem alterar negativamente a qualidade do grão.

## **2.6 Microssatélites ou SSRs**

Ainda sob o contexto dos avanços da tecnologia do DNA, principalmente o sequenciamento genômico, foi possível tornar disponível nos bancos de dados uma ampla gama de sequências para uso em diversos estudos, entre esses destaca-se os estudos para a compreensão da evolução das espécies através de sequências específicas do genoma.

Uma classe de sequências alvo de estudos nos últimos 20 anos, são os chamados elementos repetidos, dentre eles podemos citar os microssatélites, também conhecidos como SSRs (sequências de 1 a 10 nucleotídeos) repetidos não aleatoriamente no genoma. Minissatélites (> 10 nucleotídeos) são subcategorias de repetições em tandem (TRs). Além destas, as repetições intercaladas predominantes (ou remanescentes de elementos transponíveis), constituem regiões repetitivas genômicas. Os TRs são evolutivamente relevantes devido à sua instabilidade. Em geral, eles sofrem mutação a taxas entre  $10^{-3}$  e  $10^{-6}$  por geração de células, ou seja, até 10 ordens de magnitude maiores do que as mutações pontuais (Vieira et al., 2016).

De maneira geral, pode-se afirmar que a ocorrência de SSRs é menor em regiões gênicas, devido ao fato de os SSRs apresentarem uma alta taxa de mutação que pode comprometer a expressão gênica. Estudos indicam que em regiões codificantes há predominância de SSRs com motivos gênicos do tipo tri e hexanucleotídeo, resultado da pressão de seleção contra mutações que alterem o quadro de leitura (Zhang et al., 2004). O surgimento de novos alelos ocorre devido as mutações escaparem da correção pelo sistema de reparo de incompatibilidade de DNA. Por esse motivo, diferentes alelos podem existir em um determinado locus SSR, o que significa que SSRs são mais informativos do que outros marcadores moleculares, incluindo SNPs.

A importância desses elementos ocorre devido serem os marcadores mais amplamente usados para genotipagem de plantas nos últimos 20 anos pois são marcadores genéticos multialélicos altamente informativos, codominantes, que são experimentalmente reprodutíveis e transferíveis entre espécies relacionadas (Mason, 2015). Especificamente, os SSRs são úteis para espécies selvagens (i) em estudos de diversidade medida com base na distância genética; (ii) estimar o fluxo gênico e as taxas de cruzamento; e (iii) em estudos evolutivos, sobretudo para inferir relações genéticas intraespecíficas. Por outro lado, para plantas cultivadas, os SSRs são comumente usados para (i) construir mapas de ligação; (ii) mapear loci envolvidos em características quantitativas (QTL); (iii) estimar o grau de parentesco entre genótipos; (iv) realizar seleção assistida por marcador; e (v) definir as impressões digitais de DNA de cultivares (Jonah et al., 2011; Kalia et al., 2011).

Evolutivamente, apesar da ampla aplicabilidade dos SSRs como marcadores genéticos desde sua descoberta na década de 1980 (Tautz e Renz, 1984), pouco se sabe sobre a importância biológica dos SSRs, especialmente em plantas. Um alta

frequencia de SSRs existe em regiões transcritas de diversas espécies, especialmente em regiões não traduzidas - UTRs (Morgante et al., 2002). Curiosamente, existem dados substanciais indicando que as expansões ou contrações de SSR em regiões de codificação de proteínas podem levar a um ganho ou perda da função do gene por meio de mutação de frameshift ou mRNAs tóxicos expandidos. Além disso, as variações de SSR em 5'-UTRs podem regular a expressão gênica afetando a transcrição e a tradução, mas as expansões nas 3'-UTRs causam deslizamento da transcrição e produzem mRNA expandido, que pode interromper o splicing e pode interromper outras funções celulares. Os SSRs intrônicos podem afetar a transcrição do gene, o splicing do mRNA ou a exportação para o citoplasma. SSRs tripletos localizados em UTRs ou íntrons também podem induzir o silenciamento de genes do tipo mediado por heterocromatina. Todos esses efeitos podem eventualmente levar a mudanças fenotípicas (Li et al., 2004).

Por isso, diversos estudos já foram conduzidos no sentido de identificar SSRs em espécies prospectando iniciadores para serem utilizados na caracterização de espécies/populações e diversidade genética em diferentes espécies de plantas de uso econômico ou inclusive aquelas espécies que carecem de estudos moleculares. Além disso, outros estudos focam no entendimento da importância desses elementos e o impacto das variações durante a evolução (Vieira et al., 2016).

O estudo desses elementos em genomas acessórios de plantas se torna ainda mais importante, já que esses genomas geralmente são pequenos e apresentam forte pressão de seleção para variações. Adicionalmente, o impacto da variação de um SSR em um genoma mitocondrial ou plastidial pode ter reflexos extremamente negativos durante a evolução. Alguns estudos comparativos já foram conduzidos em genomas acessórios e revelaram aspectos evolutivos importantes como grande número de repetições de monômeros em comparação com dímeros e trímeros; repetições di-, tri-, tetra-, penta-, hexâmeros estão presentes em todos os genomas de plastídio investigados; e organismos mais basais como as algas mostram ampla variação de número e distribuição de SSRs entre os genomas (George et al., 2015). Porém ainda há a necessidade de revelar novos aspectos através de estudos comparativos entre as espécies principalmente relacionados a evolução desses elementos no mtDNA das plantas, que codificam proteínas associados ao metabolismo energético.



## **2.7 *Avena sativa* L.**

Aveia (*Avena sativa* L.) é um alimento saudável e bastante consumido no mundo, devido conter no seu grão teores de beta-glucano, uma fibra que pode reduzir ativamente o nível de colesterol de lipoproteína de baixa densidade e diminuir o risco de doença cardíaca coronária (EFSA, 2010). Por isso, é cultivada podendo ser útil como pasto para gado, como fertilizante verde para o solo e para a produção de grãos. Até recentemente, a produção de aveia no Brasil era baixa devido principalmente à ausência de cultivares bem adaptadas. Programas de melhoramento foram implementados para aumentar a adaptabilidade, rendimento de grãos e conteúdo de proteínas (Pedó et al., 1998). Adicionalmente a preservação e o uso de germoplasma de espécies de aveia selvagem são essenciais para o melhoramento da aveia cultivada, sendo necessário o desenvolvimento de marcadores moleculares.

Recentemente, foram sequenciados e publicados os genomas plastidiais de 26 espécies de *Avena* (Liu et al., 2020 e Fu, 2018), que estão disponíveis em bancos de dados públicos para além de estudos de evolução do genoma da aveia, também conservação e utilização de germoplasma de aveia em programas de melhoramento. Estudos tem revelado a importância dos marcadores baseados em SSRs plastidiais (Da Silva et al., 2011 e Li et al., 2008).

### **3 HIPÓTESE E OBJETIVOS**

#### **3.1 Hipóteses**

- A distribuição dos microssatélites pode explicar a evolução dos genomas mitocondriais em plantas através da análise de número e distribuição de SSRs;
- As espécies selvagens de arroz podem oferecer genes e alelos que ajudem a melhorar a qualidade de grão das espécies cultivadas;
- Variações no conteúdo de amido frente diferentes ambientes podem ser controladas através das CREs dos genes de síntese de amido em arroz;
- A amplificação de SSRs de genomas plastidiais constituem potencial ferramenta para distinção de populações e espécies de Aveia.

#### **3.2 Objetivo Geral**

Contribuir com informações acerca das sequências microssatélites em genomas acessórios e genes de síntese de amido que possam ser úteis para entender a evolução dos genomas e compor pesquisas específicas e estratégias para o melhoramento genético para qualidade de grão e tolerância a estresses em arroz.

### 3.3 Objetivos Específicos

- Revelar o número, abundância e densidade de microssatélites nas diferentes regiões dentro dos genomas mitocondriais comparando e discutindo os resultados entre 204 espécies de plantas e algas.
- Identificar, caracterizar a estrutura, localização e filogenia dos SSRGs nas espécies cultivadas e selvagens do gênero *Oryza*.
- Revelar os CREs relacionados a estresses abióticos/bióticos, hormônios, amilose e desenvolvimento celular nos promotores dos SSRGs relacionando alterações no conteúdo de amido frente a, principalmente, estresses bióticos e abióticos e as possíveis funções dos CREs já relatadas na mitigação dos efeitos dos estresses.
- Propor uma estratégia, através da identificação de um SNP, para ser testada e utilizada em programas de melhoramento para tolerância aos principais estresses que afetam a qualidade do grão (desidratação, frio e salinidade) em arroz.
- Identificar os microssatélites e propor iniciadores para as regiões SSRs nos genomas plastidiais nas espécies do gênero *Avena* disponíveis nos bancos de dados.

## **4 CAPÍTULOS**

### **4.1 Artigo 1 – An empirical analysis of mtSSRs: Could microsatellites distribution patterns explain the evolution of mitogenomes in plants?**

*Artigo publicado na revista Functional and Integrative Genomics (Springer)*



# An empirical analysis of mtSSRs: could microsatellite distribution patterns explain the evolution of mitogenomes in plants?

Karine E. Janner de Freitas<sup>1</sup> · Carlos Busanello<sup>1</sup> · Vívian Ebeling Viana<sup>1</sup> · Camila Pegoraro<sup>1</sup> · Filipe de Carvalho Victoria<sup>2</sup> · Luciano Carlos da Maia<sup>1</sup> · Antonio Costa de Oliveira<sup>1</sup>

Received: 6 July 2021 / Revised: 18 October 2021 / Accepted: 19 October 2021

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

## Abstract

Microsatellites (SSRs) are tandem repeat sequences in eukaryote genomes, including plant cytoplasmic genomes. The mitochondrial genome (mtDNA) has been shown to vary in size, number, and distribution of SSRs among different plant groups. Thus, SSRs contribute with genomic diversity in mtDNAs. However, the abundance, distribution, and evolutionary significance of SSRs in mtDNA from a wide range of algae and plants have not been explored. In this study, the mtDNAs of 204 plant and algal species were investigated related to the presence of SSRs. The number of SSRs was positively correlated with genome size. Its distribution is dependent on plant and algal groups analyzed, although the cluster analysis indicates the conservation of some common motifs in algal and terrestrial plants that reflect common ancestry of groups. Many SSRs in coding and non-coding regions can be useful for molecular markers. Moreover, mitochondrial SSRs are highly abundant, representing an important source for natural or induced genetic variation, i.e., for biotechnological approaches that can modulate mtDNA gene regulation. Thus, this comparative study increases the understanding of the plant and algal SSR evolution and brings perspectives for further studies.

**Keywords** Simple sequence repeats · Repetitive DNA · Mitogenomes · Land plants · Algae

## Introduction

Microsatellites or SSRs (simple sequence repeats) are tandem repeats that have a high mutation rate, exhibiting expansion or contraction, mainly due to replication errors caused by DNA polymerase slippage (Levinson and Gutman 1987; Schlotterer and Tautz 1992; Tautz and Schlotterer 1994; Gao et al. 2013); besides that, recombination events can allow diversification through rearrangement mainly in tracheophytes (Hecht et al. 2011).

SSRs are continuously used as robust molecular markers, marker-assisted selection, QTL mapping, population genetics, positional cloning of target genes, and DNA fingerprinting (Vieira et al. 2016). SSRs are present in both eukaryotic and prokaryotic genomes (Li et al. 2009; Oliveira et al. 2006). Moreover, they are found not only in nuclear genomes (Zhao et al. 2014) but also in cytoplasmic genomes such as plastidial (Provan et al. 2001; Devey et al. 2009; Sonah et al. 2011; Jiang et al. 2012; George et al. 2015) and mitochondrial genomes (mtDNA) (Soranzo et al. 1999; Rajendrakumar et al. 2006; Fauron et al. 2004; Kuntal and Sharma 2011; Zhao et al. 2016; Filiz 2014; Liu et al. 2014;

✉ Antonio Costa de Oliveira  
acostol@terra.com.br

Karine E. Janner de Freitas  
karinejanner@gmail.com

Carlos Busanello  
carlosbuzza@gmail.com

Vívian Ebeling Viana  
vih.viana@gmail.com

Camila Pegoraro  
pegorarocamilanp@gmail.com

Filipe de Carvalho Victoria  
filipevictoria@unipampa.edu.br

Luciano Carlos da Maia  
lucianoc.maia@gmail.com

<sup>1</sup> Plant Genomics and Breeding Center, Technology Development Center, Federal University of Pelotas, Pelotas, RS, Brazil

<sup>2</sup> Center for Antarctic Vegetation Studies (NEVA), Federal University of Pampa, São Gabriel, RS, Brazil



Raju et al. 2015). SSRs are abundant in mitochondrial genomes (mtSSRs), presenting a non-randomized intra- and intergenic distribution (Hancock 1999; Bajaj et al. 2015). Also, variation in SSR length can be observed (Soranzo et al. 1999; Kuntal and Sharma 2011; Filiz 2014) contributing to genome diversity.

The mtDNAs available in the GenBank database (NCBI Genome Information) indicate a wide range in sizes across different taxonomic groups of plants that tend to have large genomes when compared to metazoans (Gissi et al. 2008) and fungi (Fauron et al. 2004). In plants, some factors interfere with the mtDNA complexity, such as presence of introns and repeated elements. In addition, mtDNA experiences frequent gene gain/loss/transfer/duplication, and genome rearrangements (Kitazaki and Kubo 2010), which agree with the large number of variations in genome size.

Plant mtDNAs are known to have genes that encode proteins associated with energy metabolism (Race et al. 1999). However, little is known about mutations due to DNA polymerase slippage as a causal mechanism of mtDNA polymorphisms. Thus, investigating the distribution and consequences of variations and mutations in mtDNAs may provide new insights into mtSSRs in plants. Different studies have identified mononucleotides (George et al. 2015; Kuntal et al. 2012), dinucleotides (Rajendrakumar et al. 2006; Anand et al. 2019), and trinucleotides (Filiz 2014) as the most abundant type of SSR in mtDNA of different groups of plants. On the other hand, SSRs are often disrupted by single base substitution; therefore, short mononucleotides ( $\geq 6$  nt) and these types also were mined in mtDNA of specific groups of plants (Rajendrakumar et al. 2006), and as already reported, abundant in plant chloroplast genomes (George et al. 2015).

The importance of SSRs in generating genomic variability and in evolution of mtDNAs may be more pronounced than previously thought. Considering such hypothesis and seeking to assess the importance of mtSSRs of a wide range of existing plant and algal species, this study aimed to locate and compare the abundance and distribution of SSRs in

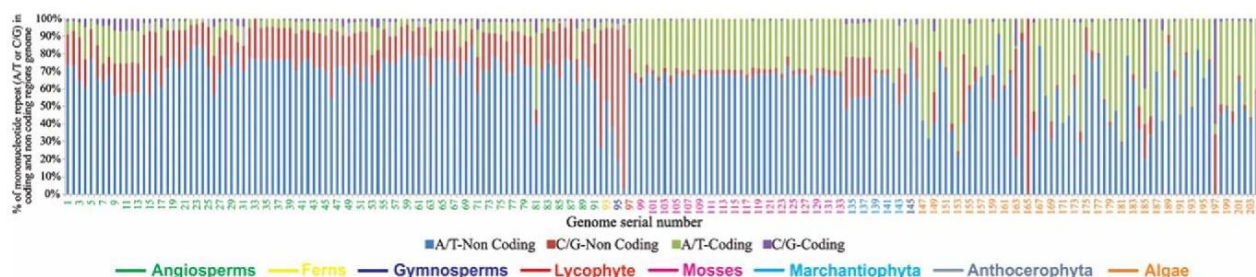
complete mtDNAs of plants and algae. Our results provide a comparative diagnosis of mtSSRs in different plant taxonomic groups, discussing the findings in terms of evolution of land-based plants.

## Results

### SSR1 (mononucleotides) in algal and plant mtDNAs: features and overrepresentation

The different taxonomic groups differed in their specific repetitions in mononucleotide type (SSR1- motif composed of one nucleotide A, or T, C, G, repeated more than six times) (Supplementary file 1 - Tables S1 and S2). SSR1s were found in higher number in groups such as algae, mosses, Marchantiophyta, Anthoceroophyta, Lycophyta, and ferns. In mtDNAs of gymnosperms and angiosperms, SSR1 types were the second most abundant type of SSR. Among the SSR1 types, poly (A) or (T) was much more abundant than poly (C) or (G) (Supplementary File 1- Table S2). Poly (C) or (G) was detected in higher numbers in ferns, gymnosperms, and Lycophyta (Figure 1). In algae, a lower G/C content was observed.

The species selected for the analysis, identified by their respective genome serial numbers in text and tables, showed wide variation in mtDNA size (algae: 12,998–56,574 bp; Anthoceroophyta: 184,908–209,482 bp; Marchantiophyta: 106,358–186,609 bp; mosses: 100,725–141,276 bp; Lycophyta: 413,530 bp; ferns: 364,070–372,339 bp; gymnosperms: 346,544–978,846 bp; flowering plants: 191,481–1,999,602 bp) (Supplementary file 1 – Table S1). Thus, to facilitate the comparison among genomes of different sizes, we standardized SSRs per 1 kilobase (kb) of genome. Relative abundance and relative density were calculated. Relative abundance (RA) gives the number of motifs per genome kb, while relative density (RD) gives the total length (in nucleotide) that contributes to 1 kb SSR of the genome (see “Material and methods”).



**Figure 1** Mononucleotide (SSR1) distribution in the mitochondrial genome for the different taxonomic groups of plants and algae. Groups of plants and algae are represented with f. The serial number

(SN) below the bars represents the species according to Supplementary file 1 - Tables S1–S4



A total of 142,417 mononucleotide repeating more than 6 times ( $\text{SSR1} \geq 6$ ) were distributed in the mtDNA of all 204 analyzed species (Supplementary file 1 - Table S1). When comparing the mtDNA from plants and algae, a wide variation in the number of  $\text{SSR1} \geq 6$  was observed. For example, in angiosperms and gymnosperms, the minimum numbers identified were 406  $\text{SSR1}$  in *Brassica carinata* (SN 9) and 757 in *Ginkgo biloba* (SN 95). The maximum numbers of  $\text{SSR1}$  were observed in *Corchorus capsularis* (SN 22) with 2,883  $\text{SSR1}$  and 1,421  $\text{SSR1}$  in the lycophyte *Phlegmariurus squarrosus* (SN 97). On the other hand, in algae, the lowest incidence was of five  $\text{SSR1}$  in *Polytoma uvella* (SN 185) and *Chlamydomonas leiostraca* (SN 154), whereas 1030, the highest detected, was found in *Roya obtusa* (SN 195). These differences in  $\text{SSR1}$  number among the algal species also extend to relative abundance values (minimum RA: 0.29  $\text{SSR1}/\text{kb}$  in *Polytoma uvella*) (maximum RA: 14.83  $\text{SSR1}/\text{kb}$  in *Roya obtusa*), and relative density (minimum RD: 1.83 nt/kb in *Polytoma uvella*) (maximum RD: 103.87 nt/kb *Roya obtusa*) in the same species (Supplementary file 1 - Table S1). However, the higher RA (Figure 2) and RD values were observed in mosses (RA: 5.43, range 4.19 to 5.94  $\text{SSR1}/\text{kb}$ ) (RD: ranged 28 to 43 nt/kb) and some liverwort species (RA: 4.35, range 3.15 to 5.82  $\text{SSR1}/\text{kb}$ ) (RD: 42 nt/kb) (Supplementary file 1 - Table S1).

In order to assess whether  $\text{SSR1}$ s were over- or under-represented in plant and algal mtDNA, an analysis of the expected number of  $\text{SSR1}$  was performed. The observed number of  $\text{SSR1}$  was overrepresented in 0.25–6.72 times, i.e., higher than the expected number of  $\text{SSR1}$ , as calculated by De Watcher's equation (Watcher 1981) (see Eq. 1). Supplementary file 1 - Table S1 (Z-score) presents the statistical significance of the O/E ratio. The largest overrepresentation of  $\text{SSR1}$  occurred in the algal species *Klebsormidium nitens* (SN 169) and *Lobosphaera incisa* (SN 170). Underrepresentation was also detected in five species of algae, being *Chlamydomonas leiostraca* (SN

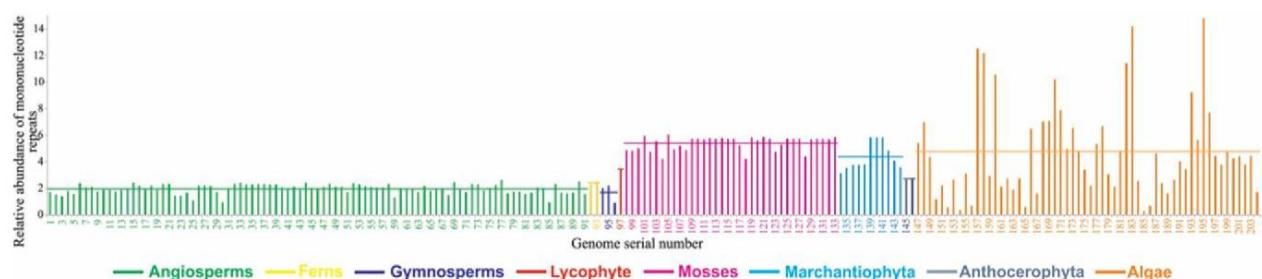
154) the one showing the highest underrepresentation ( $-3.30$ ) (Supplementary file 1 - Table S1).

To identify correlations between genomic features (genome size and GC content) and among repeated sequences ( $\text{SSR1}$ , RA, and RD), a linear regression was performed (Table S6 - A). A strong positive correlation between  $\text{SSR1}$  number and genome size ( $R^2 = 0.79$ ,  $P < 0.05$ ) was detected; however, a very low positive correlation with GC content ( $R^2 = 0.08$ ,  $P < 0.05$ ) was observed. Furthermore, it was noted that RA of  $\text{SSR1}$  was low positively correlated with genome size ( $R^2 = 0.21$ ,  $P < 0.05$ ) and with GC content ( $R^2 = 0.46$ ,  $P < 0.05$ ). Similarly,  $\text{SSR1}$  RD presented a low correlation with genome size ( $R^2 = 0.21$ ,  $P < 0.05$ ) and with GC content ( $R^2 = 0.41$ ,  $P < 0.05$ ) (Supplementary file 1 - Table S6 - A).

### Features of $\text{SSR2-6}$ in the mitochondrial genome of plants and algae

A total of 188,007  $\text{SSR2-6}$  repeating more than 3 times ( $\text{SSR2-6} \geq 3$  times) have been identified in plant and algal mtDNAs.  $\text{SSR2-6}$  repetitions were more abundant in angiosperms, gymnosperms, ferns, and Anthocerophyta (Figure 3 and Supplementary file 1 - Table S1).

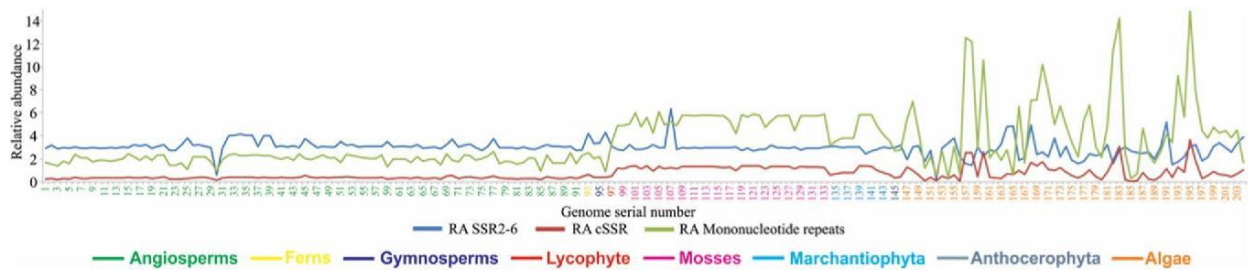
Dinucleotides ( $\text{SSR2}$ -motif composed of two nucleotide AT, TC, GC..., repeated more than three times) showed the higher number and abundance in angiosperms and gymnosperms, and were the second type showing higher number and abundance in ferns, Lycophyta, and Anthocerophyta. In mosses, Marchantiophyta, and algae,  $\text{SSR1}$  followed by trinucleotides ( $\text{SSR3}$ -motif composed of three nucleotide repeated more than three times) presented the highest RA. However, when comparing the groups, tetranucleotides ( $\text{SSR4}$ -motif composed of four nucleotide repeated more than three times) and pentanucleotides ( $\text{SSR5}$ -motif composed of five nucleotide repeated more than three times) were more abundant in ferns (Figure 4 and Supplementary file 1 - Table S2).



**Figure 2** Graphical representation of the relative abundance of mononucleotide SSRs ( $\text{SSR1}$ ) in the mitochondrial genome of species among the different taxonomic groups of plants and algae. The groups of plants and algae are represented with different colors. The

serial number below the bars represents the species according to Supplementary file 1. The mean relative abundance is represented by a trace on the bars of the graph





**Figure 3** Graphical representation of the relative abundance of dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide-SSRs (SSR2-6), composite SSRs (cSSR) and for comparative perspective mononucleotide (SSR1), in the mitochondrial

genome of different plant and algal groups. The groups of plants and algae are represented with different colors. The serial number below the bars represents the species according to Supplementary file 1 - Table S1

Angiosperms	SSR2 >	SSR1 >	SSR3 >	SSR4 >	SSR5 >	SSR6 >	SSR7 >	SSR8 >	SSR9 >	SSR10
Gymnosperms	SSR2 >	SSR1 >	SSR3 >	SSR4 >	SSR5 >	SSR6 >	SSR7 >	SSR9 >	SSR8 >	SSR10
Ferns	SSR1 >	SSR2 >	SSR4 >	SSR3 >	SSR5 >	SSR9 >	SSR6 >	SSR7 >	SSR8 >	SSR10
Lycophytes	SSR1 >	SSR2 >	SSR3 >	SSR4 >	SSR5 >	SSR8 >	SSR6 >	SSR10 >	SSR7 >	SSR9
Anthocrophyta	SSR1 >	SSR2 >	SSR3 >	SSR4 >	SSR5 >	SSR6 >	SSR9 >	SSR7 >	SSR8 >	SSR10
Mosses	SSR1 >	SSR3 >	SSR2 >	SSR4 >	SSR5 >	SSR6 >	SSR7 >	SSR10 >	SSR9 >	SSR8
Marchantiophyta	SSR1 >	SSR3 >	SSR2 >	SSR4 >	SSR5 >	SSR9 >	SSR6 >	SSR7 >	SSR8 >	SSR10
Algae	SSR1 >	SSR3 >	SSR2 >	SSR5 >	SSR4 >	SSR6 >	SSR7 >	SSR8 >	SSR10 >	SSR9

**Figure 4** Most abundant range of types SSR among different groups of plants and algae

The highest number of SSR2-6 was observed in the large genomes as *Corchorus capsularis* mtDNA (5408 SSRs) (SN 22) (Supplementary file 1 - Table S1) what is according to the correlation analysis below. Few variations in SSR2-6 were detected in bryophytes; values ranging from 259 in the liverwort *Codiophorus laevigatus* (SN 140) to 693 in the moss *Funaria hygrometrica* (SN 107) were detected. The algal group showed the smallest number of SSR2-6, 26 in *Polytomella piriformis* (SN 189). Also, the algal group presented a wide variation in RA (ranging from 0.25 SSR2-6 (SN 152) to 5.16 SSR2-6 (SN 191)) (Supplementary file 1 - Table S1).

Still, some features of SSR1 and SSR2-6 (SSRs motifs from two to six nucleotides repeated more than three times) were conserved between mosses and Marchantiophyta. High SSR1 RD was observed in these two taxa possibly due to the smaller size of mtDNA and larger sizes of SSR1 motifs (Supplementary file 1 - Table S1 - columns R to Z). Similarly, SSR1 RA and SSR2-6 RA values were higher due to the higher number of SSR and smaller genome size of moss and Marchantiophyta mtDNA. The SSR3 type was the second most abundant type in these two taxa, and SSR6 (motif composed of six nucleotide repeated more than three times) and SSR4 types were very poor and sometimes absent in the non-coding region.

Finally, the most common repeated motif AT/TA was not shared with the sister group (Anthocrophyta).

Using a linear regression analysis, we detected that the number of SSR2-6 (Table S6 - B) was strongly positively correlated with genome size ( $R^2 = 0.93$ ,  $P < 0.05$ ) and weakly positively correlated with GC content. RA has a low positive correlation with genome size ( $R^2 = 0.03$ ,  $P > 0.05$ ) and GC content ( $R^2 = 0.02$ ,  $P < 0.05$ ). Similarly, RD has a low positive correlation with genome size ( $R^2 = 0.04$ ,  $P > 0.05$ ) and GC content ( $R^2 = 0.02$ ,  $P < 0.05$ ) (Supplementary file 1 - Table S6 - B).

To verify the preference of dimer type in the studied groups, an analysis of preferential motifs was performed. It was possible to observe that there was variation between the different plant groups (Supplementary file 1 - Table S2 - column AH to AM). For example, on average in angiosperms and ferns, AG/GA and CT/TC dinucleotide motifs were preferred, while in algae, Marchantiophyta, moss, Lycophyta, and gymnosperms AT/TA was the most common. The CT/TC and AG/GA dimer repeats were the most common repeat motifs between the two hornwort species. The least common repeat motifs among the groups, in order of lowest occurrence, were CA and TG/GT, respectively.



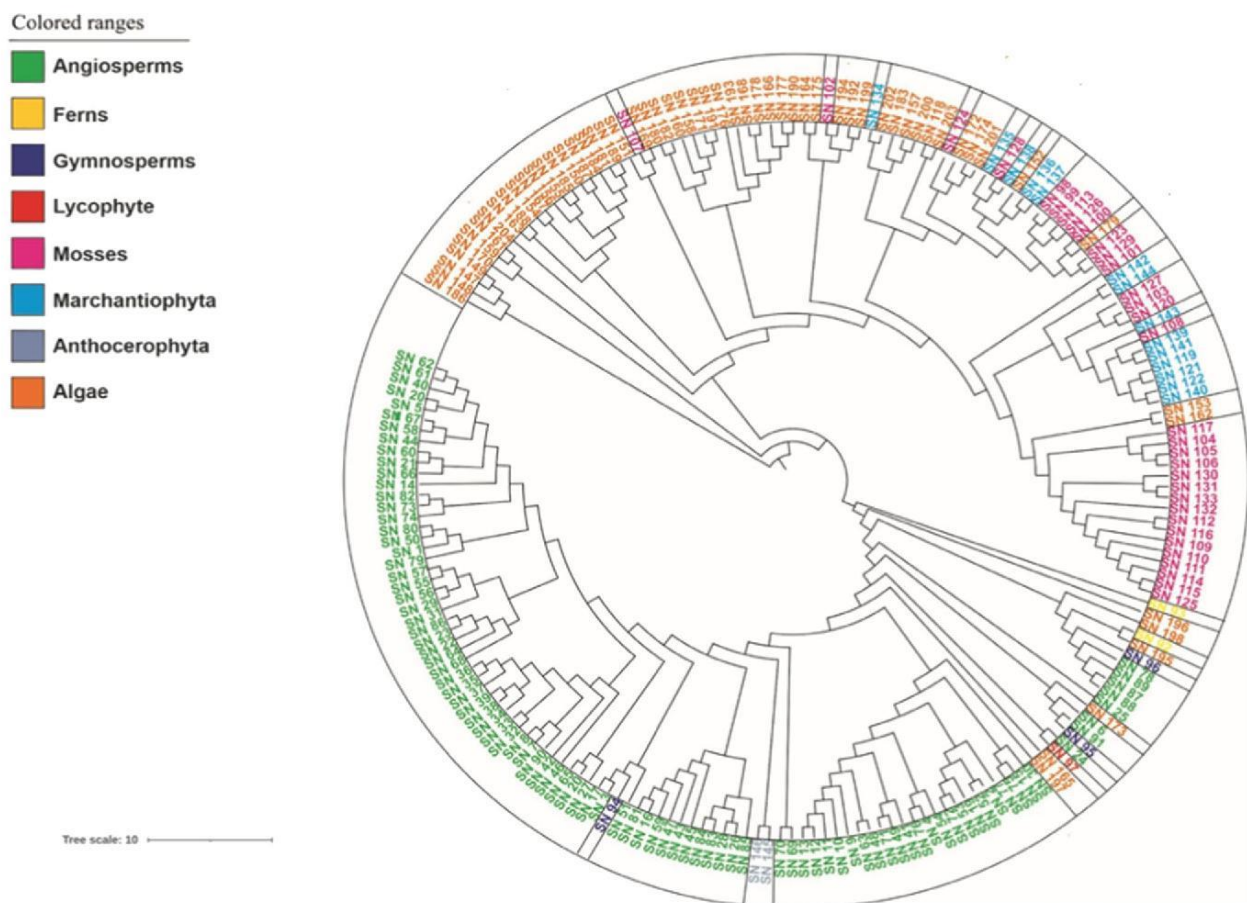
### Cluster analysis of mitochondrial SSR2-6 confirms differences between non-vascular and vascular plants

The Kullback-Leibler symmetrized divergence analysis based on the SSR2-6 percentage clearly reveals the distinction of avascular plants, except for the Anthoceroophyta that were grouped with the vascular plants (Figure 5). Also, six species of algae (SN 197, 165, 173, 195, 198, and 196) were grouped with vascular plants. The Marchantiophyta were grouped with the moss species forming two large clusters. The ferns formed a distinct clade together with Lycophyta, flowering plants, and gymnosperms, with these plants standing out from all other land plants. When compared to the profile of all repetitive DNAs found in the main strains of terrestrial and Chlorophyta plants, it is evident that the profiles show congruence with taxonomy, such as monocots and dicots together. However, some inconsistencies have been observed, such as Anthoceroophyta, Marchantiophyta, and

mosses in the same clade with *Populus* sp., as a sister clade for other angiosperms. These inconsistencies do not diverge in relation to the current understanding of the phylogenetic relationships of terrestrial plants; apparently, taxonomically more distant groups have the same RE (repetitive elements) profile, such as bryophytes and *Populus* (supplementary file 2), since the entire terrestrial plant lineage has the same ancestry. This result is very similar to the data related to SSR in the present study (Figure 5).

### SSR1 and SSR2-6 distributions in coding and non-coding regions

To understand the effect of the presence of SSRs, the location of these sequences in the coding and non-coding regions of plant and algal mtDNAs was performed. In general, SSR1 and SSR2-6 placed in the coding portion of the genome represent only 12% of the total evaluated SSRs; and therefore, most SSR repeats were distributed in the non-coding portion



**Figure 5** Cluster analysis based on dinucleotide, trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide-SSRs (SSR2-6) in the mitochondrial genome of plants and algae. Kullback-Leibler sym-

metrized divergence analysis was used (Kullback 1951). The evolutionary tree was built using the UPGMA method within the MEGA-X software (Kumar et al. 2018)

of the mitochondrial genome of plants and algae (Supplementary file 1 - Tables S1, S2, S3, and S4). In Supplementary file 3, it is possible to observe the distribution of SSRs in one species of each studied group.

Of the total SSR1 identified in all mtDNA species from each group of plants and algae, 37%, 25%, 28%, and 14% were placed in coding regions for algae, liverwort, moss, and hornwort, respectively. Also, 17%, 12%, 5%, and 9% of SSRs were placed in the coding regions of Lycophyta, gymnosperms, ferns, and flowering plants, respectively (Supplementary file 1 - Tables S3 and S4).

On average, algae have the highest percentages of SSR1 and SSR2-6 in coding regions, followed by moss and Marchantiophyta (Figure 6). The most derived plants, such as flowering plants and gymnosperms, presented higher numbers of SSRs in non-coding regions. However, species such as *Utricularia reniformis* (SN 81), *Capsicum annum* (SN 17), *Oryza minuta* (SN 54), and *Geranium maderense* (SN 30), beyond the genus *Beta* (SN 7 and 8) and *Brassica* (SN 9-13), have high SSR1 numbers in coding regions when compared to other flowering plant species (Supplementary file 1 - Table S3). Regarding SSR2, in gymnosperms and flowering plants that have the largest number of these elements, the occurrence was mainly in the non-coding regions, except for *U. reniformis* (SN 81) which has almost half of the SSR2 in the coding region. Moreover, in general, this species showed a high number of SSR1 and SSR3-6 in coding regions (Supplementary file 1 - Tables S3 and S4). Furthermore, SSR1, followed by SSR2 and SSR3, is the preferred motif in the coding regions of the mtDNAs of plants and algae. However, some species of flowering plants, gymnosperms, ferns, and algae presented more SSR2 than SSR1 in coding regions (Figure 6 and Supplementary file 1 - Tables S3 and S4).

Supplementary file 5 presents the structure of five mtDNA from the algal group, three being from phylum Chlorophyta (*Chlorella variabilis*, *Microsporum stagnorum*, *Dunaliella*

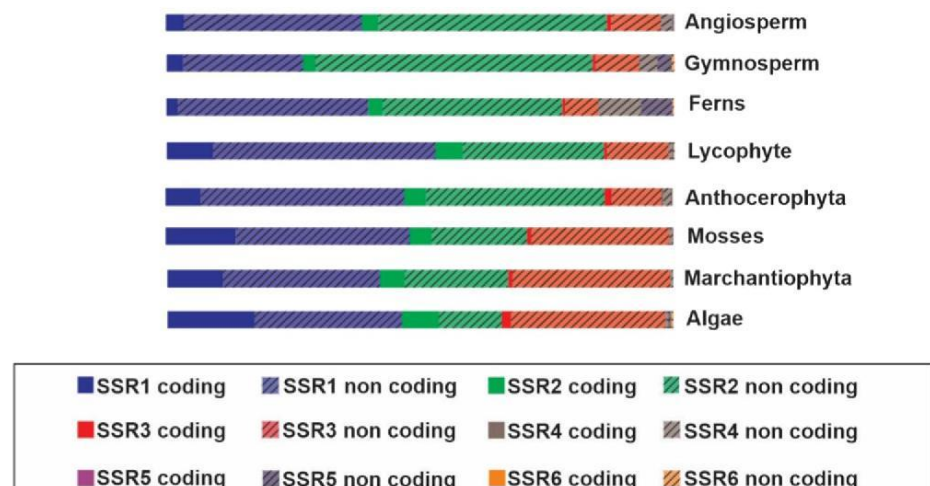
*viridis*) and two from Charophyta (*Roya obtusa* and *Nytella hyaline*), and shows the structural differences in mtDNA between the two phyla analyzed beyond the structural differences in the same phylum of Chlorophyta. The numbers of genes, introns, and genome size were drastically different among species, and reflect differences in SSR distribution, number of SSRs, density, long motifs, overrepresentation of SSR1, and AT content.

### Features of cSSRs in mitochondrial genomes of plants and algae

A total of 26,538 composite simple sequence repeats (cSSRs), also known as composite microsatellites, were identified in all analyzed mtDNAs (Supplementary file 1 - Table S1). Most cSSRs were composed of “2-” or “3- different motifs” (data from IMEX software, not shown). Apart from these, no cSSR expansion was observed in mtDNA among species.

The angiosperm group presents the highest number of cSSRs mainly in large mtDNAs as that observed for *Corchorus capsularis* (SN 22). It was confirmed through correlation analysis since cSSRs were positively correlated with genome size ( $R^2 = 0.67$ ,  $P < 0.05$ ) and low positively correlated with GC content ( $R^2 = 0.06$ ,  $P < 0.05$ ) (Supplementary file 1 - Table S6 - C). cSSR RA was more pronounced in mosses and some Marchantiophyta possibly due to the greater presence of SSR in coding regions (Supplementary file 1 - Tables S3 and S4). As observed for SSR1 and SSR2-6, algae presented greater variability in cSSR number and RA. The highest cSSR RD was observed in some algae and in *Brassica nigra* (SN 12), representing in these species the longest cSSR sequence lengths. cSSR RA was poorly correlated with GC content and poorly correlated with genome size. cSSR RD showed low correlation with genome size,

**Figure 6** SSR distribution based on percentage in the coding and non-coding regions of the mitochondrial genome of the eight taxonomic groups of plants and algae. SSR1, mononucleotide; SSR2, dinucleotide; SSR3, trinucleotide; SSR4, tetranucleotide; SSR5, pentanucleotide; SSR6, hexanucleotide





and was poorly correlated with GC content (Supplementary file 1 - Table S6 - C).

### Long SSRs in the mitochondrial genome of plants and algae

Of the 204 mitochondrial genomes analyzed from plant and algal species, 111 showed SSR1 that repeat 13 times or more ( $\text{SSR1} \geq 13$  nt) (see “Material and methods”, similar values from long SSR are used (George et al. 2015)). These comprehend 60 species of flowering plants; five species from gymnosperms, ferns, and Lycopphyta; 19 moss species; 9 Marchantiophyta; two Anthoceroophyta; and 16 algae species (Supplementary file 1 - Table S3). Of the total number of long SSRs, 2.48% were placed within exons, 1.49% were within open reading frames (ORFs), and 9.14% were within intronic regions (Supplementary file 1 - Table S5). The total number of  $\text{SSR1} \geq 13$  nt ranged from 1 (47 species) to 18 (*Liriodendron tulipifera* (SN 46)). The longest SSR1 was found in the hornwort *Phaeoceros laevis* (SN 145) which presented 37 nt containing poly (G). Of the total long SSR1, 35 were located in coding regions, distributed in 31 species between the groups (Supplementary file 1 - Tables S1 and S3). The largest SSR1 located in the coding region had 19 nt, composed of poly (G), found in three genomes: *Phoenix dactylifera* (SN 58), *Treubia lacunosa* (SN 144), and *Phaeoceros laevis* (SN 145). Algae, ferns, and gymnosperms do not have long SSR1 in their coding region. Mosses were the group of plants with the longest SSR1 in the coding region, with 15 species presenting these elements. Mosses and Marchantiophyta were also the groups showing greater occurrence of  $\text{SSR2} > 10$  bp in the mtDNA, among which 10 long SSR2 were located within introns, being *Calypogeia suecica* (SN 138) and *Calypogeia neogaea* (SN 137) the ones showing the longer repeated motifs, with the AT/TA dimer repeated 29 times (Supplementary file 1 - Table S5).

Regarding the SSR3 type, long motifs were found in small numbers when compared to SSR1 and SSR2 types and the largest number occurred in angiosperm species. The longest SSR3 was identified in *Tripsacum dactyloides* species (SN 78), which presented a 40-fold repeated CTA motif located 22,129 nt upstream of *rrn26* gene sequence. Two species of flowering plants had the longer trimer motifs in the region containing an ORF of the putative protein YP\_004927816. The species *Welwitschia mirabilis* (SN 96) from the gymnosperm group also had one of the longest SSR3, a CGA motif repeated 18 times. Similarly, the AAG motif was repeated 16 times in *Closterium bailyanum* algae (SN 161), and was located in an ORF region of the putative reverse transcriptase-maturase coding gene.

Long SSR4, SSR5, and SSR6 were present in large numbers in the mtDNAs of gymnosperm and fern species. Long SSR4 were more frequent in ferns, and *Psilotum nudum* (SN

93) had 38 long SSR4 motifs. However, the longest SSR4 was found in the Anthoceroophyta *Nothoceros aenigmaticus* (SN 146), the TATG motif which was repeated 31 times. Many SSR4 are part of the coding region; furthermore, long SSR4 were distributed in the intronic region of the species *Chlorokybus atmophyticus* (SN 159), *Megaceros aenigmaticus* (SN 146), *Calypogeia suecica* (SN 138), *Ginkgo biloba* (SN 95), and *Psilotum nudum* (SN 93) (Supplementary file 1 - Table S5). Large numbers of long SSR5 were observed in angiosperms, but mainly in gymnosperms and ferns, and the largest reason was found in *Chlorokybus atmophyticus* (SN 159) algal species (ATGCA = 39 upstream to *cox1* gene). Many of the long SSR5 were distributed in intron and exon regions.

The Anthoceroophyta species and Lycopphyta did not show long SSR5. Long SSR5 were more abundant than any other type in the mtDNA of plants and algae. Long SSR6 were found mainly in flowering plant species. The longest SSR6 repeat was found in the *Tripsacum dactyloides* (SN 78) with the ATAAGA motif that repeated 36 times. The mosses, Marchantiophyta, and Lycopphyta showed no long SSR6 (Supplementary file 1 - Table S5). In the coding region, the largest expansion of SSR6 was 7 repeat units long.

### Distribution of SSR7-10 and the long SSR7-10

In general, SSR7-10 (SSR motifs from seven nucleotides to ten repeated more than three times) occur in all plant groups (Supplementary file 1 - Table S2 and S5) but large SSRs are facilitated in larger genomes such as gymnosperms and angiosperms. Some of the SSR7-10 occur within ORFs (1.21%), exons (3%), and important non-coding regions such as introns (7.8%) in plant and algal mtDNA. Heptanucleotides (SSR7s- motif composed of seven nucleotides repeated more than three times) were identified in large numbers in spermatophyte mtDNA, including *Cucumis sativus* (SN 24) species that had 11 heptamers in its mtDNA. *Zea perennis* (SN 89) presented the longest SSR7, the SSR TATAGT A repeated 12 times. Octanucleotides (SSR8s) were more frequent in flowering plant and algal mtDNA, and two species showed SSR8s in intronic regions (*Dunaliella salina* (SN 164) with AATGTTAT = 3 times and *Dunaliella viridis* (165) with TAACTACC = 5 times). Two long SSR8s were identified in *Ulva linza* (SN 201), GAGAAAGC = 24 times, and GAGAAAGC = 12 times. Ferns were identified as having the greatest abundance of nonanucleotides (SSR9s), with the species *Ophioglossum californicum* (SN 92) presenting 36 nonanucleotides, with the predominance of SSR GGAGGAGTT in its mtDNA. The longest SSR9 was observed in the *Lobosphaera incisa* (SN 170), GAG GGCTAC = 13 times. Five SSR9 (motif composed of eight nucleotides repeated more than three times) were found within genes, including *rps10* and *atp1* in algal genomes,



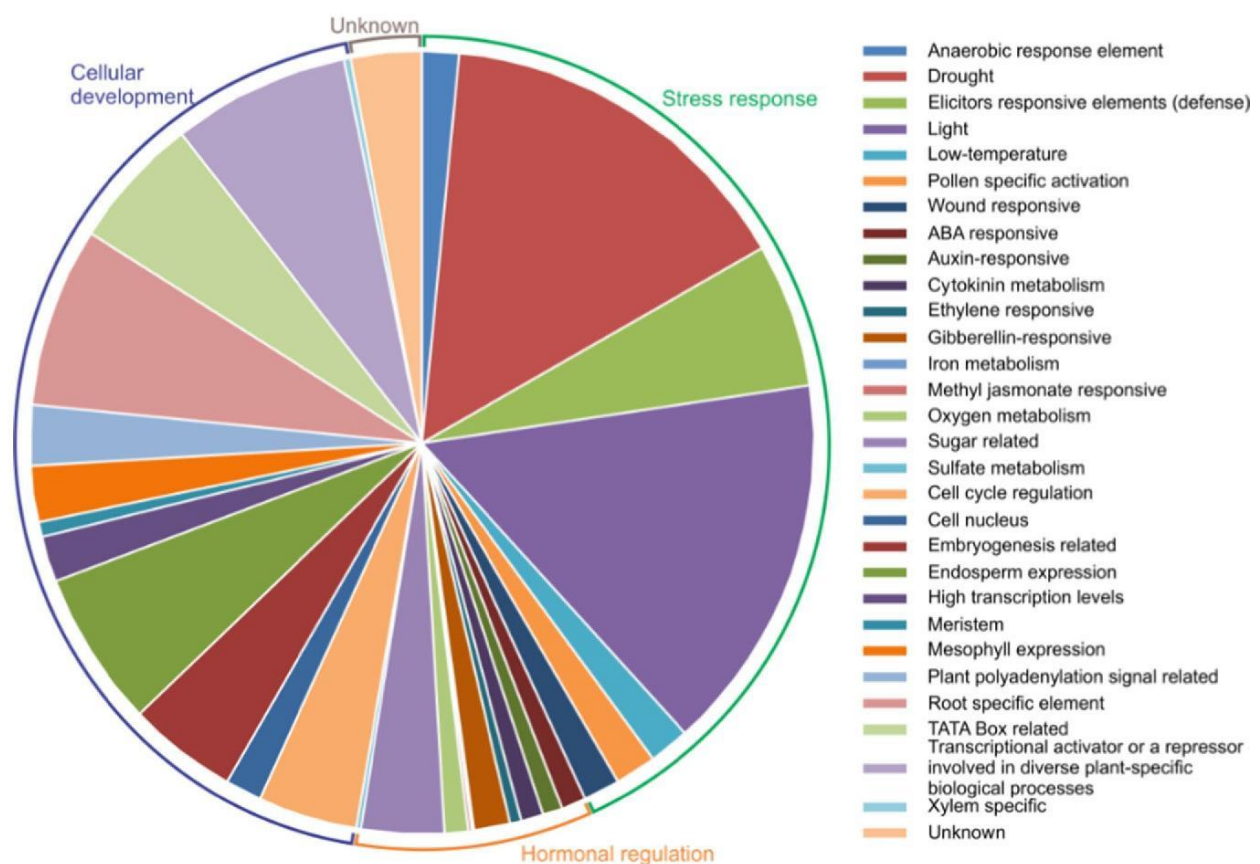
and *rtl* in Marchantiophyta. The other two nonamers were found in intron regions of *Ophioglossum californicum* (SN 92). An increase in the number of decanucleotides (SSR10) was observed in algae as well as a long decamer has been found in *Chlorokybus atmophyticus* (SN 159) (ACTGCA GTAA = 9 times). Some long SSR10s (motif composed of ten nucleotides repeated more than three times) were found within introns, exons, and ORFs in algae. Mosses and Marchantiophyta had almost no SSR7-10 in mtDNA, and further expansion in the coding region of these groups occurs as SSR5s (Supplementary file 1 - Table S4). In algae, the largest expansion occurred in a SSR9 (motif composed of nine nucleotides repeated more than three times) motif, repeated 3 times (Supplementary file 1 - Table S4).

### *cis* regulatory element (CRE) analysis

From the identification of mtSSRs in genomes analyzed in this study, we were able to verify the possible location of these sequences in regulatory regions of genes. SSR3-6 showed large expansion upstream of mitochondrial genes,

near promoter regions (Supplementary file 1 - Table S5). However, only a few conserved SSR3-6 were observed upstream to homologous and orthologous genes among the studied species. After identification of regulatory regions containing SSRs, the possible presence and function of *cis*-regulatory elements (CREs), which play an important role in the regulation of gene expression, were analyzed. We identified 537 CREs distributed in the promoter regions containing SSR<sup>3-6</sup> in the mitochondrial genes that were selected in this study. The metabolic pathways in which CREs act were grouped into different functional categories (Figure 7). CREs, in either the (−) or (+) position, that are known as responsive to cellular components were the most common among plant and algal groups and had SSRs in promoter regions, followed by those CREs involved in stress and hormonal regulation. Light-related CREs were found in greater number. The identified CREs ranged from 2 to 12 bp in size (data not shown).

The CREs that are part of SSRs in the promoter regions of the genes chosen for analysis are described in Table 1. Through the analysis of the presence of SSR3-6



**Figure 7** Graph presenting the *cis* regulatory elements found in promoter regions of plant genes and algae containing trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide- SSRs (SSR3-6) in those regions

**Table 1** Table of *cis* regulatory elements (CREs) that are part of trinucleotide, tetranucleotide, pentanucleotide, and hexanucleotide- SSRs (SSR3-6) in putative plant and algal gene promoters

Functional category	CREs	Sequence	Occurrence	Function
Stress response	ANAERO2CONSENSUS	AGCAGC	AlgaeTRI	Anaerobic responsive
Cell development	RYREPEATVFLEB4	CATGCATG	AlgaeTETRA	Endosperm
Cell development	RYREPEATLEGUMINBOX	CATGCAY	AlgaeTETRA	Endosperm
Stress response	REALPHALGLHCB21	AACCAA	AlgaePENTA	Light regulated
Stress response	MYB1AT	WAACCA	AlgaePENTA	Dehydration responsive
Hormonal regulation	CAREOSREP1	CAACTC	AlgaeHEXA	Gibberellin responsive
Stress response	POLLEN1LELAT52	AGAAA	AngTRI	Pollen specific
Cell development	CAATBOX1	CAAT	AngPENTA	Common <i>cis</i> -acting element in promoter
Cell development	ROOTMOTIFTAPOX1	ATATT	AngHEXA	Root specific
Cell development	CACTFTPPCA1	YACT	AngHEXA	Mesophyll expression
Cell development	NODCON1GM	AAAGAT	AngHEXA	Root specific
Cell development	OSE1ROOTNODULE	AAAGAT	AngHEXA	Root specific
Cell development	DOFCOREZM	AAAG	AngHEXA	Transcriptional activator or a repressor
Stress response	GT1CONSENSUS	GRWAAW	AngHEXA	Light
Stress response	IBOXCORE	GATAAG	AngHEXA	Light
Stress response	GATABOX	GATA	AngHEXA	Light
Cell development	MARTBOX	TTWTWTTWTT	FernTRI	Scaffold attachment region
Cell development	POLASIG1	AATAAA	FernTRI	Plant polyadenylation signal
Cell development	TATA BOX5	TTATTT	FernTETRA	Tata Box region
Cell development	MARA BOX1	AATAAAYAAA	FernTETRA	Cell nucleus
Cell development	POLASIG1	AATAAA	FernTETRA	Plant polyadenylation signal
Stress response	GATA BOX	GATA	FernHEXA	Light
Cell development	ARR1AT	NGATT	BryoPENTA	Transcriptional activator or a repressor

*Algae*, algae; *Ang*, angiosperm; *Fern*, ferns; *Bryo*, bryophytes

in CREs, it was possible to observe that many of these elements are involved in fundamental processes such as cell development, stress responses, and hormonal signaling. In addition, they are part of the binding site of transcription factors such as DOFs (DOFCOREZM), bZIPs (RYREPEATVFLEB4), and MYBs (MYB1AT) that play an important role in plant growth and development, and in response to biotic and abiotic stresses.

Many of the SSR3-6 identified in the promoters can possibly act as regulatory elements (Supplementary file 4A) while others contain regulatory elements (Supplementary file 4B). Supplementary file 4A shows the SSR SSC trimer, which can act as a regulatory element located in the *mttB* (*SecY-independent transporter*) gene promoter; its product plays a role in intracellular transmembrane transport mediating protein transport. This CRE called ANAERO2CONSENSUS/AGCAGC responds to anaerobic conditions. In Supplementary file 4B, a TCT TTA hexamer repeat SSR region can be seen which contains several CREs at the (–) position of the *rpl16* (*ribosomal protein L16*) promoter region.

## Discussion

Here, we present a complete comparative diagnosis of mtSSRs in different taxonomic groups of plants and algae. The data presented show the large number of SSRs existing in mtDNAs of plants and algae with potential involvement in increasing genotypic and phenotypic diversity among the species.

### SSR1 and SSR2 features demonstrate bias in plant groups

In this study, we showed that the number of mtSSRs is directly related to genome size as also observed before (Rajendrakumar et al. 2008); however, the groups of plants and algae demonstrated some bias for specific types of SSRs (Figure 4) as already described (Kuntal and Sharma 2011). We demonstrated that SSR1s were the most abundant (RA) type in non-vascular plants and



algae, Lycophyta, and ferns as indicated in Figures 1 and 2, and Supplementary file 1 - Tables S1 and S3. Higher abundance of SSR1 was previously reported in mtDNAs of some plant and algal species (Kuntal and Sharma 2011). SSR1s have as profile occurrence mainly in non-coding regions (Supplementary file 1 - Table S1) (Kuntal and Sharma 2011). Overrepresentation of SSR1 occurred in most mtDNAs, mainly in mosses and Marchantiophyta, and in some algal species (Supplementary file 1 - Table S1). It is suggested that the frequent increase and deterioration of SSR1 may occur in the plastid genomes of above taxa, as they have a high replacement rate (George et al. 2015; Ceplitis et al. 2005; Jakobsson et al. 2007). An overrepresentation of SSR1 in mtDNAs may occur as in the plastid genome, although the variation in the replacement rate of mt SSR1 is unknown. Bias was also observed for moss, Marchantiophyta, and Anthoceroophyta in relation to the high SSR1 overrepresentation, reflecting also high RA and high density (RD) in coding regions (Figure 6 and Supplementary file 1 - Table S3). SSR1s also have larger sizes (>8 nt) in these small genomes; for example in these three plant groups cited above, the longest SSR1 for all species was found in *Phaeoceros laevis* (Supplementary file 1 - Table S1). It has already been observed in plastid genome of these species (George et al. 2015). SSR1s that are in coding regions may also be under some selection pressure (Kuntal and Sharma 2011), and are useful as molecular markers. mtDNAs of flowering plants have a significant frequency of long SSR1 possibly in non-coding regions, mainly in Malvaceae species (Supplementary file 1 - Table S1). This is in line with previous studies that have shown that longer motifs in *Gossypium* species are common and usually composed of A/T (Wang et al. 2015). Thus, despite their abundance and the fact that long SSR1s have been identified in mtDNAs, their mutation rate is a cause of concern, although their use as molecular markers has already been reported (Rajendrakumar et al. 2008). In spermatophytes, SSR2 was the most abundant type (RA) of SSRs, as observed in the mtDNA of most analyzed grasses (Figures 3, 4, and 6 and Supplementary file 1 - Table S2) (Filiz 2014). This type also is abundant in coding regions (Filiz 2014) (Supplementary file 1 - Table S4), mainly *Brassica*. In some species of spermatophytes, it was frequent in large intergenic (Victoria et al. 2011) as well as non-coding regions such as introns (Srivastava et al. 2019). This is in accordance with the correlation between the size of mtDNA and the abundance of dinucleotides (Rajendrakumar et al. 2008). Moreover, this type of SSR is characterized by high mutation rates, contributing to its high density (RD of SSR2-6 - Supplementary file 1 - Table S1) in non-coding regions in spermatophytes, facilitated by low pressure selection mainly in regions of introns and intergenic regions. In addition,

this low pressure facilitates the occurrence of other long motifs, such as observed in *Tripsacum dactyloides*, that presented the largest trinucleotide repeat (Supplementary file 1 - Table S5).

The possible reasons that certain types of motifs are more abundant in certain groups of plants can be explained by a set of genomic events and preference/selectivity of certain motifs that occurred during evolution, e.g., rearrangements, gene gains/losses, and number of introns (Cole et al. 2018). Preference for certain repeats composed of A/T or C/G was also cited (Tian et al. 2011; Qin et al. 2015). For SSR1, Toth et al. (Sousa et al. 2020) suggested that the poly(A) tails of densely scattered retroposed sequences and processed pseudogenes are responsible for this higher proportion of A/T-rich repeats, which may be the evolutionary driver of A/T mononucleotide SSRs. For *Karlin and Burge* (Karlin and Burge 1995), SSR2 abundance appears to reflect the species-specific properties of DNA modification, replication, and repair mechanisms and Da Maia et al. (Maia et al. 2009) complement that this bias is common in expressed regions of grass genes. For use, markers based on SSR2 were more polymorphic than those based on SSR3 as cited by Oliveira et al. (Oliveira et al. 2006).

Another relevant issue is that most mtDNAs are distinct from plastid genomes (George et al. 2015), since abundance does not decrease with increasing SSR motif size (Figure 4). This evidence became clear also when observed SSR3 that were more abundant than SSR2 in moss, Marchantiophyta, and algae; SSR4 and SSR5 were more abundant than SSR3 in ferns and gymnosperms. This profile marks a distribution bias of mtSSR in plant analyzed species.

A/T content is an important genomic trait because it is a factor in determining the frequency of SSRs in genomes; the percentages of AT motifs in SSRs increase with increasing genome AT content, GC content being therefore not correlated with SSRs (Tian et al. 2011; Kalia et al. 2011; Morgante et al. 2002) (Supplementary file 1 - Table S6A and B). In general, SSR1 containing A/T was more frequent in coding and non-coding regions of the analyzed species, except for the ferns and gymnosperms, which presented higher G/C frequency. High G/C content has also been observed in plastid and nuclear genomes of ferns and gymnosperms (George et al. 2015; Kuntal and Sharma 2011; Kalia et al. 2011) (Supplementary file 1 - Table S2 and Figure 1). AT/TA was the most common type of dimer among gymnosperms, Lycophyta, and non-vascular plants, although the preference for this type of motif has changed to other motifs such as CT/TC and AG/GA in Anthoceroophyta, ferns, and flowering plants. The same was reported in mtDNAs (Kuntal and Sharma 2011), but not in the nuclear genome (Qin et al. 2015). AG/TC is the most common type of SSR2 in mtDNA of mono- and dicotyledons (Kuntal and Sharma 2011; Victoria et al. 2011). GC/CG was uncommon in plant mtDNA and



less frequent in algae, presenting the same profile in plastid genomes (George et al. 2015). However, in nuclear genomes, it occurs quite often, especially in non-vascular plant species (Victoria et al. 2011).

### Can mitochondrial SSR features reflect the evolution of organisms from algae to land plants?

The distribution of SSRs could be tracked according to the plant taxonomic group, and reflects the several ancient divergences of land plant evolution (Figures 1, 2, 3, 4, 5, and 6 and Supplementary file 1). Each analyzed plant group has a non-random distribution and particular traits on the mtDNA structure and the number and distribution of SSRs reflect these features. This bias was described by Song et al. (Song et al. 2021) in gene-coding sequences. Despite being a cytoplasmic genome such as the chloroplast genome, mitochondrial genome of plants and algae sometimes differs in many features from chloroplast regarding SSR distribution (George et al. 2015).

Initially addressing to the algae, among this group of species, few conserved features and great variation from high to low number, abundance, and density of SSR that do not occur in other plant species are observed in Figures 2 and 3. It may be a result of differences in the mtDNA and in selection pressure on SSR sequences. Algae have special features with respect to the other taxonomic groups: (1) abundance and variation of SSR1 (Figure 2) but few long SSR1 $\geq$  13 and species showing no overrepresentation (Supplementary file 1- Tables S1 to S4); (2) greater cSSR and SSR2-6 variation (Figure 3) and some species showing large SSR5 repetitions; (3) algae have large repeats of SSR10s (supplementary file 1- Table S6), being curious since there is a tendency, as the smaller the motif size the higher its frequency (George et al. 2015); (4) large number of SSR in the coding portion (Figure 6). These large variations among species reflect the lack of distribution pattern and the clustering observed in Figure 5. That fact can be attributed to the distinct evolutionary pattern of mtDNA, not only between the Chlorophyta and Streptophyta, but also the species within the Chlorophyta. One of the evolutionary patterns shows genomes with ancient characteristics (Gray et al. 1999) that have kept clear traces of eubacterial ancestry, and the derived pattern (Pombert et al. 2006) has been attributed to Chlorophyta mtDNAs that radically deviate from the ancestral pattern, with little or no evidence of their traits. Such patterns imply radical changes in structure, gene content, gene organization, amount of introns, and variation in genome size, all these locations and their variations determining the frequency in coding and non-coding regions, abundance, and size of the SSR. This evidence was clearly detected in our study because features and genome organization of mtSSR of Charophyta not only are different from those of the other phylum

Chlorophyta, but also these differences extend among Chlorophyta species (Supplementary file 5), for example, *Dunaliella viridis* that have high G/C content even comparable to gymnosperms besides high dinucleotide number, and the same occurs in *Microspora stagnorum*; this aspect being completely different from other charophyte species being positioned with the spermatophytes. In chloroplast, differences in SSR distribution were reported in green algae, red algae, and apicomplexans (George et al. 2015). Algae have extensive regulatory gene flexibility that allows this basal group to inhabit distinct environments and survive fluctuations in nutrient availability (Grossman et al. 2007). In addition, there is still the bias factor already widely discussed in repetitive elements in several other organisms, such as fish and amphibians (Cabañas et al. 2020; Wang et al. 2019) as well as in plants that show very similar characters and ongoing low recombination in mtDNAs in all aspects (Dong et al. 2019). These peculiarities may be reflecting the clustering distribution patterns between algae (such as *Dunaliella* and *Microspora*) or Anthoceroophyta and angiosperms. It may explain the difference in the features of SSRs presented by algae, showing a great dynamism of SSRs, which may play an important role in the existence and evolution of these organisms.

The different types of SSRs were conserved in the bryophyte group in genome size, number of SSRs, and consequently RA and RD (Figures 1, 2, 3, 4, and 5 and Supplementary file 1). A few SSR expansions were observed, except for SSR1 in coding regions (Supplementary file 1- Table S3). Also, few expansions were observed in non-coding regions, as observed with long SSR2 repeats in intron region of *Calypogeia* species (Supplementary file 1- Table S5). This complex conservation can be a reflex of the preferential and selective inclusion of SSRs during evolution, i.e., occurred due to the evolutionary advantage that they confer to mtDNA. Conserved profile characterizes the bryophyte species and is different from the events in spermatophytes (Zhao et al. 2016; Liu et al. 2016). This can be explained by the fact that bryophytes still retain many of the structural and molecular traits that made possible the origin of land plants, which during the transition process underwent significant changes in structure and function. All of these adaptations were accompanied by changes in genome sequence (Rensing et al. 2007) that remained during evolution.

Another striking point is related to similar distribution and abundance of SSR1, SSR2-6, cSSR, and SSR7-10 (Figures 2, 3, 4, 5, and 6) in mosses and Marchantiophyta, only that corroborates the Setaphyta monophyly (Sousa et al. 2020). Among these features, few long SSR2-6 repeats were observed between these two groups, except few dinucleotides (Supplementary file 1 – Table S5). The implications of excess short repetitions (<8 repeated units) are extremely



important for genetic stability and the evolution of additional genomic traits such as codon usage. The bryophyte lato sensu group is currently divided into two clades, divisions in which the Anthocerophyta are the sister group of Setaphyta clade (mosses + Marchantiophyta). Meanwhile, species of Anthocerophyta differ in their SSR traits, even being more similar to angiosperms as can be seen in the phylogenetic analysis (Figure 5). This may reflect the conservation of SSRs during evolution, keeping them under constant selective pressure, as they possibly play a fundamental role in the efficient adaptive evolution of these two groups of plants. This agrees with the single event origin hypothesis for all land plant lineages (Zhong et al. 2015; Vries et al. 2016; Bowles et al. 2020). The issue of ancestry is still very controversial under a phylogeny based on tree-thinking approaches; we may even speculate that from the philosophical point of view, *Arabidopsis* may be more ancestral than Marchantiophyta, considering the greater similarities between the genomes of these flowering plants and the Anthocerophyta (Rich and Delaux 2020). In addition, unlike Anthocerophyta, mosses and Marchantiophyta have not been recombined since the Devonian Period (Xue et al. 2010; Liu et al. 2011; Field and Wills 1998), which may explain the similar distribution of SSRs. An evolutionary gap has already been pointed out between the Marchantiophyta/moss relative to the Anthocerophyta. This can be explained by the wide extinctions between Anthocerophyta and moss during evolution (Zhao et al. 2016). On the other hand, if we interpret the Setaphyta clade as monophyletic (Sousa et al. 2020), this gap is a new argument held in favor to the hypothesis of Anthocerophyta as a sister group for all other bryophytes. However, the scarcity of genomic data representing Anthocerophyta and Marchantiophyta makes it difficult to further assess this issue. Repetitive DNA analyzes joint mosses to groups of land plants that are not taxonomically close. *Physcomitrella patens* and *Populus* have greater similarity in the distribution of their RE than that of Anthocerophyta and Marchantiophyta (Supplementary file 2). This could be explained by the higher occurrence of polyploidy in mosses, especially in the Funariaceae family, where *Physcomitrium* is classified (Medina et al. 2018), polyploidy being an ancient event in poplar (Sterck et al. 2005). However, it is necessary to analyze more genomes to have an accurate answer on this topic. In addition, introns are sites known for their high SSR density (Zhang et al. 2006). The intron distribution between the bryophyte groups is a specific trait of each of the three strains, forming an individual set of introns that is not shared among the strains. Loss of three mitochondrial introns ensures Marchantiophyta monophyletic origin as the sister group of embryophytes, which consistently occurs in other strains (Knoop 2010). Thus, it is possible that different evolutionary mechanisms act on the Anthocerophyta mtDNA and may be responsible

for distinguishing mitochondrial SSR features between the Anthocerophyta and the other bryophyte strains.

Furthermore, a separation between nonvascular and vascular plants is seen, that is why in ferns and their sister group Lycopphyta, the distribution of SSR1, SSR2-6, SSR7-10, and cSSR (Figures 2, 3, 4, and 5) is similar to spermatophytes, and the highest abundance of SSR5, and long SSR4, SSR5, and SSR6 are observed only in the first group including gymnosperms (Supplementary file 1- Tables S1 to S4). DNA polymerase slippage during the replication process is the major factor for SSR sequence expansion (Levinson and Gutman 1987). As already observed (Qin et al. 2015), SSR sequence expansion needs further study, as the different features of SSRs may reflect the different fidelity of DNA polymerase between ferns and the other groups. On the other hand, mtDNA of ferns has abundant expansions of SSR4, SSR5, and SSR6, mainly in introns (Supplementary file 1- Table S5), and which are shared with the gymnosperm species *Welwitschia mirabilis*. The large number of mitochondrial introns in ferns (Guo et al. 2017), and the mixed structural features that ferns share with Lycopphyta and gymnosperms (Guo et al. 2016) make them a transition group. In this context, ferns contributed evolutionarily to the structural diversity observed in gymnosperms and angiosperms, also facilitating the abundance and expansions of SSR2-6 in mtDNA as observed in our analysis (Guo et al. 2016, 2017). According to Toth et al. (Toth et al. 2000), strand-slippage theories alone cannot explain SSR distribution in the genome as a whole; enzymes and other proteins involved in various aspects of DNA-processing (i.e., replication and repair) and chromatin remodeling may be responsible for the taxon specificity of SSR abundance. And Harr et al. (Harr et al. 2002) comment that the mismatch repair system may have an important role in shaping genome composition. On the other hand, in this study, it was possible to observe bias in SSR distribution in all plant groups, but some other additional features (bias in SSR4, SSR5, SSR6 and the great introns that accumulate long SSRs — result of numerous intron gains due to recombinations (Hecht et al. 2011; Wynn and Christensen 2019; Kozik et al. 2019) — added to our clustering analysis) can give clues that rearrangement occurred impacting mtSSR mainly between ferns and gymnosperms as commented by Cole et al. (Cole et al. 2018). Evidence of genome rearrangement can be also observed among closely related species of angiosperm as *Beta*, *Zea*, and *Silene* as commented by Cole et al. (2018) but it does not seem to have impacted SSR distribution as observed in our agrupament analysis. The number of SSRs is correlated with genome size while RA and RD were not correlated possibly because specific regions of mtDNA as introns and some non-coding regions present more SSRs. On the other hand, as occurred with SSR1 in mosses, the highest RA and RD occurred possibly to the greatest number



of mitochondrial genes and to a preferential and selective inclusion in these genes.

### Non-vascular plants show tendency to adaptive evolution verified through the cSSRs

The largest variation in cSSR abundance occurred in algae, and species representing non-vascular plants showed the highest cSSR abundance values (Figure 3 and Supplementary file 1- Table S1). A similar trend was observed in plastid genome of basal plants (George et al. 2015), which showed the adaptive evolution possibly due to the reduced number of genes involved in DNA repair, since the absence of DNA polymerase repair may increase the rate of change in SSRs. cSSRs are believed to originate from imperfections following eukaryotic SSRs (Chen et al. 2011). However, the repair activity as well as the genes involved in the repair process seems to have gradually increased when we look at the backbone of plant evolution with respect to the number of cSSR (Supplementary file 1 - Table S1). Our data also indicate that the RA of cSSR is greater where the RA of SSR1 is higher, indicating that many cSSRs may be composed of SSR1 as observed mainly in mosses (Supplementary file 1- Table S1). In addition, cSSRs were the least abundant SSR type among all plant groups when compared to SSR2-6 and SSR1. This seems to be a trend in cytoplasmic genomes as it has also been observed in plastids (George et al. 2015).

### Presence of SSRs in coding regions is related to plant and algal evolution

The understanding of SSR distribution in coding and non-coding regions can help elucidate the possible role of SSRs in gene regulation and genome organization. Also, abundant markers for genetic, genomic, and evolutionary studies can be obtained. The majority of SSRs detected here show preference for the non-coding fraction of embryophyte genomes (Figure 6); despite the fact many algae present a high number of SSR1 and SSRs2-6 in the coding region, it is possible that these SSRs are related to CDS and RNA processing (Zhao et al. 2014). Thus, there is a relationship between the presence of SSRs in the coding region of mtDNAs and plant evolution; the more recent the evolution of a given plant, the lower the percentage of SSRs in the coding region of mtDNA (Figure 6). This fact can be attributed to the negative selection that occurred during the plant evolutionary process against reading frame mutations in coding regions. In addition, the decay of SSRs in the coding region occurs due to natural selection (Gao et al. 2013). Despite this, SSRs within genes may be subjected to severe restrictive selection on plant mtDNAs, possibly due to the potential harm of mutations (Metzgar et al. 2000) causing changes in sequence of important cell respiration chain genes. Significant

differences in the occurrence of mitochondrial SSRs also exist even among closely related species mainly in algal and vascular species, suggesting that the abundance of SSR1 and SSR2-6 may change relatively fast during evolution (Supplementary file 1- Table S1). Still, in the coding region of mtDNAs, the higher frequency is due mainly by SSR1, followed by SSR2 and SSR3, although there are exceptions as noted in *Brassica* and *Beta* genera (Supplementary file 1 -table S4). This pattern was previously reported in the plastid genome (George et al. 2015), indicating that accessory genomes are rich in SSR1 and SSR2 even in the coding region, unlike the nuclear genome where motifs multiple of three in the coding region are more frequent (Li et al. 2004). In this study, regardless of the repetition class, long SSRs were also present in the coding region of mtDNAs (Supplementary file 1 -Table S5), in which one trimer was the longest SSR observed. Many genes, such as *NADH* genes, are associated with cellular respiration, and may be under strong selective pressure. Modifications in mtDNA gene structure may lead to important plant changes, as reported in rice (Mignouna et al. 1987). The presence of many SSRs in ORFs, especially in algae, shows how this group reacts to fluctuations in the environment and its high adaptation dynamics (Li et al. 2004). It also shows that SSRs in ORFs act as a molecular device for adaptation to environmental stresses.

### Minisatellites are frequent in important mtDNA regions

In mtDNA, SSR7-10s are present in all plant groups, but occurrences of these types are widely variable within and among families. Many SSR7-10s are present in exons, introns, and ORFs (Supplementary file 1- Table S5), possibly being that these larger motifs play an important role in the mtDNAs of the studied species mainly in generation of polymorphism as reported by Merritt et al. (Merritt et al. 2015). In plastid genomes, SSR7-10s are observed only in algae and are not present in coding regions (George et al. 2015). The possible functions of these motifs are still unknown, but taking into account the importance of these regions for polymorphism generation, it is suggested to use them for the study of genetic diversity. Despite the reported decrease in efficiency of DNA polymerase in PCR analysis as it is considerably large, it can be compensated by the better discrimination power of the alleles and greater accuracy in the genotypic identification of individuals (Kocaman et al. 2020).

Long SSR7-10 > 5 observed only in algae, except for the 12-fold occurrence of an angiosperm heptamer, shows that selection pressure operates against long SSR7-10 types. This effect was previously reported by our group (Maia et al. 2009) in monocot and dicot EST regions but appears



to occur in the mtDNA of a wide range of plant species. SSR7-10 is the less abundant SSR type in mtDNAs and its occurrence does not appear to have a clear pattern.

### SSRs can act as *cis* regulatory elements and can affect gene expression

*cis* regulatory elements (CREs) are essential regulatory units required for gene transcription and help coordinate environmental stimulus being associated with response to various stresses (Zhang et al. 2006). Also, SSRs evolve rapidly, indicating that they provide an opportunity for rapid adaptive change in these regulatory regions or have a specific role in gene regulation. The possible functions of SSRs in mitochondrial gene promoter regions have not yet been shown. Here, most CREs that were close to SSR are related to cell development response and stress. In addition, many of the CREs are part of SSRs identified in the mitochondrial gene promoter regions. Our analyses suggest that the expansion of SSR motifs in the promoter region of mitochondrial genes may affect the expression of these genes in response to certain conditions and may potentiate or inactivate them. Certain SSRs may act as CREs, controlling gene expression through transcription factor binding. In *Triticum aestivum*, it has been reported that perfect tandem repeats, when present in multiple copies on the *TaALMT1* promoter, control the expression of a gene that is associated with tolerance to aluminum by enhancing *TaALMT1* expression (Gao et al. 2013). Similarly, the *waxy* genes from *Oryza sativa* contain an SSR motif (CT) in the 5'UTR region and SSR polymorphism is associated with amylose content (Li et al. 2004).

Considering the importance of CREs for plant breeding, it is suggested that certain long SSR motifs identified in the mtDNA promoter regions may act as CREs and increase gene expression; however, experiments should be performed to validate this hypothesis. Male sterility in the plants may be associated with inactivation of a mitochondrial gene by variation of an SSR sequence that acts as CREs (Mignouna et al. 1987). In *Oryza sativa* lines, this phenomenon assists in the commercial production of hybrid seeds (Ishii et al. 2001). It is also emphasized that altering the normal functioning of mitochondria in plants, i.e., a mutation altering or silencing the expression of a gene, can lead to changes in nuclear gene expression, causing serious damage to the plant and delaying stress signaling (Rhoads and Subbaiah 2007). Indeed, it should be noted that if an SSR motif is maintained during evolution to regulate transcription of a gene or to target a binding protein for one or more nuclear processes (such as chromatin organization, replication, transcription, and DNA recombination), its abundance and distribution are controlled (Zhang et al. 2006); and therefore, these regions that possess SSR may be under strong selective pressure.

### Mitochondrial SSRs can be efficient for use as molecular markers

Despite the low mutation rate, plant mtDNAs underwent major changes that affected genome features during evolution. As observed above, these changes can influence distribution of SSR between vascular and non-vascular plants, but in the evolutionary dynamics in each group of plant and algae as observed in cluster analysis. Cluster analysis based on SSR2-6 in mtDNAs of 204 plant species indicates that SSR sequences have regular distribution relative to plant groups. It means that they follow the evolution of the main lineages of green plants, including lineages that remain more conserved since the origin of land plants, as in the case of the Anthocerothyta, where the RE and SSR profiles are more similar to the flowering plants. mtSSRs from plants also appear to have accumulated mutations during evolution. This is evident when we observe the position of some species of algae grouped with the mosses or flowering plants (Figure 5).

In general, despite that mtDNAs do show a high rate of sequence reorganization during evolution mainly in algae, gymnosperms, and angiosperms (Agarwal et al. 2008) which would make SSR markers not interesting for phylogenetic analyses, we show through cluster analysis that its use is possible. Since many SSRs have accumulated mutations that are conserved at the family level, and some even at sister groups (Figure 5), variable long motifs were detected among species: in *Racomitrium* genus, the (TA)<sub>12</sub> motif with variation of 3 nt in the gene *cox1*. Also, variations in (AT)<sub>15</sub> inside an *NADH* intron were detected in some *Calypogeia* species. Variations in (GTTT)<sub>4</sub> upstream of *matR* were detected in some *Gossypium* species (Supplementary file 1 - Table S5). The large number and abundance of mtSSRs among mtDNAs, in addition to overrepresentation of SSR1, can be used as a possible source of variation as molecular markers. Several other studies have already successfully indicated the use of mtSSR regions in phylogenetic studies (Gupta and Varshney 2000; Rajendrakumar et al. 2008; Uthapaisanwong et al. 2017; Madhav et al. 2015) and studies of genetic structure (Sakaguchi et al. 2018), and also there are reports of use to distinguish CMS aberrant lines in pigeon pea.

An important issue related to mtSSR markers and chloroplast SSRs is that mtSSR markers may not show as much diversity as chloroplast SSRs in *Oryza*, which can be extended to other plant species in certain loci as well (Nishikawa et al. 2005). That conserved profile can be observed especially in genetic regions (Supplementary file 1 - Tables S3 and S4), being therefore more interesting the use of intergenic mtSSR markers for phylogenetic studies and differentiation of populations (Agarwal et al. 2008). On the other hand, mtSSRs present in coding regions of mtDNAs are extremely advantageous because they are



compatible among species due to the conservation of transcribed regions (Varshney et al. 2005). In comparison with conventional marker genes, SSRs may, at some point of view, be more advantageous as they provide greater levels of variation (Merritt et al. 2015). Taking these trends into account, markers based on mtSSRs may be useful in distinguishing closely related species. These findings point out that SSR markers may increase their application in future plant genetic and genomic studies.

## Material and methods

### Sequences of mitochondrial genomes of plants and algae

A total of 204 complete mitochondrial genomes of plant and algal species deposited in GenBank (NCBI Genome Information) were assessed. Of this total of mitochondrial genomes, 91 are from species of flowering plants, 3 gymnosperms, 2 ferns, 1 Lycopphyta, 48 species of bryophyte sensu lato (35 mosses, 11 Marchantiophyta, 2 Anthoceroophyta), and 58 algae species (50 Chlorophyta, 6 charophyte, 2 streptophyte). The accession number, genome size, and GC content of mitochondrial genomes are showed in Supplementary file 1 - Table S1. The annotation file (CDS feature) was used to differentiate SSRs in coding and non-coding regions. The total number of SSRs in non-coding regions was determined by subtracting the total number of SSRs in the mitochondrial genome from the total number of SSRs in the coding region.

### Identification and investigation of SSRs

The microsatellite (SSRs) search was performed using the SSR Locator software, a computational tool with an interface for Windows users (Maia et al. 2008). The algorithm used for perfect and imperfect SSR searches was written in Perl and consists of the generation of a matrix that mixes A (adenine), T (thymine), C (cytosine), and G (guanine) in all possible composite arrangements between 1 and 10 nucleotides. The script instructions perform readings on fasta files, searching all possible arrangements in each database sequence (Maia et al. 2008). Initially, it was developed for mining SSRs in EST/cDNAs sequences but can be efficient for accessory genomes. Mono- to decanucleotides can be identified, as well as perfect to imperfect motifs (Maia et al. 2008). Perfect and imperfect SSR mononucleotides (SSR1), di-, tri-, tetra-, penta-, and hexamers (SSR2-6) and perfect hepta-, octa-, nona-, and decamer (SSR7-10) minisatellites have been identified in mitochondrial plant genomes. This software identifies SSRs at a distance of 100 bp, but in this case because they are smaller genomes this value has been set to 10 bp. Only those repetitions where

the motif was repeated continuously for 3 or more times for SSR2-6 and SSR7-10, and repeated 6 or more times for SSR1, were considered. Similar threshold values have been previously used in plastid and mitochondrial genomes (George et al. 2015; Rajendrakumar et al. 2006). SSRs are defined as perfect, imperfect, and composite. The perfect SSR is a tandem repeat of exact copies of an SSR. SSRs are often interrupted by substitutions or indels resulting in an imperfect SSR sequence. A composite SSR is a region that contains two or more different perfect/imperfect SSRs that are overlapped or separated by a few nucleotides. In this study, composite SSRs were examined using the advanced mode of IMEX software (Mudunuri and Nagarajaram 2007). Some parameters have been set to - repeat type: perfect; - size of repetition: all; - minimum number of repetitions: 6,3,3,3,3,3; - maximum allowable distance between two SSR (dMAX) set to 10. This same threshold was used by other reports (Alam et al. 2017). Long SSRs were considered mononucleotide > 13, dinucleotide > 10, trinucleotide > 7, tetranucleotide > 5, pentanucleotide > 4, hexanucleotide > 4. Long SSR7-10s were considered > 5. As for composite SSRs (cSSR), they were called m1\_xn\_m2 “2-microsatellites” and m1\_xn\_m2\_xt\_m3 “3-microsatellites” (Chen et al. 2011). The analysis of preferred dimer motifs was performed by counting each specific dimer motif in each genome and then transformed into percentages.

### Calculation of expected number of SSR1s

The expected number of SSRs has been estimated for SSR1 due to their greater abundance in the analyzed genomes. Thus, seeking to assess whether SSR1s were over- or under-represented in the mitochondrial genome of plants, we compared the observed number of SSR1 ( $O$ ) with the expected number of SSR1 ( $E$ ) as an  $O/E$  ratio. The statistical significance of the  $O/E$  representation was accessed using the  $Z$  score defined as  $(O - E) / \sqrt{E}$  (Mrazek 2006). The expected number of SSR<sup>1</sup> is  $Mt$  ( $M$  is an SSR motif with repeat number  $t$ , and its size is  $L$ ) in a genome of size  $G$  calculated using Eq. 1, described by De Watcher (Watcher 1981):

$$\text{Exp}(Mt) = f(M) t [1 - f(M)] [G'(1 - f(M))] + 2L \quad (1)$$

Where  $\text{Exp}(Mt)$  is the expected number of  $Mt$ , and  $f(M)$  is the probability of  $M$ .

$G'$  defined as:

$$G' = G - tL - 2L + 1$$

A Python script was developed to automatically calculate the expected number of SSR1 from all 204 species analyzed. The script sums the total number and also each base type of the parsed sequence, then checks to see if the sum of all bases equals the total bases, therefore calculating the frequency of



each base of the sequence. The expected number of SSR<sup>1</sup> is given for each repetition (1–13) by the sum for each base frequency (fA, fT, fG, fC) using the equation (Frequency \*\* 1Repetition) \* (1-Frequency) \* ((BasesGenome - 1Repeat - 1) \* (1-Frequency) + 2). Thus, the script output identifies the species under analysis, the genome size, and the expected number of mononucleotide from 1 repeat to 13.

### Statistical analyses

To enable comparison between genomes of different sizes (relative abundance - RA) (Eq. 2), the total number of SSR classes was normalized to the number of SSRs per sequence kb. The estimated relative density (RD) (bp/kb) for each genome is defined as the sum of the total size (in nucleotides) contributing to each SSR per 1000 nucleotides of the genome analyzed (Eq. 3).

$$\text{Relative Abundance} : \frac{N^{\circ} \text{SSR}}{\text{genomesize}} \times 1000 \quad (2)$$

$$\text{Relative Density} : \frac{\sum \text{SSR}}{\text{genomesize}} \times 1000 \quad (3)$$

To identify the correlation between genomic characteristics and repeated sequences, linear regression was used with the aid of the Microsoft Office Excel 2007 program.

### Analysis of cis regulatory elements (CREs)

The long SSR3-6 located upstream to the gene, in promoter regions, were evaluated regarding the presence of *cis* regulatory elements (CREs) using the software PlantCARE (Rombaux et al. 1999) and SOGO (New PLACE) (Higo et al. 1999). Mitochondrial genes that contained tri-, tetra-, penta-, and hexanucleotide in the promoter of mitochondrial genes in algae, bryophytes (moss, Marchantiophyta, and Anthoceroophyta), ferns (Lycophyta and ferns), and spermatophytes (seed plants) were considered for analysis, representing all plant groups. A total of 14 promoter regions (4 from algae, 2 from bryophytes, 4 from ferns, and 4 from flowering plants) were analyzed. Only those SSR regions located less than 1500 bp from the gene transcription start site (TSS) were considered for analysis. In bryophyte species, only 2 promoters were selected, as we no longer have SSR species in the promoters with less than 1500 bp. Only elements with Matrix score  $\geq 5$  were considered.

### Cluster analyses of SSRs in plant and algal mitochondrial genomes

The cluster analysis was constructed using SSR2-6s from mitochondrial genomes of plants and algae of different species with symmetric Kullback-Leibler divergence analysis

(Kullback 1951). The difference between two species was quantitatively measured with the percentage of SSRs [ $p(x)$  and  $q(x)$ ] in two species respectively (Eq. 4).

$$\left( \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x q(x) \log \frac{q(x)}{p(x)} \right) \times \frac{1}{2}, \quad (4)$$

where  $x$  represents the SSR class (di-, tri-, tetra-, penta-, and hexanucleotide). Peer-to-peer analyses were performed among the 204 species resulting in a matrix of distances. Clustering analysis was conducted using the UPGMA method with the MEGA-X software (Kumar et al. 2018) according to the symmetrized Kullback-Leibler divergence analysis. After the creation of the clustering tree, it was adjusted in the iTOL online platform (Letunic and Bork 2016).

Graphical maps of mtDNA were prepared (Supplementary file 3). The graphical representation of the distribution of SSRs in the genomes (inner part of the maps) was performed using GenomeVx software (Conant and Wolfe 2008). On the outside of the maps, the genes of each species that were obtained using the OGDRAW software are displayed (Lohse et al. 2007). Linear graphical map of five mtDNAs from algal group (Supplementary file 5) was performed with OGDRAW software (Lohse et al. 2007). In the left inside, rectangles were inserted showing characteristics of the mtSSRs of each analyzed species, and above the linear maps the legend referring to the mitochondrial genes.

In order to make comparisons between the most representative repetitive elements in the genomes of the strains chosen for the present work, global analyses of repetitive DNA in all genomes and in mitochondrial genomes were performed. Mitochondrial raw reads were mapped from genomic reads of the main lineages of plant species, obtained at the NCBI/SRA, as follows: *Chlamydomonas nivalis* (SRX7092453); *Anthoceros agrestis* (ERX3577695); *Marchantia polymorpha* (SRX1669110); *Physcomitrium patens* (SRX1528135); *Selaginella tamariscina* (SRX3337410); *Ginkgo biloba* (SRX2319321); *Populus nigra* (SRX8743786); *Arabidopsis thaliana* (SRX972170); *Oryza sativa* (PRJNA702010). The repetitive DNA element profiles for each lineage were obtained using the RepeatProfiler pipeline (Novák et al. 2017; Negm et al. 2020).

### Conclusion

This study is the first large-scale analysis of SSR in mitochondrial genomes of plants and algae, where we included more than 200 species. Our comparative analysis shows that the distribution of mitochondrial SSRs depends on plant group analyzed, since the clustering places some species of algae and vascular plants together when assessing the



distribution of SSRs, reflecting the conservation of some common motifs in algae and terrestrial plants, making evident the ancient divergences of land plant evolution. In addition, each group of plants has SSR bias of types and distribution, showing algae as a very diverse group and vascular plants, presenting numerous and long motifs reflecting the large size of their mtDNA. The use of mtSSR markers from coding and non-coding regions makes them amenable to distinguishing closely related species. In general, mitochondrial SSRs are highly abundant and may represent an important source for the study of genetic variation and biotechnological studies since they can be associated to the gene regulation of mtDNA. Thus, this comparative study increases the understanding of the evolution of mtSSRs in plants and algae and brings perspectives for further studies.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s10142-021-00815-7>.

**Author contribution** K.E.J.F. wrote the main manuscript text and C.B. prepared figures. F.C.V. conducted the RE analysis and contributed to the main manuscript text. V.E.V., C. P., and L.C.M. made contributions to the discussion. A.C. led the research. All the authors reviewed the manuscript.

**Funding** This research received no external funding from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq -process n° 407591/2018-4), and from Fundação de Amparo a Pesquisa do Rio Grande do Sul (FAPERGS).

## Declarations

**Informed consent** Not applicable

**Conflict of interest** The authors declare no competing interests.

## References

- Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4:203–221
- Schlotterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* 20:211–215
- Tautz D, Schlotterer C (1994) Simple sequences. *Curr Opin Gen Dev* 4:832–837
- Gao C, Ren X, Mason AS, Li J, Wang W, Xiao M, Fu D (2013) Revisiting an important component of plant genomes: microsatellites. *Funct Pollut Biol* 40:645–661
- Li L, Wang B, Liu Y, Qiu YL (2009) The complete mitochondrial genome sequence of the hornwort *Megaceros aenigmaticus* shows a mixed mode of conservative yet dynamic evolution in early land plant mitochondrial genomes. *J Mol Evol* 68:665–678
- Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC (2006) Origin, evolution and genome distribution of microsatellites. *Gen Mol Biol* 29:294–307
- Zhao Z, Guo C, Sutharzan S, Li P, Echt CS, Zhang J, Liang C (2014) Genome-wide analysis of tandem repeats in plants and green algae. *G3: Ge Gen Genet* 4:67–78
- Provan J, Powell W, Hollingsworth PM (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends Ecol Evol* 16:142–147
- Devey DS, Chase MW, Clarkson JJ (2009) A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. *Taxon* 58:7–15
- Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, Rana JC, Singh NK, Sharma TR (2011) Genome-wide distribution and organization of microsatellites in plants: an insight into marker development in *Brachypodium*. *Plos One* 6:e21298
- Jiang JX, Wang ZH, Tang BR, Xiao L, Ai X, Yi ZL (2012) Development of novel chloroplast microsatellite markers for *Miscanthus* species (Poaceae). *Am J Bot* 99:e230–233
- George B, Bhatt BS, Awasthi M, George B, Singh AK (2015) Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Curr Genomics* 61:665–677
- Soranzo N, Provan J, Powell W (1999) An example of microsatellite length variation in the mitochondrial genome of conifers. *Gen* 42:158–161
- Rajendrakumar P, Biswal AK, Balachandran SM, Srinivasarao K, Sundaram RM (2006) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformatics* 23:1–4
- Fauron C, Allen J, Clifton S, Newton K (2004) Plant mitochondrial genomes. In: Daniell H., Chase C. (eds) *Molecular biology and biotechnology of plant organelles*. Springer, Dordrecht
- Kuntal H, Sharma V (2011) In silico analysis of SSRs in mitochondrial genomes of plants. *Omic* 15:783–789
- Zhao CX, Zhu RL, Liu Y (2016) Simple sequence repeats in bryophyte mitochondrial genomes. *Mitochond DNA Part A* 27:191–197
- Filiz E (2014) SSRs mining of Brassica species in mitochondrial genomes: bioinformatic approaches. *Hortic Environ Biotechnol* 54:548–553
- Liu Y, Medina R, Goffinet B (2014) 350 My of mitochondrial genome stasis in mosses, an early land plant lineage. *Mol Biol Evol* 31:2586–2591
- Raju GV, Rao PS, Rao CS, Sekhar VC, Mudunuri SB (2015) Microsatellite repeats in mitochondrial genomes: a bioinformatic analysis. *Proc Int Conf Adv Res Comp Sci Eng* 40:1–5
- Hancock JM (1999) Microsatellites and other simple sequences: genomic context and mutational mechanisms. In *Microsatellites: evolution and applications*. Edited by Oxford: Goldstein DB and Schlotterer C, 1–9.
- Bajaj D, Saxena MS, Kujur A, Das S, Badoni S, Tripathi S (2015) Genome-wide conserved non-coding microsatellite (CNMS) marker-based integrative genomics for quantitative dissection of seed weight in chickpea. *J Exp Bot* 66:1271–1290
- Gissi C, Iannelli F, Pesole G (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. *Heredity* 101:301–320
- Kitazaki K, Kubo T (2010) Cost of having the largest mitochondrial genome: evolutionary mechanism of plant mitochondrial genome. *J Bot* 620137.
- Race HL, Herrmann RG, Martin W (1999) Why have organelles retained genomes? *Trends Gen* 15:364–370
- Kuntal H, Sharma V, Daniell H (2012) Microsatellite analysis in organelle genomes of Chlorophyta. *Bioinformatics* 8:255–259
- Anand K, Kumar S, Alam A, Shankar A (2019) Mining of microsatellites in mitochondrial genomes of order Hypnales (Bryopsida). *Proc Sci Tod* 6:635–638
- De Watcher R (1981) The number of repeats expected in random nucleic acid sequences and found in genes. *J Theor Biol* 91:71–98

- Ceplitis A, Su Y, Lascoux M (2005) Bayesian inference of evolutionary history from chloroplast microsatellites in the cosmopolitan weed *Capsella bursa pastoris* (Brassicaceae). *Mol Ecol* 14:4221–4233
- Jakobsson M, Säll T, Lind-Halldén C, Halldén C (2007) Evolution of chloroplast mononucleotide microsatellites in *Arabidopsis thaliana*. *Theor Appl Genet* 114:223–235
- Wang Q, Fang L, Chen J, Hu Y, Si Z, Wang S, Zhang T (2015) Genome-wide mining, characterization, and development of microsatellite markers in *Gossypium* species. *Sci Rep* 5:10638
- Victoria FC, Da Maia LC, De Oliveira AC (2011) In silico comparative analysis of SSR markers in plants. *BMC Plant Biol* 11:15
- Srivastava S, Avvaru AK, Sowpati DT et al (2019) Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genom* 20:153
- Tian X, Strassmann JE, Queller DC (2011) Genome nucleotide composition shapes variation in simple sequence repeats. *Mol Biol Evol* 28:899–909
- Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177:309–334
- Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Gen* 30:194
- Qin Z, Wang Y, Wang Q, Li A, Hou F, Zhang L (2015) Evolution analysis of simple sequence repeats in plant genome. *PLoS One* 10:e0144108
- Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. *Science* 283:1476–1481
- Pombert JF, Beauchamp P, Otis C, Lemieux C, Turmel M (2006) The complete mitochondrial DNA sequence of the green alga *Oltmannsiellopsis viridis*: evolutionary trends of the mitochondrial genome in the Ulvophyceae. *Curr Gen* 50:137–147
- Grossman AR, Croft M, Gladyshev VN, Merchant SS, Posewitz MC, Prochnik S, Spalding MH (2007) Novel metabolism in *Chlamydomonas* through the lens of genomics. *Curr Opin Plant Biol* 10:190–8
- Liu F, Hu Z, Liu W, Li J, Wang W, Liang Z, Wang F, Sun X (2016) Distribution, function and evolution characterization of microsatellite in *Sargassum thumbergii* (Fucales, Phaeophyta) transcriptome and their application in marker development. *Sci Rep* 6:18947
- Rensing SA, Ick J, Fawcett JA, Lang D, Zimmer A, Van de Peer Y (2007) An ancient genome duplication contributed to the abundance of metabolic genes in the moss *Physcomitrella patens*. *BMC Evol Biol* 7:130
- Sousa F, Civián P, Brazão J, Foster PG, Cox CJ (2020) The mitochondrial phylogeny of land plants shows support for Setaphyta under composition-heterogeneous substitution models. *Peer J* 8:e8995
- Zhong B, Sun L, Penny D (2015) The origin of land plants: a phylogenomic perspective. *Evol Bioinform Online* 11:137–141
- De Vries J, Stanton A, Archibald JM, Gould SB (2016) Streptophyte terrestrialization in light of plastid evolution. *Trends Proc Sci* 21(6):467–476
- Bowles AMC, Bechtold U, Paps J (2020) The origin of land plants is rooted in two bursts of genomic novelty. *Curr Biol* 30(3):530–536.e2
- Rich MK, Delaux PM (2020) Plant evolution: when *Arabidopsis* is more ancestral than *Marchantia*. *Curr Biol* 30(11):R642–R644
- Xue JY, Liu Y, Li L, Wang B, Qiu YL (2010) The complete mitochondrial genome sequence of the hornwort *Phaeoceros laevis*: retention of many ancient pseudogenes and conservative evolution of mitochondrial genomes in hornworts. *Curr Gen* 56:53–61
- Liu Y, Xue JYY, Wang B, Li L, Qiu YLL (2011) The mitochondrial genomes of the early land plants *Treubia lacunosa* and *Anomodon rugelii*: dynamic and conservative evolution. *PLoS One* 6:e25836
- Field D, Wills C (1998) Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces. *Proc Nat Acad Sci* 95:1647–1652
- Medina R, Johnson M, Liu Y, Wilding N, Hedderson TA, Wickett N, Goffinet B (2018) Evolutionary dynamism in bryophytes: phylogenomic inferences confirm rapid radiation in the moss family Funariaceae. *Mol Phylogenet Evol* 120:240–247
- Sterck L, Rombauts S, Jansson S, Sterky F, Rouzé P, Van de Peer Y (2005) EST data suggest that poplar is an ancient polyploid. *New Phytol* 167(1):165–70
- Zhang L, Zuo K, Zhang F, Cao Y, Wang J, Zhang Y, Tang K (2006) Conservation of noncoding microsatellites in plants: implication for gene regulation. *BMC Gen* 7:323
- Knoop V (2010) Looking for sense in the nonsense: a short review of non-coding organellar DNA elucidating the phylogeny of bryophytes. *Trop Biol* 31:51–60
- Guo W, Zhu A, Fan W, Mower JP (2017) Complete mitochondrial genomes from the ferns *Ophioglossum californicum* and *Psilotum nudum* are highly repetitive with the largest organellar introns. *New Phytol* 213:391–403
- Guo W, Grewe F, Fan W, Young GJ, Knoop V, Palmer JD, Mower JP (2016) *Ginkgo* and *Welwitschia* mitogenomes reveal extreme contrasts in gymnosperm mitochondrial evolution. *Mol Biol Evol* 33:1448–60
- Chen M, Zeng G, Tan Z, Jiang Zhang J, Zhang C, Peng J (2011) Compound microsatellites in complete *Escherichia coli* genomes. *FEBS Lett* 585:1072–1076
- Metzgar D, Bytof J, Wills C (2000) Selection against frameshift mutations limits microsatellite expansion in coding DNA. *Gen Res* 10:72–80
- Li B, Xia Q, Lu C, Zhou Z, Xiang Z (2004) Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes. *Genom Prot Bioinf* 2:24–31
- Mignouna H, Virmani SS, Briquet M (1987) Mitochondrial DNA modifications associated with cytoplasmic male sterility in Rice. *Theor Appl Gen* 74:666–669
- Li YC, Korol AB, Fahima T, Nevo E (2004) Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 21:991–1007
- Da Maia LCD, Souza VQD, Kopp MM, Carvalho FIFD, Oliveira ACD (2009) Tandem repeat distribution of gene transcripts in three plant families. *Genet Mol Biol* 32:822–833
- Ishii T, Xu Y, McCouch SR (2001) Nuclear- and chloroplast-microsatellite variation in A-genome species of rice. *Genome* 44:658–666
- Rhoads DM, Subbaiah CC (2007) Mitochondrial retrograde regulation in plants. *Mitoch* 7:177–194
- Da Maia LC, Palmieri DA, De Souza VQ, Kopp MM, de Carvalho FIF, Costa de Oliveira A (2008) SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *Int J Plant Gen* 412696.
- Mudunuri BS, Nagarajaram AH (2007) IMEX: imperfect microsatellite extractor. *Bioinformatics* 23:1181–1187
- Alam CM, Iqbal A, Tripathi D, Sharfuddin C, Ali S (2017) Microsatellite diversity and complexity in eighteen *Staphylococcus* phage genomes. *Gene Cell Tissue* 4:3
- Mrazek J (2006) Analysis of distribution indicates diverse functions of simple sequence repeats in *Mycoplasma* genomes. *Mol Biol Evol* 23:1370–1385
- Rombauts S, Déhais P, Van Montagu M, Rouzé P (1999) PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res* 27:295–296
- Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res* 27:297–300



- Kullback S (1951) Leibler RA (1951) On information and sufficiency. *Ann Math Statist* 22:79–86
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Mol Biol Evol* 35:1547–1549
- Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res* 44:242–5
- Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* 24:861–862
- Lohse M, Drechsel O, Bock R (2007) Organellar Genome DRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Curr Gen* 52:267–274
- Novák P, Robledillo LA, Koblízková A, Vrbová I, Neumann P, Macas J (2017) TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nuc Acids Res* 45(12)
- Negm S, Greenberg A, Larracuente AM, Sproul JS (2020) Repeat-Profiler: a pipeline for visualization and comparative analysis of repetitive DNA profiles. *Mol Ecol Resour*
- Vieira ML, Santini L, Diniz AL, Munhoz C (2016) Microsatellite markers: what they mean and why they are so useful. *Genet mol biol* 39(3):312–328
- Sakaguchi S, Takahashi D, Setoguchi H, Isagi Y (2018) Genetic structure of the clonal herb *Tanakaea radicans* (Saxifragaceae) at multiple spatial scales, revealed by nuclear and mitochondrial microsatellite markers. *Plant Species Biol* 33(1):81–87
- Khera P, Saxena R, Sameerkumar CV et al (2015) Mitochondrial SSRs and their utility in distinguishing wild species, CMS lines and maintainer lines in pigeonpea (*Cajanus cajan* L.). *Euphytica* 206:737–746
- Gupta P, Varshney R (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113:163–185
- Agarwal M, Shrivastava N, Padh H (2008) Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep* 27:617–631
- Nishikawa T, Vaughan DA, Kadowaki K (2005) Phylogenetic analysis of *Oryza* species, based on simple sequence repeats and their flanking nucleotide sequences from the mitochondrial and chloroplast genomes. *Theor Appl Genet* 110(4):696–705
- Rajendrakumar P, Biswal AK, Balachandran SM, Sundaram RM (2008) In silico analysis of microsatellites in organellar genomes of major cereals for understanding their phylogenetic relationships. *In Silico Biol* 8(2):87–104
- Uthapaisanwong P, Somyong S, Tangphatsornruang S, Yoocha T, Jantasuriyarat C (2017) Development and characterization of simple sequence repeats derived from mitochondrial genome of oil palm using next generation sequencing. *Thai J Sci Technol* 6(3):288–300
- Madhav MS, Rajendrakumar P, Sivaraju K, Vishalakshi B, Devi SR (2015) Phylogenetic reconstruction of five Solanaceous species by genome-wide analysis of simple sequence repeats in organellar genomes and their utility in establishing species relationships of genus *Nicotiana*. *Curr Trends Biotechnol Pharmacol* 9(2):107–116
- Kocaman B, Sevim TOY, Marakli S (2020) Application of different molecular markers in biotechnology. *J Sci Lett* 2(2):98–113
- Merritt BJ, Culley TM, Avanesyan A, Stokes R, Brzyski J (2015) An empirical review: characteristics of plant microsatellite markers that confer higher levels of genetic variation. *Appl Plant Sci* 17;3(8):apps.1500025
- Varshney R, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. *Trends Biot* 23:48–55
- Cabañas N, Becerra A, Romero D et al (2020) Repetitive DNA profile of the amphibian mitogenome. *BMC Bioinformatics* 21:197
- Wang X, Zhang Y, Zhang H et al (2019) Complete mitochondrial genomes of eight seahorses and pipefishes (Syngnathiformes: Syngnathidae): insight into the adaptive radiation of syngnathid fishes. *BMC Evol Biol* 19:119
- Dong S, Zhaom C, Zhang S, Zhang L, Wu H, Liu H, Zhu R, Jia Y, Goffinet B, Liu Y (2019) Mitochondrial genomes of the early land plant lineage liverworts (Marchantiophyta): conserved genome structure, and ongoing low frequency recombination. *BMC Genomics* 20(1):953
- Ding S, Wang S, He K et al (2017) Large-scale analysis reveals that the genome features of simple sequence repeats are generally conserved at the family level in insects. *BMC Genomics* 18:848
- Hecht J, Grewe F, Knoop V (2011) Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biol Evol* 3:344–58
- Cole LW, Guo W, Mower JP, Palmer JD (2018) High and variable rates of repeat-mediated mitochondrial genome rearrangement in a genus of plants. *Mol Biol Evol* 35(11):2773–2785
- Toth G, Gaspari Z, Jurka J (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res* 10(7):967–81
- Harr B, Todorova J, Schlotterer C (2002) Mismatch repair-driven mutational bias in *D. melanogaster*. *Mol Cell* 10(1):199–205
- Wynn EL, Christensen AC (2019) Repeats of unusual size in plant mitochondrial genomes: identification, incidence and evolution. *G3 (Bethesda)* 9(2):549–559
- Kozik A, Rowan BA, Lavelle D, Berke L, Schranz ME, Michelmore RW, Christensen AC (2019) The alternative reality of plant mitochondrial DNA: one ring does not rule them all. *PLoS Genet* 30;15(8):e1008373
- Song X, Yang Q, Bai Y et al (2021) Comprehensive analysis of SSRs and database construction using all complete gene-coding sequences in major horticultural and representative plants. *Hort Res*, 8.
- Karlin S, Burge CB (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 11(7):283–90

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

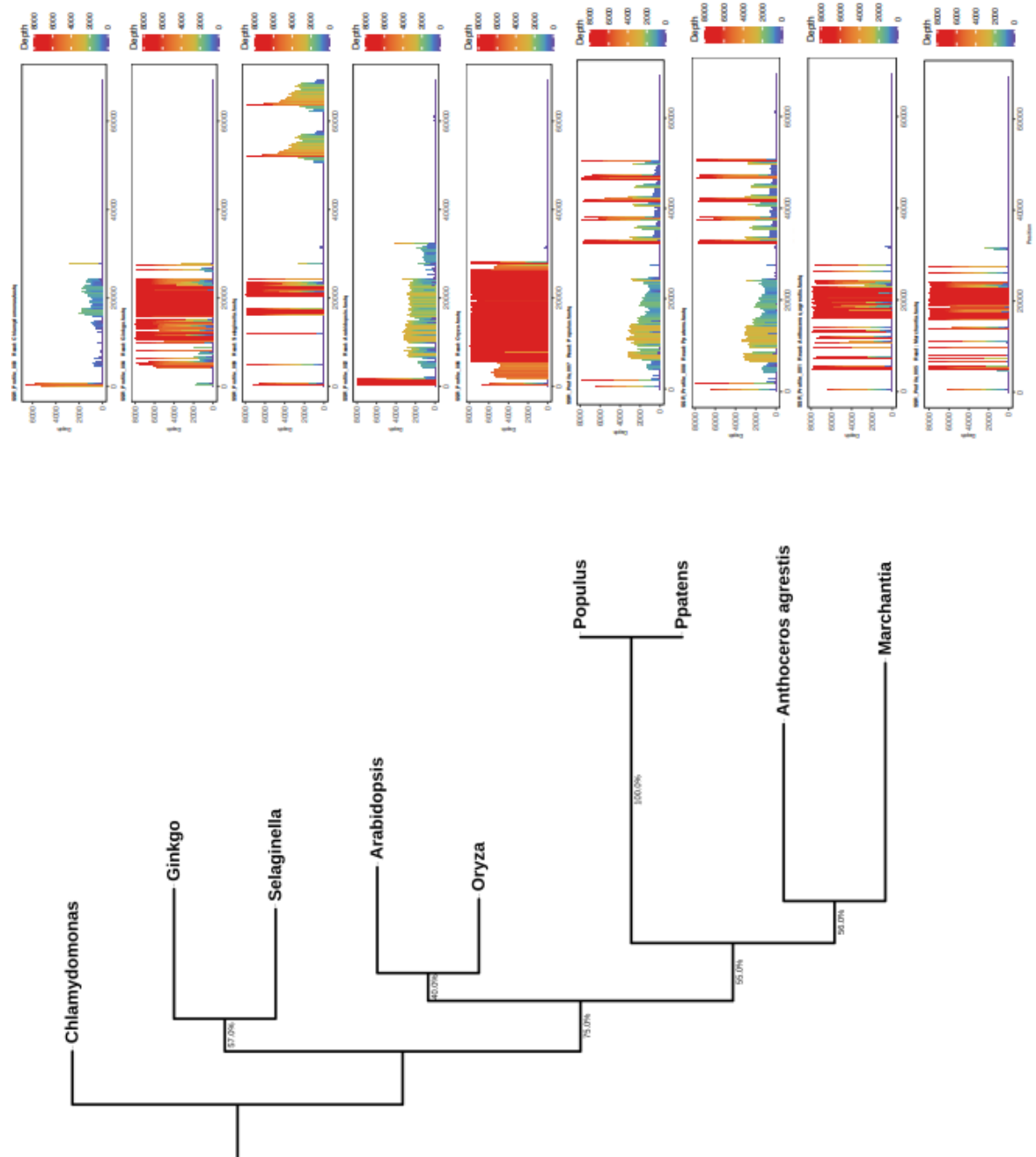
## Supplementary file1 (XLSX 350 KB)

**Table S1.** Overview of various mitochondria genomes and total number, relative abundance, density of microsatellites, and expected number of SSR<sup>1</sup> in the selected genomes. **Table S2.** Distribution of repeats types in mitochondria genomes. **Table S3.** Distribution of SSR<sup>1</sup> repeats in coding region of genome and non-coding. **Table S4.** Distribution of SSR<sup>2-10</sup> types in coding region of genome. **Table S5.** Distribution of long SSR<sup>2-6</sup> in mitochondria genomes from various plants groups **Table S6.** Linear correlation between genome size and GC content with SSR characteristics. (XLSX 350 kb).

Arquivo disponibilizado de forma separada disponível em formato .xlsx.

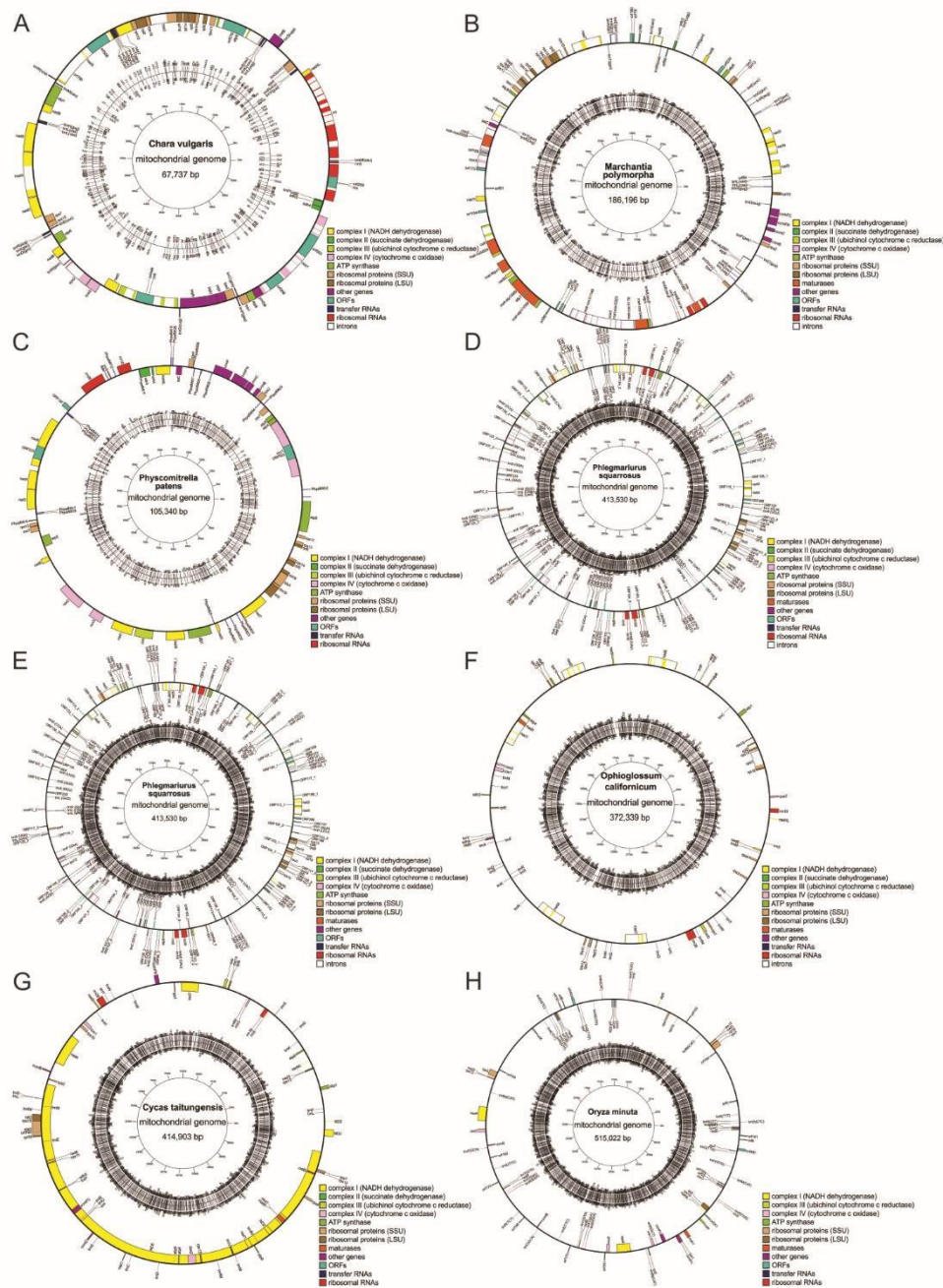


Supplementary file2 (PDF 6571 KB)



**Supplementary file2. A.** A tree shows the relationships among main green plants genomes inferred from variant signals in the profiles **B**. Position and depth of repetitive DNA in the main genome of an organism representing algae and an organism for each plant group.

**Supplementary file3 (PDF 4754 KB)**



**Supplementary file3.** Distribution of SSRs in the mitochondrial genome of an organism representing algae and an organism for each plant group. **A:** *Chara vulgaris* (algae); **B:** *Marchantia polymorpha* (liverwort); **C:** *Physcomitrium patens* (moss); **D:** *Phaeoceros laevis* (hornwort); **E:** *Phlegmarius squarrosus* (lycophyte); **F:** *Ophioglossum californicum* (fern); **G:** *Cycas taitungensis* (gymnosperm) and **H:** *Oryza sativa* (flowering plant).

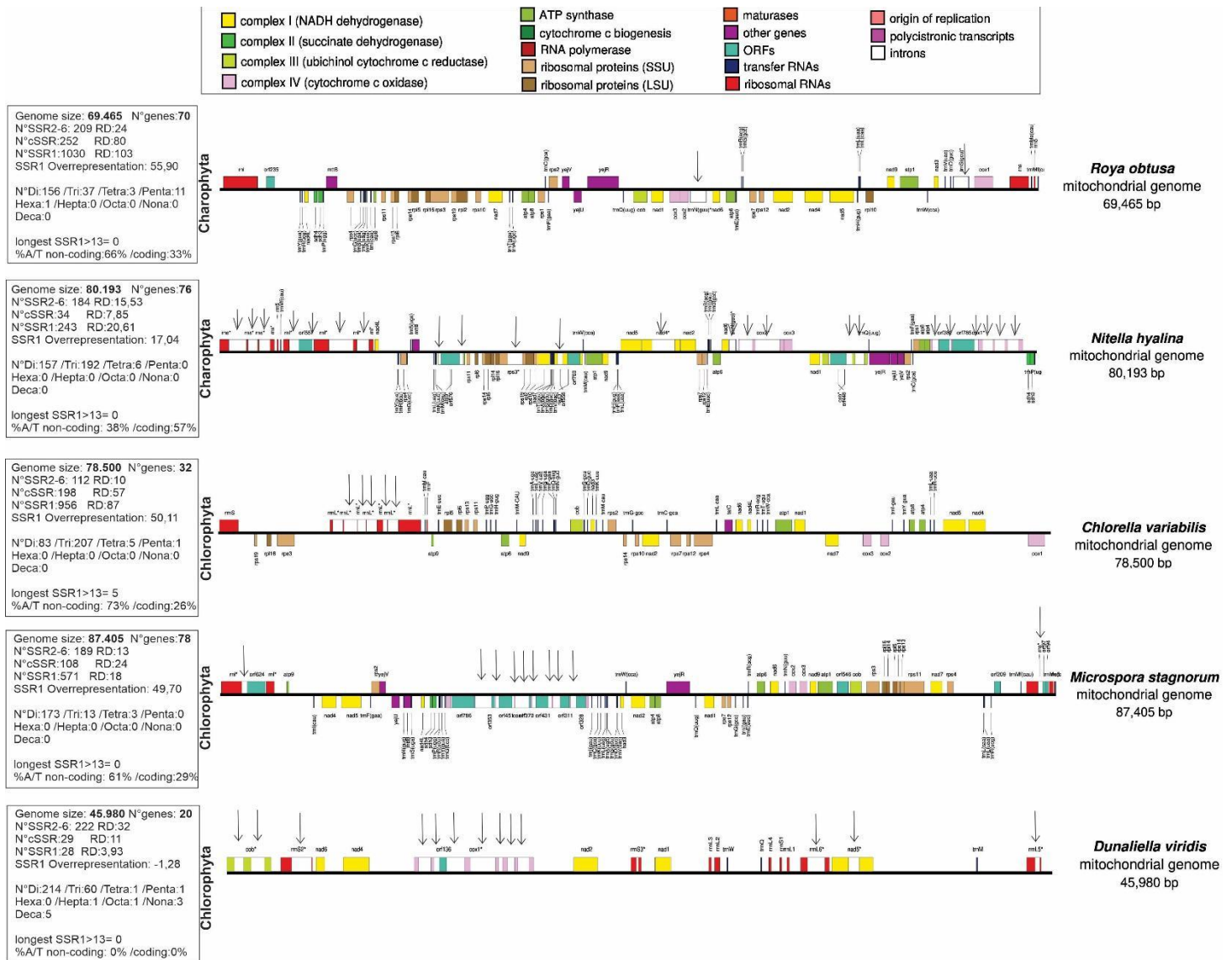
## Supplementary file4 (PDF 290 KB)

a) GCNAGCAGCAGCAGCAGCAGCAGCAGTGCAGTGCAGCACAGTAAACCGAA  
 (+) ANAERO2CONSENSUS S000478 854 AGCAGC  
 (+) ANAERO2CONSENSUS S000478 857 AGCAGC  
 (+) ANAERO2CONSENSUS S000478 860 AGCAGC  
 (+) ANAERO2CONSENSUS S000478 863 AGCAGC  
 (+) ANAERO2CONSENSUS S000478 866 AGCAGC  
 (+) ANAERO2CONSENSUS S000478 869 AGCAGC  
 (-) CACTFTPPCA1 S000449 875 YACT  
 (-) CACTFTPPCA1 S000449 880 YACT  
 (-) CACTFTPPCA1 S000449 890 YACT  
 (+) LTRE1HVBLT49 S000250 896 CCGAAA

b) TTATCTTTATCTTTATCTTTATCTTTATCTTTATCTTTATCTTT  
 (-) IBOXCORE S000199 851 GATAA  
 (-) GATABOX S000039 852 GATA  
 (-) NODCON1GM S000461 853 AAAGAT  
 (-) OSE1ROOTNODULE S000467 853 AAAGAT  
 (-) DOFCOREZM S000265 855 AAAG  
 (-) TAAAGSTKST1 S000387 855 TAAAG  
 (-) GT1CONSENSUS S000198 856 GRWAAW  
 (-) IBOXCORE S000199 857 GATAA  
 (-) GATABOX S000039 858 GATA  
 (-) NODCON1GM S000461 859 AAAGAT  
 (-) OSE1ROOTNODULE S000467 859 AAAGAT  
 (-) DOFCOREZM S000265 861 AAAG  
 (-) TAAAGSTKST1 S000387 861 TAAAG  
 (-) GT1CONSENSUS S000198 862 GRWAAW  
 (-) IBOXCORE S000199 863 GATAA  
 (-) GATABOX S000039 864 GATA  
 (-) NODCON1GM S000461 865 AAAGAT  
 (-) OSE1ROOTNODULE S000467 865 AAAGAT  
 (-) DOFCOREZM S000265 867 AAAG  
 (-) TAAAGSTKST1 S000387 867 TAAAG  
 (-) GT1CONSENSUS S000198 868 GRWAAW  
 (-) IBOXCORE S000199 869 GATAA  
 (-) GATABOX S000039 870 GATA  
 (-) NODCON1GM S000461 871 AAAGAT  
 (-) OSE1ROOTNODULE S000467 871 AAAGAT  
 (-) DOFCOREZM S000265 873 AAAG  
 (-) TAAAGSTKST1 S000387 873 TAAAG  
 (-) GT1CONSENSUS S000198 874 GRWAAW  
 (-) IBOXCORE S000199 875 GATAA  
 (-) GATABOX S000039 876 GATA  
 (-) NODCON1GM S000461 877 AAAGAT  
 (-) OSE1ROOTNODULE S000467 877 AAAGAT  
 (-) DOFCOREZM S000265 879 AAAG  
 (-) TAAAGSTKST1 S000387 879 TAAAG  
 (-) GT1CONSENSUS S000198 880 GRWAAW  
 (-) IBOXCORE S000199 881 GATAA  
 (-) GATABOX S000039 882 GATA  
 (-) NODCON1GM S000461 883 AAAGAT  
 (-) OSE1ROOTNODULE S000467 883 AAAGAT  
 (-) DOFCOREZM S000265 885 AAAG  
 (-) TAAAGSTKST1 S000387 885 TAAAG  
 (-) GT1CONSENSUS S000198 886 GRWAAW  
 (-) IBOXCORE S000199 887 GATAA  
 (-) GATABOX S000039 888 GATA  
 (-) NODCON1GM S000461 889 AAAGAT

**Supplementary file4.** Microsatellite identified by being part of or acting as a cis regulatory element. A: It has an AGC trimer motif that acts as a cis regulatory element in the mitochondrial mttB gene promoter of the *Closterium baillyanum* algae species. B: It has a TCTTTA hexamer motif that contains several cis regulatory elements in its structure. This microsatellite motif is found in the mitochondrial gene promoter rpl16 in the angiosperm species *Tripsacum dactiloides*.

## Supplementary file5 (PDF 290 KB)



**Supplementary file 5:** Linear graphical map of five mtDNA from algae group. In the left inside rectangles are inserted the characteristics of the mtSSRs of each analyzed species, and above the linear maps the legend referring to the mitochondrial genes.

## 4 CAPÍTULOS





### 4.2 Artigo 2 – ***Starch Synthesis-Related Genes (SSRG) Evolution in the Genus Oryza***

*Artigo publicado na revista Plants (MDPI)*



## Article

# Starch Synthesis-Related Genes (SSRG) Evolution in the Genus *Oryza*

Karine E. Janner de Freitas <sup>1</sup>, Railson Schreinert dos Santos <sup>2</sup> , Carlos Busanello <sup>1</sup> ,  
Filipe de Carvalho Victoria <sup>3</sup> , Jennifer Luz Lopes <sup>1</sup>, Rod A. Wing <sup>4,5</sup> and Antonio Costa de Oliveira <sup>1,\*</sup> 

<sup>1</sup> Centro de desenvolvimento Tecnológico—CDTec, Graduate Program in biotechnology, Capão do Leão Campus, Federal de Pelotas, Pelotas 96160, Brazil; karinejanner@gmail.com (K.E.J.d.F.); carlosbuzza@gmail.com (C.B.); jenniferlopesagronomia@gmail.com (J.L.L.)

<sup>2</sup> Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar), Alegrete 97541, Brazil; railson.schreinert@iffarroupilha.edu.br

<sup>3</sup> Núcleo de Estudos da Vegetação Antártica—NEVA, Campus São Gabriel Federal do Pampa (UNIPAMPA), São Gabriel 97030, Brazil; filipevictoria@unipampa.edu.br

<sup>4</sup> The School of Plant Sciences, Ecology & Evolutionary Biology, Arizona Genomics Institute, Tucson, AZ 97030, USA; rwing@ag.arizona.edu

<sup>5</sup> Center for Desert Agriculture, King Abdullah University of Science & Technology, Thuwal 23955, Saudi Arabia

\* Correspondence: acostol@terra.com.br

**Abstract:** Cooking quality is an important attribute in Common/Asian rice (*Oryza sativa* L.) varieties, being highly dependent on grain starch composition. This composition is known to be highly dependent on a cultivar's genetics, but the way in which their genes express different phenotypes is not well understood. Further analysis of variation of grain quality genes using new information obtained from the wild relatives of rice should provide important insights into the evolution and potential use of these genetic resources. All analyses were conducted using bioinformatics approaches. The analysis of the protein sequences of grain quality genes across the *Oryza* suggest that the deletion/mutation of amino acids in active sites result in variations that can negatively affect specific steps of starch biosynthesis in the endosperm. On the other hand, the complete deletion of some genes in the wild species may not affect the amylose content. Here we present new insights for *Starch Synthesis-Related Genes* (SSRGs) evolution from starch-specific rice phenotypes.

**Keywords:** *Leersia perrieri*; phylogeny; starch synthesis; cooking quality



**Citation:** de Freitas, K.E.J.; dos Santos, R.S.; Busanello, C.; de Carvalho Victoria, F.; Lopes, J.L.; Wing, R.A.; de Oliveira, A.C. *Starch Synthesis-Related Genes* (SSRG) Evolution in the Genus *Oryza*. *Plants* **2021**, *10*, 1057. <https://doi.org/10.3390/plants10061057>

Academic Editor: Sung-Ryul Kim

Received: 27 February 2021

Accepted: 14 April 2021

Published: 25 May 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Common rice (*Oryza sativa* L.) is a food of great importance worldwide, especially in Asian countries, where it is an important part of local culture. Being widely consumed and having different forms of preparation makes “quality” something different in each country around the world. Nevertheless, no matter what grain quality means, its demand is increasingly becoming a priority for international export markets worldwide.

Today, cooking has become one of the most important research components in several rice breeding programs, where characteristics such as amylose content (AC) and gelatinization temperature (GT), which have major effects on cooking quality (CQ) and consumption, are controlled by physicochemical properties of starch in rice grain endosperm [1].

The ratio of amylase to amylopectin as well as the structure of amylopectin itself can vary greatly between different rice genotypes [2]. Generally, grains with higher amylose content present a harder non-sticky texture after cooking, being preferred in several countries. Such feature is usually evaluated during grain development in different cultivars [3]. However, the genetic events that lead to this type of grain are not well understood, and genotypes that deliver such grains are not easily obtained. That is the reason it is so



important to understand the behavior of grain-quality-related genes, which enable more efficient and precise breeding applications.

The 27 known *Oryza* species span over 15 million years of evolution which we can take advantage of, since it constitutes a rich source of genetic variation. Though a better understanding of the genomic differences between these species is essential for such purpose, the recent publication of the genomes of 13 rice species has opened the door to a series of new studies that make it possible to enrich the germplasm that can be used for breeding [4,5]. The possibility of using these wild species to improve grain quality should also be considered, but what would be the first genes to start such an analysis?

The answer to this question is directly linked to the formation of amylose and amylopectin, which are two types of polysaccharides that form starch granules, and both have very complex biosynthesis processes. The ratio of these polysaccharides has a major influence on the appearance and structure of starch granules and consequently affects the quality of food production and industrial applications [1,6]. *Starch Synthesis-Related Genes (SSRGs)* are involved in this complex starch biosynthesis process, and can be divided into four classes: ADP-glucose pyrophosphorylase (AGPase), starch synthase (SS), starch branching enzyme (SBE) and starch de-branching enzyme (DBE). A very simplified explanation can be described, and this begins in cytosol for synthesis of amylopectin, in which AGPase synthesizes ADP-glucose from Glc1P and ATP through AGPs (AGPS2a), which plays a catalytic function, while AGPLs (AGPL1, AGPL3, AGPL4) are mainly responsible for modulating the allosteric regulatory properties [7,8]. Already in plastid, the elongation of glucan chains occurs through 8 SS enzymes: SSI plays a role in short chains; SSII1/2/ALK in medium chains; SSIII1/2 in long chains; and SSIV1/2 in terminal chains for formation of 1,4- $\alpha$ -D-glycosyl [9,10]. Meanwhile, the other SS enzymes, GBSSII and Waxy, synthesize amylose (10). After the formation of 1,4- $\alpha$ -D-glycosyl, the enzymes from SBE class, SBE1 and SBE3 [9], work by branching the chains until the formation of 1,6- $\alpha$ -D-glycosyl. From there, the enzymes from DBE class—PUL [11], ISA [12] and DPE1 [13]—de-branch short external chains of glucans and also influence the activity of  $\alpha$ -amylase and  $\beta$ -amylase [6].

In the first studies with SSRGs, it was identified that the Waxy gene (Granule-bound starch synthase I—from SS class) is directly involved with AC [13], while ALK (Starch synthase II-3—from SS class) is involved with TG [9]. The following studies contributed to the understanding that each SSRG is involved with one of the main starch-related quality characteristics (AC, GT, CQ) [9]. Despite this, it is also known that all SSRGs act in a coordinated way, forming a fine regulatory network in which the absence or change in the performance of a gene can ultimately lead to grain malformation [6,14], making it difficult to predict the nature of the starch resultant from biotechnological modifications. This makes it necessary to gain more in-depth knowledge about each of these genes through Targeted and Open-Ended Approaches [15]. Some of these have already been studied in some rice genotypes, achieving high-amylose content by: (1) transgenic knockdown of SBE1, SBEIIa and SBEIIb [16]; (2) CRISPR/Cas9-Mediated targeted mutagenesis of Starch Branching Enzymes [17]; (3) editing of rice isoamylase gene ISA1 [18]; and (4) downregulating of SSII2 caused lower AC in the endosperm [19]. Despite that, many studies regarding the variation of isozyme show that some variations in gene structure may not be so beneficial, mainly the non-synonymous substitutions that can affect the active domain of the isozyme. As presented through TILLING, in which mutations in the region of the exon 9/intron and exon 10 caused AGPL subunit (AGPase class), mutants severely shriveled with low weight and starch content [7].

Considering the importance of SSRGs in the control of CQ and the limited exploration of the information recently made available to the scientific community on *Oryza* genomes, an evolutionary analysis is needed to reveal the role of adaptive mechanisms before and after rice domestication. It will thus help to understand the complexity of the evolution of enzymes involved in the starch synthesis pathways, and further provide the basis for approaches that can generate new phenotypes through new strategies to modify starch synthesis. We therefore selected a set of SSRGs according to Zeng et al. [20], to explore

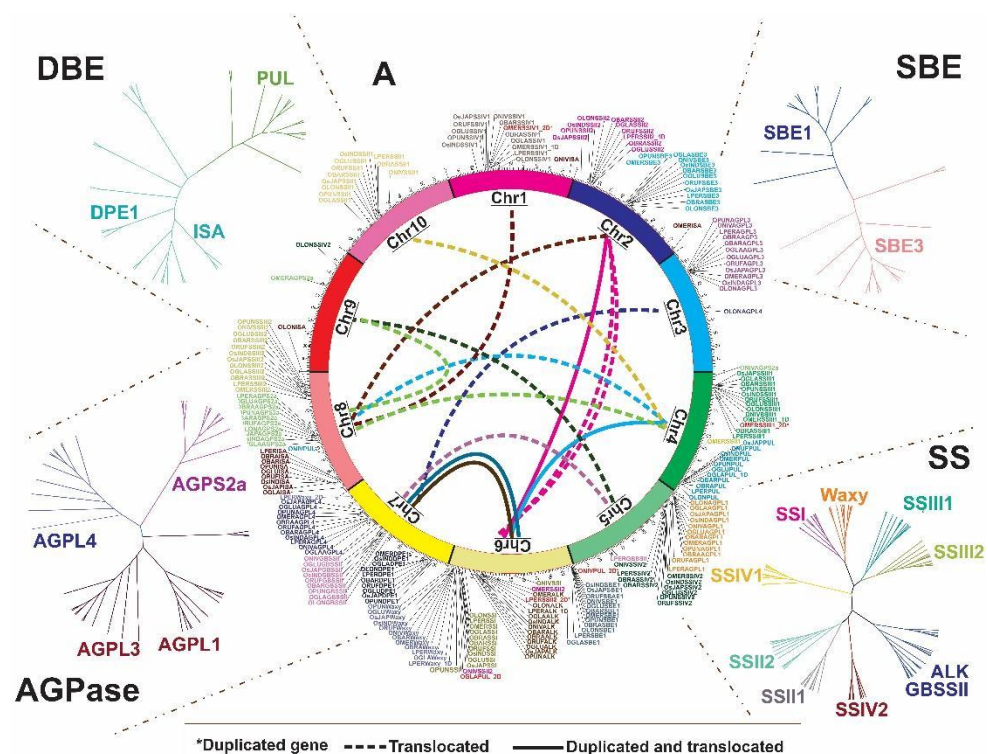


their molecular evolution across the genus *Oryza* in which we observe differences in the structure of genes and proteins that can imply changes in the content of amylose and amylopectin. Table S1 presents all SSRGs identified in 11 *Oryza* species and *Leersia perrieri*, which is the nearest outgroup of the genus *Oryza*.

## 2. Results

### 2.1. AGPase Subunits

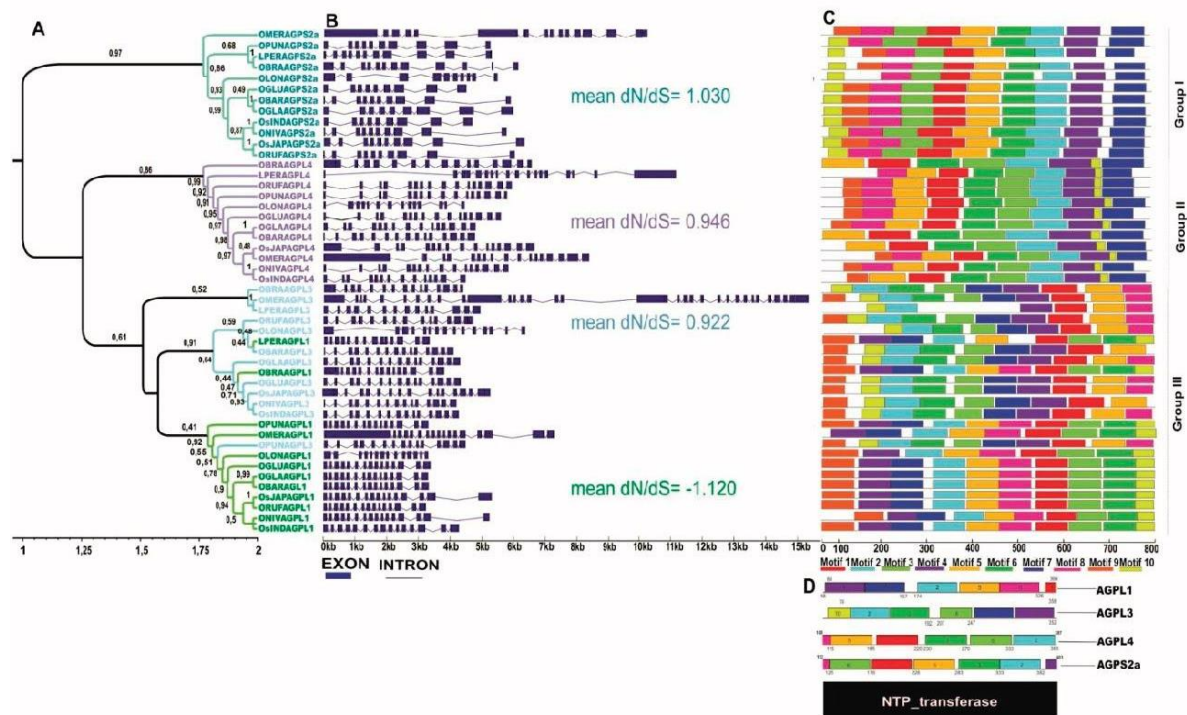
The phylogenetic tree of the ADP-glucose pyrophosphorylase (AGPase genes), which includes both large (AGPL1, AGPL3, AGPL4) and small (AGPS2a) subunits, identified 48 genes across the *Oryza* and the outgroup (*L. perrieri*), which revealed three different groups (Figures 1 and 2). The first group is formed by the AGPS2a genes with both large exon and intron structures (Figure 2B). This group is believed to have the highest similarity to the ancestor of every *Oryza* AGPase gene. *O. meridionalis* (AA) also contains the same large exon structure in the second group formed by AGPL4 and, likewise, in the third clade which is a mixture formed by AGPL3 and AGPL1, respectively (Figure 2A).



**Figure 1.** Gene localization and phylogeny of SSR proteins. Center. Circular map of SSRGs in a single representation from chromosomes 1 to 10 in *Oryza* species. Duplicated genes are marked with an asterisk. Around the map. Phylogenetic analysis of DBE (De-branching enzymes) (represented by DPE1 (Disproportionating enzyme), ISA (Isoamylase), PUL (Pullulanase); SBE (Starch branching enzymes) (represented by SBE1 and SBE3); SS (Starch synthase) (represented by SSI, SSII1, SSII2, SSIII1, SSIII2, SSIV1, SSIV2, Waxy (Granule-bound starch synthase I), GBSSII (Granule-bound starch synthase II); and AGPase proteins. Clade groups are indicated by different colors followed by the name.

Positive selection pressure was identified in the AGPS2a gene (Figure 2B). The conserved motifs of the four analyzed AGPase subunits form a signature pattern, revealing that motifs 9 and 10 are not detectable in the NTP transferase domain of AGPS2a; the same occurs for motifs 8 and 9 in AGPL4; 7, 8 and 9 in AGPL1; and 3, 6, 8 and 9 in AGPL3, which are not found in some *Oryza* species (Figure 2C,D).





**Figure 2.** Phylogenetic relationship, genetic structure and analysis of conserved motifs in *AGPase* genes of species of the genus *Oryza*. (A) Phylogenetic protein tree. The branches of the AGPS2a, AGPL4, AGPL3 and AGPL1 proteins are marked in dark green, purple, light blue and light green, respectively. The bootstrap values are indicated in the phylogenetic tree. (B) Exon-intron structure of the *AGPase* genes in *Oryza*. (C) Protein motifs indicated in different colors. (D) Protein domain shared by *AGPase* proteins.

Recombination analysis based on the alignment does not show any evidence of recombination in the AGP partition. On the other hand, positive selective pressure ( $dN/dS > 1$ ) was detected in some sites of sequence alignments, suggesting diversifying selection (Figure S1).

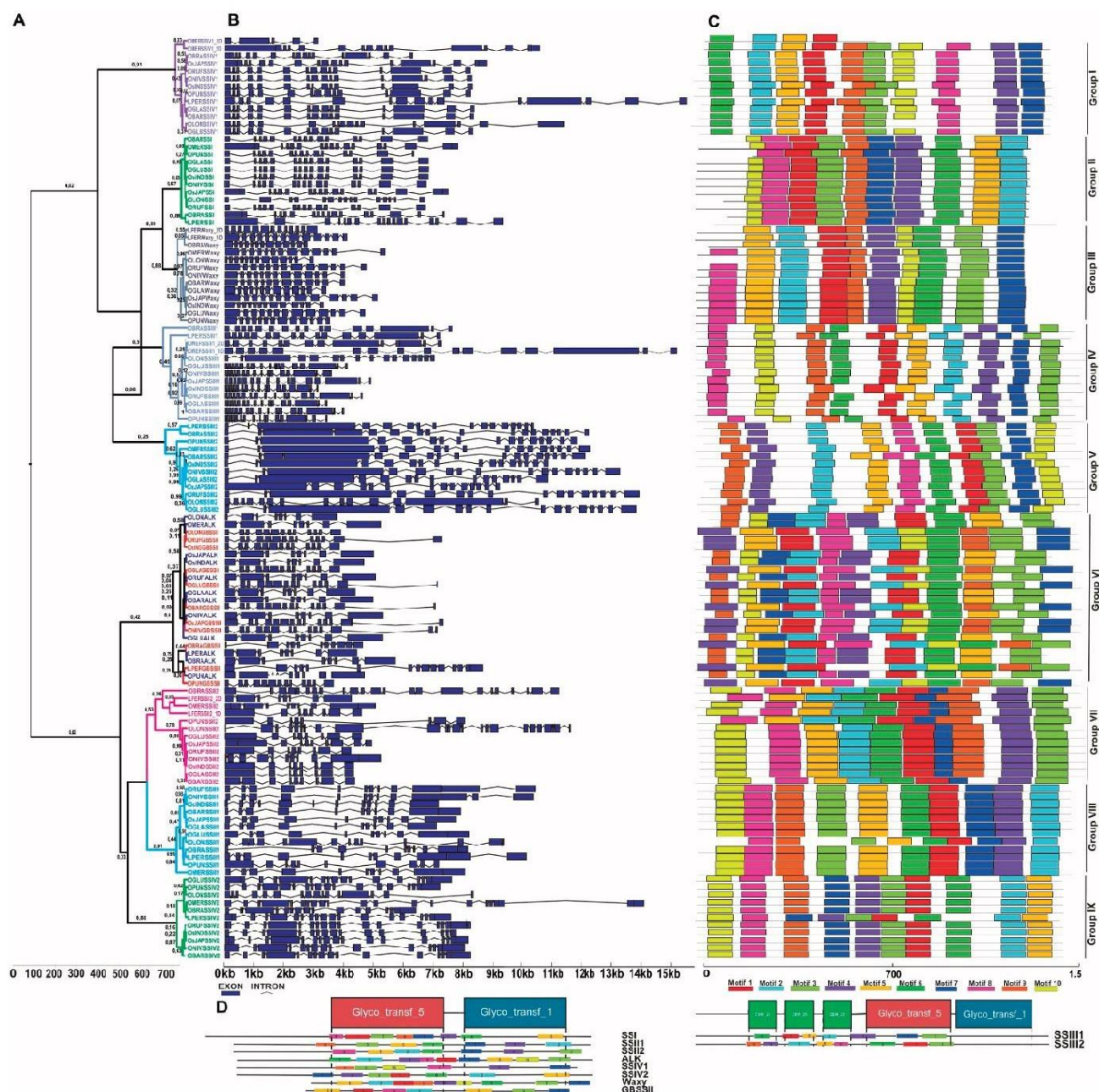
In relation to gene position, all *Oryza* species have *AGPS2a* positioned in Chr. 8; of note, we observed that *OMERAGPS2A* and *ONIVAGPS2A* are located on two different chromosomes (i.e., Chr. 9 and Chr. 4, respectively (Figure 1)). Possible differential Mobile Element Insertion (MEI) events related to these loci was investigated, a region of 50 kb up- and downstream of these genes were aligned, showing high similarity between *OMERAGPS2A* (AA) to *AGPS2a* of other species, which means that this change in position probably did not occur through transposable element (TE) insertion (Figure S2).

## 2.2. Starch Synthesis (SS) Genes

A total of 92 protein coding starch synthase (SS) genes (SSI, SSII1, SSII2, ALK, SSIII1, SSIII2, GBSSII, Waxy, SSIV1 and SSIV2) were found across the 12 genomes data set, while its phylogenetic analysis allowed the identification of nine different clades based on sequence similarity. Clades I, II, III, IV, V, VI, VII, VIII and IX typically represent SSIV1, SSI, Waxy, SSIII1, SSIII2, GBSSII / ALK, SSII2, SSII1 and SSIV2, respectively (Figures 1 and 3).

Some genes that have long exons near the 5' or 3' UTRs, as observed in few SS proteins of *L. perrieri*, *O. longistaminata*, *O. brachyantha* and *O. meridionalis*, seem to be ancestors of other species SSs (Figure 3A,B). The duplication of gene *OMERSSIV1\_2D* (Clade I) is a probable result of a sub-functionalization since it does not contain motifs 4, 7 and 8, which represent the catalytic domain of starch synthase (Glyco\_transf\_5) and (Glyco\_transf\_1) (Figure 3C,D).





**Figure 3.** Phylogenetic relationship, genetic structure and conserved motifs/domain analysis in SS genes of *Oryza* species. (A) Phylogenetic protein tree and bootstrap values of SSI, SSIII, SSII2, ALK, SSIII1, SSIII2, GBSSII, Waxy, SSIV1 and SSIV2. (B) Exon-intron structure of SS genes in *Oryza*. (C) Arrangement of the 10 most frequent motifs found in the analyzed proteins. (D) Composition and distribution of domains and conserved motifs of SS proteins.

In addition to *OMERSSIV1\_2D*, another recent duplication was identified in the *O. meridionalis* *SSIII1* gene, but in this case both the original and duplicated copies look functional, containing all the motifs that are part of its characteristic domain. However, the large size of *OMERSSIII1\_1D* (7844 bp longer than the original copy) is something that deserves more investigation, especially when we take into account the highly conserved profile of these genes (Figure 3A–C). It is also important to notice that the same large domain occurs in duplicated copies of *SSII2* and *Waxy* in the outgroup *L. perrieri*.

Some *Oryza* species and *L. perrieri* show changes in chromosome position of the SS genes relative to *O. sativa* (Figure 1), such as *OMERSSIII1* from Chr. 10 to 4 (Figure S3),

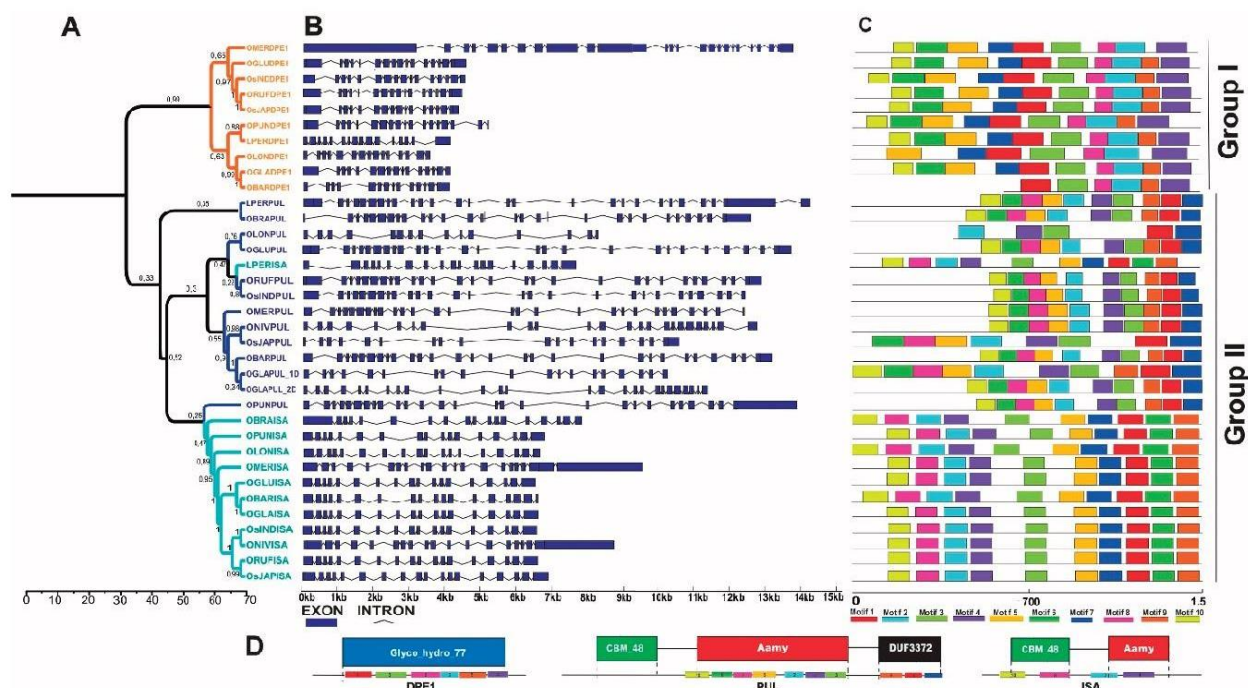


*OMERSSII2* from Chr. 2 to 6, *ONIVSSII2* from Chr. 2 to 6 (Figure S4) and *OLONSSIV2* from Chr. 5 to 9 (Figure S5). An alignment analysis shows that, for *OMERSSII1* and *OMERSSII2*, the change did not occur through a differential TE insertion, since an analysis of 50kb upstream and downstream of each gene shows a lack of or just partial synteny (fragments from approximately 40 Kb) between the other *Oryza* loci. In case of partial synteny, a significant presence of TEs in this region was not identified using that with RiTE-DB. Interestingly, *OMERSSII2* contains an inverted region of 50 kb that denotes an unusual rearrangement by translocation and inversion of blocks up- and downstream of the gene (Figures S3 and S4).

Recombination events are found in both *ALK* and *Waxy* gene copies (Figures S6 and S7), while for *Waxy* stronger evidence of breakpoints can be identified (Figure S6). However, the same was not observed for the other *SS* copies, where 117 were found to be under positive selection (Figure S8) with no recombination events detected.

### 2.3. De-Branching Enzymes (DBE)

The DBE (De-branching enzymes) genes are classified as *DPE1*, *PUL* and *ISA*. In total, 35 DBE genes were identified in *Oryza* and *L. perrieri* in the 12 genomes data set (Figures 1 and 4). The phylogenetic analysis showed that the DBE proteins can be grouped in two clades, one that comprises *DPE1* (Group I) and the other consisting of a mixed group composed of *PUL* and *ISA* proteins (Group II) (Figure 4A).



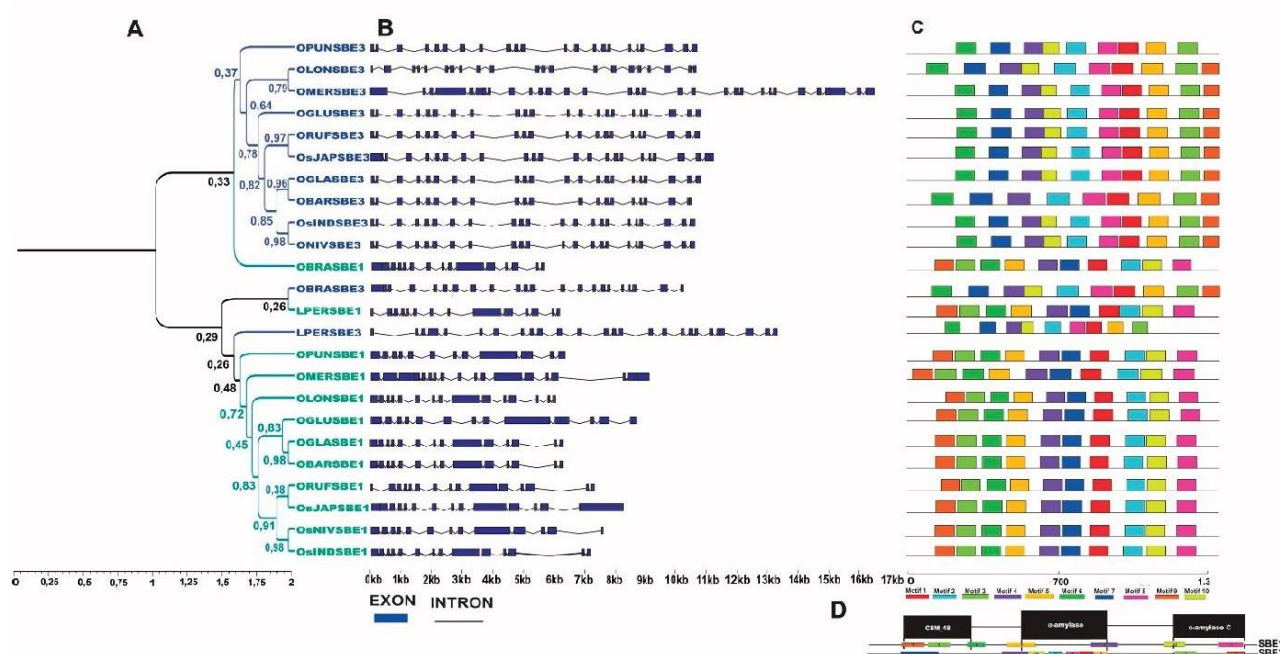
**Figure 4.** Phylogenetic relationship, genetic structure and conserved motifs/domain analysis in DBE genes of *Oryza* species. (A) Phylogenetic protein tree and bootstrap values of DPE1, PUL and ISA with branches marked in orange, blue and green, respectively. (B) Exon-intron structure of DBE genes in *Oryza*. (C) Arrangement of the 10 most frequent motifs found in the analyzed proteins. (D) Compositions and distributions of domain structures and conserved motifs of DBE proteins.

*DPE1*, despite forming a conserved clade, presents some variations in its two subgroups. First, *O. meridionalis* (AA) shows the longest gene structure, with more than eight exons, being the longest in its 5' UTR, something that contrasts with the usual short structure of *DPE1* genes (Figure 4B). Furthermore, *OMERDPE1*, *OBARDPE1* and *OLONDPE1* lack motif 9, which is part of glycoside hydrolase family 77 domain (Glico\_transf\_77), a domain responsible for cleaving the starch granule into smaller glucan molecules (Figure 4C,D).

On the other hand, *ISA*, different from *PUL*, contains long and frequent introns in its gene structure; besides this, it also possesses every single motif that forms the formerly discussed protein signature (Figure 4B,C). We identified an event in *O. glaberrima* where *PUL* (Figure 1) is duplicated and translocated from Chr. 4 to Chr. 6 (Figure S9). Recombination analysis was performed, but no recombination could be inferred (Figure S10).

#### 2.4. Starch Branching Enzymes (SBE)

In total, 24 SBE (Starch Branching Enzymes) genes were identified in *Oryza* and *L. perrieri* (Figures 1 and 5). According to the position of *L. perrieri* in the phylogenetic tree, *SBEs* are defined as a mixed clade, which presents a very conserved gene structure and protein signature that comprises *SBE3* and *SBE1* (Figure 5A–C). Although the conserved motif analysis showed that motif 9 is not present in *OPUNSB3* and *OMERSBE3*, they contain many more exons than the *SBE3*s of other *Oryza* species (Figure 5C).



**Figure 5.** Phylogenetic relationship, genetic structure and conserved motifs/domain analysis in *SBE* genes of *Oryza* species. (A) Phylogenetic protein tree and bootstrap values of *SBE1* and *SBE3* with branches marked in green and blue, respectively. (B) Exon-intron structure of *SBE* genes in *Oryza*. (C) Arrangement of the 10 most frequent motifs found in the analyzed proteins. (D) Compositions and distributions of domain structures and conserved motifs of *SBE* proteins.

*SBE* proteins are characterized by a modular architecture composed of an N-terminal domain with a carbohydrate-binding module family 48 (CBM48), a central  $\alpha$ -amylase domain, as well as a  $\alpha$ -amylase C-terminal domain (Figure 5D).

Positive selection can be seen through the alignment of these genes, with 64 sites undergoing a diversification process (Figure S11), but no recombination events were detected.

### 3. Discussion

#### 3.1. AGPase Subunits

In *Oryza* and other plants, the AGPase protein subunit is characterized by a core region that is important for catalytic activity, called the nucleotidyl transferase domain (NTP transferase) that is important in providing the substrate for starch biosynthesis. The absence of specific motifs can affect the endosperm starch synthesis limiting the reaction converting Glucose 1-Phosphate (Glc-1-P) and Adenosine triphosphate (ATP) to ADP-glucose and inorganic pyrophosphate (PPi) in amyloplasts, directly reflecting the control



of carbon flux into the starch accumulation pathway, consequently causing a shrunken endosperm in rice [1,6,21].

The evolution of the subunits of ADP proteins in *Oryza* is notably different from that reported in the literature for counterparts in other plant species [22]. This is probably due to different rates of selective pressure between species, which makes even more complex the study of the diversification of *AGPS2a* (Figure 2B).

In other plant species, such as dicots, the small subunits are under higher purification selection, thus remaining more conserved over time than the large subunits, which are primarily responsible for most of the diversification of *AGPase* genes [22,23]. The large subunits concentrate most of the positive selection, showing a great variability. However, in our results, when comparing the *AGPase* copies only inside *Oryza* genus, the opposite is observed. One explanation for this would be that in *AGPase* genes the large subunits concentrate most of the duplications [24]. However, Non-Homologous Recombination (NHR) cannot yet be ruled out, since both our data and previous reports indicate that NHR can be more frequent than MEI in *Oryza* species [25]. On the other hand, contrasting evolutionary patterns are expected between paralogues, and in the case of *AGPase* in *Oryza*, some duplications are already known and accompanied by a change in cell compartmentalization (from plastids to cytosol) and in their regulating properties [26].

In relation to gene position change, Non-Homologous Recombination (NHR) is likely to have occurred, placing this *OMERAGPS2A* large block (upstream + gene + downstream) in Chr. 9. The locus from *O. nivara* in Chr. 4 has only a small ortholog block that corresponds to the end of the upstream region and the start of the downstream region. Small up- and downstream fragments similar to specific LTR-TEs were found using the Rice Transposable Elements database (RiTE-db), but it is unlikely that these are responsible for a translocation event. As previously reported, the most frequent events responsible for changing copy number variations and gene position to other chromosomes are mediated by either transposable elements, through MEI or NHR, for both *Oryza* and *Arabidopsis* [23,27].

### 3.2. Starch Synthesis (SS) Genes

The phylogenetic analysis showed that, in most *Oryza* species, SS isoforms have undergone different degrees of gene duplication, something that is also observed in most plant species. *Oryza* clades I, IV, V, IX possess a different genetic origin from clades II, III, VI, VII and VIII and, since paralogous genes tend to slowly accumulate variations over time, it is easy to notice a large variation when we compare SS genes between these two clades [28–31]. The distinct spatial pattern of starch deposition within a caryopsis, which is also related to differences in the temporal expression pattern between early (SSIII1, SSII2, GBSSII) and late (ALK, SSIII2, Waxy) expressed genes [10], is probably the result of variations accumulated over time. Overall, the phylogenetic tree analysis reveals a highly conserved structure for both gene and amino acid sequences, suggesting a strong evolutionary relationship between species in each SS.

Taking into account that sequence variation in SSRGs have a great influence in rice amylose content, gelatinization temperature and amylopectin chain length [32], although important, it is hard to understand the roles of each SS isoform in each of the characters, due to the high sequence variation among these genes. Furthermore, it is even more complicated when we consider its diversity of genes in starch biosynthesis. The structural features of the genes and duplicated copies denote that wild *Oryza* species can be used as a rich source of variability that can improve starch quantity and quality, mainly through modifications of amylopectin synthesis chains [1].

Expressed specifically in the developing rice endosperm and leaves, SSIII 1 and 2 include three other repeated domains in addition to the starch synthase domain. An N-terminal Carbohydrate-Binding Module (CBM) domain is a contiguous amino acid sequence within a carbohydrate-active enzyme with carbohydrate-binding activity (Figure 3C,D). Although no lack of protein motifs was observed that could affect the catalytic domain in SSIII, in *O. sativa* this domain synthesizes long chains, and a deficiency in



SSIII1, that is, the second major enzyme [33], can indirectly enhance both the SS-I and GBSS-I gene transcripts. On the other hand, a survey of amino acid motifs of SS isoforms reveals that certain motifs are absent in certain *Oryza* species, as it is possible to notice in OsINDSSIV1, OLONSSIV1, OBARSSII2 and OMERSSII2, which are part of the two C-terminal domains. This may affect the catalytic performance of the chain-elongation reaction of  $\alpha$ -1-4-glucosidic linkage, which can further complicate the interplay between SS, SBE and DBE [34,35].

Waxy is believed to be the main enzyme that controls high amylose content in *Oryza* species and, with GBSSII, presents tissue-specific expression in a complementary manner between endosperm and non-endosperm tissues, causing different characteristics with respect to amylose content, and branch length distribution in amylopectin [36]. Thus, the differential action of these two enzymes affects the final amylose content in the endosperm. Despite this, the absence of GBSSII (Table S1) does not influence the high content of amylose in the endosperm (about 35%) of *O. meridionalis* [37]. Despite the evolutionary advantage that the presence of the two enzymes (Waxy and GBSSII) confer for starch biosynthesis, Waxy enzymes without GBSSII seem to be enough for high amylose accumulation in *Oryza* endosperm, something that brings new perspectives for the improvement of this complex network [15,36,38].

On the other hand, the loss of SSIV2 in *O. glaberrima* during evolution does not eliminate the ability of chloroplasts in producing starch granules, since features in the N-terminal extension of SSIV enable the interaction with other proteins contributing to granule initiation. In *Arabidopsis*, when the SSIV glucosyl transferase domain is absent, a significant reduction of starch synthesis is observed [39,40].

The only SS genes that show evidence of having undergone recombination are *ALK* and *Waxy*. This agrees with previous reports, in which the diversification in SS genes was suggested to be driven by many duplication events instead of recombination events [41].

### 3.3. Debranching Enzymes (DBE)

DPE1 is a protein part of Group 1, which comprises enzymes that act in the initial phase of endosperm development [9], playing an important role in grain quality improvement programs [20]. A total absence of DPE1 was observed in *O. nivara* and *O. brachyantha*. Although there is not much clarity about the performance of DPE1 in *Oryza* chloroplasts, it is known that *Arabidopsis* plants lacking the plastidic DPE1 accumulate maltooligosaccharides (maltotriose-maltoheptaose), but not maltose, an important carbohydrate in starch formation [42].

Completely different from DPE1, regarding its phylogenetic position and structure, but also showing an important influence in the final portion of the starch synthesis pathway in *Oryza*, the enzymes PUL and ISA catalyze different reactions, but both have a conserved gene structure. Although they play unique roles in regulating the crystallization and degradation of starch, the enzymes have a close relationship in *Oryza* and share, as expected, the N-terminal O-Glycosyl hydrolase (CBM\_48) and central domain alpha-amylase (Amy), in which both degrade amylopectin. However, in some species like *O. sativa* v.g. *japonica* and *O. longistaminata*, there is still an absence of the C-terminal domain DUF\_3372 domain (Figure 4D), which characterizes the Pullulanase, and usually cleaves the  $\alpha$ -1,6-linkages of polyglucans in pullulan. This absence may affect the final endosperm amylose content. The main gene that controls amylose is *Waxy*, but as starch synthesis is a fine regulatory network, together with other enzymes like *PUL*, *AGPase*, *SSI*, *ALK*, and *SSIII2*, they control the final content of amylose (AC). However, in the absence of pullulan degradation, the final starch content may be lower, and consequently the AC is lower as well [15]. Exactly what is perceived in the *O. sativa* ssp *japonica* genotypes is that they have amylose content around 10–22% (low AC) while *O. sativa* ssp *indica* show 18–32% (high AC) [43,44].

Regarding the events of changing gene positions in chromosome of DBE genes, NHR constitutes a relatively frequent event in *Oryza* genomes, one might think that MEI could also be the responsible for such duplication and translocation, since these events frequently



generate syntenic failures between homologue chromosomes when comparing different species [45]. Here we show (Figure S9) that it is not possible that MEI insertion could have occurred in these *PUL* genes. The same event could also have occurred in the other genes that have different chromosome positions (Figure 1). No recombination inference was found. However, 43 sites were observed to be under positive selection (Figure S10) in *DBE*, *ISA* and *PUL* gene copies phylogeny. Both Noug   et al. [41] and Qu et al [6] reported *DBE* homologue diversification through the detection of strong positive selection over these genes, once again denoting the complex evolutionary history of starch biosynthesis pathway.

### 3.4. Starch Branching Enzymes (SBE)

*Oryza* species present multiple SBE isoforms, more than shown here, but these are the major genes involved in the synthesis of amylopectin [20]. In modular architecture, both C and N termini play important roles in determining the substrate preference, catalytic capacity and chain length transfer [46]. The importance of SBE1 in synthesis of B1, B2, B3 chains of amylopectin has been reported in rice mutants [47,48], while others show that SBE3 has a role in the synthesis of 1–6 branching linkage [49]. In *Oryza*, these two enzymes are in the same clade. Some residues of binding sites for maltopentaose and glucose were not conserved between SBEI and SBEII isoforms; however, these residues were mainly found in SBEIII, which seems to be the reason for such a close proximity between SBE1 and SBE3 in *Oryza* [6].

## 4. Materials and Methods

### 4.1. DNA/RNA Sequencing and Gene Prediction

DNA sequences used in this study were obtained from the complete genome sequencing of *O. rufipogon* (Cultivar: W1943; Gramene accession: PRJEB4137), *O. nivara* (IRGC:100897; AWHD000000000), *O. glumaepatula* (GEN1233; ALNU000000000), *O. glaberrima* (IRGC:96717; ADWL000000000), *O. barthii* (IRGC:105608; ABR000000000), *O. meridionalis* (OR44 (W2112); ALNW000000000), *O. punctata* (IRGC:105690; AVCL000000000), *O. brachyantha* (IRGC:101232; AGAT000000000), *O. longistaminata* (unnamed accession; PRJNA545798) and *Leersia perrieri* (A. Camus) Launert (IRGC:105164; ALNV000000000). Sequencing, assembly and annotation of these species are part of the IOMAP initiative in which Next Generation Sequencing (NGS) was used for obtaining both genomic and transcriptomic sequences.

### 4.2. Identification of SSR Genes in the *Oryza* Genus

The Starch Synthesis-Related (SSRGs) used in this study were chosen according to Zeng et al., [22]. Initially, the SSRGs of *O. sativa* ssp. *japonica* were obtained through the RAP-DB (The Rice Annotation Project Database) (<https://rapdb.dna.affrc.go.jp/index.html> (accessed on 29 April 2021)). The similarity of these SSRGs of *O. sativa* ssp. *japonica* to sequences of other *Oryza* species was evaluated through the BLAST tool [50], available in the ENSEMBL PLANTS database (<http://plants.ensembl.org/index.html> (accessed on 29 April 2021)). Only SSRGs that had the high score (values > 0) associated with high coverage (values > 0) and low e-value (value ≤ 0) were selected for analysis (Table S1). Duplicated genes were identified with these same parameters but in this case in other locations in relation to the original gene. All genes and protein SSRG sequences are available in Supplementary Files S1 and S2.

### 4.3. Phylogenetic Analysis

The 19 resulting SSRGs were subjected to ClustalW [51] global alignment, generating an initial tree built through Neighbor-joining method [52] with 10,000 bootstrap replicates, with the aid of the Molecular Evolutionary Genetics Analysis 7—MEGA 7 [53] using the Gonnet matrix. The best replacement model was obtained through analysis in MEGA7. The appropriate model was selected for use in Bayesian analysis using Bayesian Evolution-

ary Sampling Trees—BEAST [54] with 1,000,000 bootstrap replicates. The resulting tree was plotted in FigTree (<http://tree.bio.ed.ac.uk/software/figtree/> (accessed on 29 April 2021)). *Leersia perrieri* was used as the outgroup. The ratio between non-synonymous and synonymous substitutions (dNS/dS) for each SSRG was estimated using the Single Likelihood Ancestor Counting (SLAC) method, to infer the evolutionary force at work in each Open Reading Frame (ORF). The MEME test (Mixed Effects Model of Evolution) was applied to detect branches that are proportionally in higher positive selection pressure based in the Likelihood ratio test for episodic diversification (LRT). To evaluate the presence of recombination in each gene partitions used in this study, the Genetic Algorithm for Detection of Recombination (GARD) was applied; such recombinant sequences can cause misinterpretation in the phylogenetic relationships because recombination selection inference often leads to a significant increase in false positives. All analyses are available at Datamonkey [55] with *p*-value threshold of 0.1. The AUGUSTUS annotation software [56] was used to construct the structure of SSRGs and for GSDS 2.0 visualization [57]. The identification of the SSRG protein motifs was performed using Multiple Motif In Elicitation version 4.11.1 (MEME; <http://meme-suite.org/tools/meme> (accessed on 29 April 2021)) [58], considering the maximum number of motifs equal to 10. Protein domain analysis was performed in SMART database (<http://smart.embl-heidelberg.de/> (accessed on 29 April 2021)) [59].

#### 4.4. Translocation Event Analysis

To understand the origin of *ONIVAGPS2a*, *ONIVISA*, *ONIVPUL*, *ONIVSSII2*, *OGLAPU L\_2D*, *OMERAGPS2a*, *OMERISA*, *OMERSSII1*, *OMERSSII2*, *OLONAGPL4* and *OLONSSIV2*, which could have occurred due to translocation events, we performed the alignment of regions corresponding to 50 Kb upstream and 50 Kb downstream from these genes in all the analyzed *Oryza* species, using Mauve [60]. We also used the RiTE database [61] to verify if this possible event occurred due to the translocation of transposable elements (TEs).

#### 4.5. Chromosome Position Analysis

A circular MapChart-based plot was created in which the location of each gene can be seen. For the correct positioning of the gene of each species on the respective single/common chromosome, a simple percentage calculation was performed in order to establish a proportion relation when comparing the location of the genes in homoeologous chromosomes, according to the following equation:

$$\frac{\text{gene location(bps)} \times 100}{\text{chr size(bps)}} = \text{Relative position of the gene}$$

### 5. Conclusions

In summary, we identified and characterized SSRG homologs in the wild relatives of Price. Using phylogenetics and comparative genomics analyses we offer insights for the use of their gene variations in plant breeding. We confirmed the relative conservation of SSRGs between species within the AA-, BB- and FF-genomes, but structural analysis of these proteins suggest that deletions/mutations of amino acids in some active sites can result in structural variation that may negatively affect specific phases of starch biosynthesis. Direct modification of the endosperm, as usually observed in *O. sativa* ssp. *japonica*, which possesses lower AC, can likely be related to the absence of PUL C-terminal domain. The complete deletion of some genes appears not to affect the final composition of starch in the endosperm, as observed for *GBSSII* in *O. meridionalis*, *SSIV2* in *O. glaberrima*, and *DPE1* in *O. brachyantha* and *O. nivara*.

The analysis of structural features points to both absence of and duplicated copies of some motifs that can modify metabolic activity, denoting that the use of different *Oryza* species can be a rich source of variability for starch-targeted improvement in rice. These genes should now be further investigated by phenotyping different mutants and through



the characterization of starch content of both wild *Oryza* genotypes and near isogenic lines (NILs) of *O. sativa* containing introgressions of these wild relatives. Such an analysis will help us to reveal the role of each variation of these genes, thereby contributing greatly to the simplification of the improvement processes that involve this complex path.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/plants10061057/s1>, Table S1: Table S1. Full information about each of the *Starch Synthesis-Related Genes* (SSRGs) analyzed. Figure S1: Positive selection and Mixed Effects Model of Evolution (MEME) in *AGPS-AGPL* genes. A total of 34 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT). Figure S2: Alignment of the promoter region (1500 bps upstream) of *AGPS2a* genes, showing possible translocation events between *ONIVAGPS2a* and *OMERAGPS2a*. Figure S3: Alignment of the promoter region (1500 bps upstream) of *SSIII1* genes, showing possible translocation events between *OMERSSIII1* and other *Oryza* species. Figure S4: Alignment of the promoter region (1500 bps upstream) of *SSII2* genes, showing possible translocation events between *OMERSSII2* and *ONIVSSII2* and other *Oryza* species. Figure S5: Alignment of the promoter region (1500 bps upstream) of *SSIV2* genes, showing possible translocation events in *OLONSSIV2* and other *Oryza* species. Figure S6: Genetic Algorithm for Recombination Detection (GARD) for *Waxy* genes. Comparing the AICc score of the best fitting GARD model, that allows for different topologies between segments (21,426.8), and that of the model that assumes the same tree for all the partitions inferred by GARD the same tree, but allows different branch lengths between partitions (21,430.9) suggests that because the multiple tree model cannot be preferred over the single tree model by an evidence ratio of 100 or greater, some or all of the breakpoints may reflect rate variation instead of topological incongruence. Figure S7: Genetic Algorithm for Recombination Detection (GARD) for *ALK* genes. Comparing the AICc score of the best fitting GARD model, that allows for different topologies between segments (32,200.2), and that of the model that assumes the same tree for all the partitions inferred by GARD the same tree, but allows different branch lengths between partitions (32,211.2) suggests that because the multiple tree model can be preferred over the single tree model by an evidence ratio of 100 or greater, at least of one of the breakpoints reflects a true topological incongruence. Figure S8: Positive selection and Mixed Effects Model of Evolution (MEME) in *SS* genes. A total of 117 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT). Figure S9: Alignment of the promoter region (1500 bps upstream) of *PUL* genes, showing possible translocation events in *OGLAPUL\_2D* and other *Oryza* species. Figure S10: Positive selection and Mixed Effects Model of Evolution (MEME) in *DBE*, *ISA* and *PUL* genes. A total of 117 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT). Figure S11: Positive selection and Mixed Effects Model of Evolution (MEME) in *SBE* genes. A total of 46 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT). File S1: Sequence of all SSRGs used in this study.

**Author Contributions:** K.E.J.d.F. conducted the major bioinformatics analyses, data interpretation and wrote the manuscript. R.S.d.S., J.L.L., R.A.W., and C.B. contributed to performing the research and revising the manuscript. F.d.C.V. conducted the selection pressure and in silico recombination analysis. A.C.d.O. conceived the study and supervised the research. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was financed by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) -Finance Code 001, by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and by the Fundação de Amparo a Pesquisa do Rio Grande do Sul (FAPERGS).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** We would like to thank the Arizona Genomics Institute and its School of Plant Sciences, Ecology & Evolutionary Biology for the providing data and for contributing to the review of details related to the use of these.



**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Pandey, M.K.; Rani, N.S.; Madhav, M.S.; Sundaram, R.M.; Varaprasad, G.S.; Sivaranjani, A.K.P.; Bohra, A.; Kumar, G.R.; Kumar, A. Different isoforms of starch-synthesizing enzymes controlling amylose and amylopectin content in rice (*Oryza sativa* L.). *Biotechnol. Adv.* **2012**, *30*, 1697–1706. [\[CrossRef\]](#)
- Yu, G.; Olsen, K.M.; Schaal, B.A. Molecular evolution of the endosperm starch synthesis pathway genes in rice (*Oryza sativa* L.) and its wild ancestor, *O. rufipogon* L. *Mol. Biol. Evol.* **2011**, *28*, 659–671. [\[CrossRef\]](#) [\[PubMed\]](#)
- Walter, M.; Marchezan, E.; Avila, L.A. Arroz: Composição e características nutricionais. *Ciênc. Rural* **2008**, *38*, 1184–1192. [\[CrossRef\]](#)
- dos Santos, R.S.; Farias, D.d.R.; Pegoraro, C.; Rombaldi, C.V.; Fukao, T.; Wing, R.A.; de Oliveira, A.C. Evolutionary analysis of the SUB1 locus across the *Oryza* genomes. *Rice* **2017**, *10*. [\[CrossRef\]](#)
- Stein, J.C.; Yu, Y.; Copetti, D.; Zwickl, D.J.; Zhang, L.; Zhang, C.; Chougule, K.; Gao, D.; Iwata, A.; Goicoechea, J.L.; et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **2018**, *50*, 285–296. [\[CrossRef\]](#)
- Qu, J.; Xu, S.; Zhang, Z.; Chen, G.; Zhong, Y.; Liu, L.; Zhang, R.; Xue, J.; Guo, D. Evolutionary, structural and expression analysis of core genes involved in starch synthesis. *Sci. Rep.* **2018**, *8*, 12736. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tuncel, A.; Kawaguchi, J.; Ihara, Y.; Matsusaka, H.; Nishi, A.; Nakamura, T.; Kuhara, S.; Hirakawa, H.; Nakamura, Y.; Cakir, B.; et al. The rice endosperm ADP-glucose pyrophosphorylase large subunit is essential for optimal catalysis and allosteric regulation of the heterotetrameric enzyme. *Plant Cell Physiol.* **2014**, *5*, 1169–1183. [\[CrossRef\]](#)
- Ohdan, T.; Francisco, P.B., Jr.; Sawada, T.; Hirose, T.; Terao, T.; Satoh, H.; Nakamura, Y. Expression profiling of genes involved in starch synthesis in sink and source organs of rice. *J. Exp. Bot.* **2005**, *56*, 3229–3244. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tian, Z.; Qian, Q.; Liu, Q.; Yan, M.; Liu, X.; Yan, C.; Liu, G.; Gao, Z.; Tang, S.; Zeng, D.; et al. Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. *Proc. Nat. Acad. Sci. USA* **2009**, *106*, 21760–21765. [\[CrossRef\]](#)
- Hirose, T.; Terao, T. A comprehensive expression analysis of the starch synthase gene family in rice (*Oryza sativa* L.). *Planta* **2004**, *220*, 9–16. [\[CrossRef\]](#)
- Fujita, N.; Toyosawa, Y.; Utsumi, Y.; Higuchi, T.; Hanashiro, I.; Ikegami, A.; Akuzawa, S.; Yoshida, M.; Mori, A.; Inomata, K.; et al. Characterization of pullulanase (PUL)-deficient mutants of rice (*Oryza sativa* L.) and the function of PUL on starch biosynthesis in the developing rice endosperm. *J. Exp. Bot.* **2009**, *60*, 1009–1023. [\[CrossRef\]](#) [\[PubMed\]](#)
- Tsumi, Y.; Utsumi, C.; Sawada, T.; Fujita, N.; Nakamura, Y. Functional diversity of isoamylose oligomers: The ISA1 homo-oligomer is essential for amylopectin biosynthesis in rice endosperm. *Plant Physiol.* **2011**, *156*, 61–77. [\[CrossRef\]](#)
- Yamanaka, S.; Nakamura, I.; Watanabe, K.N.; Sato, Y.I. Identification of SNPs in the waxy gene among glutinous rice cultivars and their evolutionary significance during the domestication process of rice. *Theor. Appl. Genet.* **2004**, *108*, 1200–1204. [\[CrossRef\]](#)
- Abe, N.; Asai, H.; Yago, H.; Oitome, N.F.; Itoh, R.; Crofts, N.; Nakamura, Y.; Fujita, N. Relationships between starch synthase I and branching enzyme isozymes determined using double mutant rice lines. *BMC Plant Biol.* **2014**, *14*, 80. [\[CrossRef\]](#) [\[PubMed\]](#)
- Beckles, D.M. Use of biotechnology to engineer starch in cereals. In *Encyclopedia of Biotechnology in Agriculture and Food*; Emerald Group Publishing Limited: Bradford, UK, 2010; pp. 1–9. [\[CrossRef\]](#)
- Regina, A.; Li, Z.; Morell, M.K.; Jobling, S.A. *Genetically Modified Starch: State of Art and Perspectives*; Elsevier: Edinburgh, UK, 2014; pp. 13–29. [\[CrossRef\]](#)
- Sun, Y.; Jiao, G.; Liu, Z.; Zhang, X.; Li, J.; Guo, X.; Du, W.; Du, J.; Francis, F.; Zhao, Y.; et al. Generation of high-amylose rice through CRISPR/Cas9-mediated targeted mutagenesis of starch branching enzymes. *Front. Plant Sci.* **2010**, *8*, 298. [\[CrossRef\]](#)
- Shufen, C.; Yicong, C.; Baobing, F.; Guiai, J.; Zhonghua, S.; Ju, L.; Shaoqing, T.; Jianlong, W.; Peisong, H.; Xiangjin, W. Editing of rice isoamylase gene ISA1 provides insights into its function in starch formation. *Rice Sci.* **2019**, *26*, 77–87. [\[CrossRef\]](#)
- Xu, Z.; Yu, M.; Yin, Y.; Zhu, C.; Ji, W.; Zhang, C.; Li, Q.; Zhang, H.; Tang, S.; Yu, H.; et al. Generation of selectable marker-free soft transgenic rice with transparent kernels by downregulation of SSSII-2. *Crop J.* **2020**, *8*, 53–61. [\[CrossRef\]](#)
- Zeng, D.; Tian, Z.; Rao, Y.; Dong, G.; Yang, Y.; Huang, L.; Leng, Y.; Xu, J.; Sun, C.; Zhang, G.; et al. Rational design of high-yield and superior-quality rice. *Nat. Plants* **2017**, *3*, 17031. [\[CrossRef\]](#)
- Smith, A.M.; Denyer, K.; Martin, C. The synthesis of the starch granule. *Plant Mol. Biol.* **1997**, *48*, 67–87. [\[CrossRef\]](#)
- Batra, R.; Saripalli, G.; Mohan, A.; Gupta, S.; Gill, K.S.; Varadwaj, P.K.; Balyan, H.S.; and Gupta, P.K. Comparative Analysis of AGPase Genes and Encoded Proteins in Eight Monocots and Three Dicots with Emphasis on Wheat. *Front. Plant Sci.* **2017**, *8*, 19. [\[CrossRef\]](#)
- Georgelis, N.; Braun, E.L.; Shaw, J.R.; Hannah, C.L. The two AGPase subunits evolve at different rates in angiosperms, yet they are equally sensitive to activity-altering amino acid changes when expressed in bacteria. *Plant Cell* **2007**, *19*, 1458–1472. [\[CrossRef\]](#)
- Georgelis, N.; Braun, E.L.; Hannah, L.C. Duplications and functional divergence of ADP-glucose pyrophosphorylase genes in plants. *BMC Evol. Biol.* **2008**, *8*, 232. [\[CrossRef\]](#)
- Bai, Z.; Chen, J.; Liao, Y.; Wang, M.; Liu, R.; Ge, S.; Wing, R.A.; Chen, M. The impact and origin of copy number variations in the *Oryza* species. *BMC Genom.* **2016**, *17*, 261. [\[CrossRef\]](#)



26. Corbi, J.; Dutheil, J.Y.; Damerval, C.; Tenaillon, M.I.; Manicacci, D. Accelerated evolution and coevolution drove the evolutionary history of AGPase subunits during angiosperm radiation. *Ann. Bot.* **2012**, *109*, 693–708. [\[CrossRef\]](#)
27. Freeling, M.; Lyons, E.; Pedersen, B.; Alam, M.; Ming, R.; Lisch, D. Many or most genes in Arabidopsis transposed after the origin of the order Brassicales. *Genome Res.* **2008**, *18*, 1924–1937. [\[CrossRef\]](#)
28. Patron, N.J.; Keeling, P.J. Common Evolutionary Origin of Starch Biosynthetic Enzymes in Green and Red Algae. *J. Phycol.* **2005**, *41*, 1131–1141. [\[CrossRef\]](#)
29. Deschamps, P.; Colleoni, C.; Nakamura, Y.; Suzuki, E.; Putaux, J.-L.; Buléon, A.; Haebel, S.; Ritte, G.; Steup, M.; Falcón, L.I.; et al. Metabolic Symbiosis and the Birth of the Plant Kingdom. *Mol. Biol. Evol.* **2008**, *25*, 536–548. [\[CrossRef\]](#)
30. Ball, S.; Colleoni, C.; Cenci, U.; Raj, J.N.; Tirtiaux, C. The evolution of glycogen and starch metabolism in eukaryotes gives molecular clues to understand the establishment of plastid endosymbiosis. *J. Exp. Bot.* **2011**, *62*, 1775–1801. [\[CrossRef\]](#)
31. Guo, H.; Jiao, Y.; Tan, X.; Wang, X.; Huang, X.; Jin, H.; Paterson, A.H. Gene duplication and genetic innovation in cereal genomes. *Genome Res.* **2019**, *29*, 261–269. [\[CrossRef\]](#)
32. Kasem, S.; Waters, D.L.E.; Rice, N.F.; Shapter, F.M.; Henry, R.J. The endosperm morphology of rice and its wild relatives as observed by scanning electron microscopy. *Rice* **2011**, *4*, 12–20. [\[CrossRef\]](#)
33. Fujita, N.; Yoshida, M.; Asakura, N.; Ohdan, T.; Miyao, A.; Hirochika, H.; Nakamura, Y. Function and characterization of starch synthase using mutants in rice. *Plant Physiol.* **2006**, *140*, 1070–1084. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Myers, A.M.; Morell, M.K.; James, M.G.; Ball, S.G. Recent progress toward understanding the biosynthesis of the amylopectin crystal. *Plant Physiol.* **2000**, *122*, 989–998. [\[CrossRef\]](#)
35. Nakamura, Y. Towards a better understanding of the metabolic system for amylopectin biosynthesis in plants: Rice endosperm as a model tissue. *Plant Cell Physiol.* **2002**, *43*, 718–725. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Wang, W.; Wei, X.; Jiao, G.; Chen, W.; Wu, Y.; Sheng, Z.; Hu, S.; Xie, L.; Wang, J.; Tang, S.; et al. GBSS-BINDING PROTEIN, encoding a CBM48 domain-containing protein, affects rice quality and yield. *J. Int. Plant Biol.* **2019**. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Mondal, T.K.; Henry, R.J. *The Wild Oryza Genomes*; Springer: Berlin/Heidelberg, Germany, 2018.
38. Vrinten, P.L.; Nakamura, T. Wheat granule-bound starch synthase I and II are encoded by separate genes that are expressed in different tissues. *Plant Physiol.* **2000**, *122*, 255–264. [\[CrossRef\]](#)
39. Szydlowski, N.; Ragel, P.; Raynaud, S.; Lucas, M.M.; Roldan, I.; Montero, M.; Muñoz, F.J.; Ovecka, M.; Bahaji, A.; Planchot, V.; et al. Starch granule initiation in Arabidopsis requires the presence of either class IV or class III starch synthases. *Plant Cell* **2009**, *21*, 2443–2457. [\[CrossRef\]](#)
40. Zeeman, S.C.; Kossmann, J.; Smith, A.M. Starch: Its metabolism, evolution, and biotechnological modification in plants. *Ann. Rev. Plant Biol.* **2010**, *61*, 209–234. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Nougé, O.; Corbi, J.; Ball, S.G.; Manicacci, D.; Tenaillon, M. Molecular evolution accompanying functional divergence of duplicated genes along the plant starch biosynthesis pathway. *BMC Evol. Biol.* **2014**, *14*, 103. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Critchley, J.H.; Zeeman, S.C.; Takaha, T.; Smith, A.M.; Smith, S.M. A critical role for disproportionating enzyme in starch breakdown is revealed by a knock-out mutation in *Arabidopsis*. *Plant J.* **2001**, *26*, 89–100. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Lang, N.T.; Buu, B.C. Quantitative analysis on amylase content by DNA markers through backcross populations of rice (*Oryza sativa* L.). *Omonrice* **2004**, *12*, 13–18.
44. Ayabe, S.; Kasai, M.; Ohishi, K.; Hatae, K. Textural properties and structures of starches from *indica* and *japonica* rice with similar amylose content. *Food Sci. Technol. Res.* **2009**, *15*, 299–306. [\[CrossRef\]](#)
45. Ammiraju, J.S.; Lu, F.; Sanyal, A.; Yu, Y.; Song, X.; Jiang, N.; Pontaroli, A.C.; Rambo, T.; Currie, J.; Collura, K.; et al. Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell* **2008**, *20*, 3191–3209. [\[CrossRef\]](#)
46. Kuriki, T.; Stewart, D.C.; Preiss, J. Construction of chimeric enzymes out of maize endosperm branching enzymes I and II: Activity and properties. *J. Biol. Chem.* **1991**, *272*, 28999–29004. [\[CrossRef\]](#)
47. Satoh, H.; Nishi, A.; Fujita, N.; Kubo, A.; Nakamura, Y.; Kawasaki, T.; Okita, T.W. Isolation and characterization of starch mutants in rice. *J. Appl. Glycosci.* **2003**, *50*, 225–230. [\[CrossRef\]](#)
48. Satoh, H.; Nishi, A.; Yamashita, K.; Takemoto, Y.; Tanaka, Y.; Hosaka, Y.; Sakurai, A.; Fujita, N.; Nakamura, Y. Starch-branching enzyme I-deficient mutation specifically affects the structure and properties of starch in rice endosperm. *Plant Physiol.* **2003**, *133*, 1111–1121. [\[CrossRef\]](#)
49. Chen, M.H.; Huang, L.F.; Li, H.M.; Chen, Y.R.; Yu, S.M. Signal peptide-dependent targeting of a rice  $\alpha$ -amylase and cargo proteins to plastids and extracellular compartments of plant cells. *Plant Physiol.* **2004**, *135*, 1367–1377. [\[CrossRef\]](#)
50. Altschul, S.F.; Gish, W.; Miller, W.; Meyers, E.W.; Lipman, D.J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [\[CrossRef\]](#)
51. Larkin, M.A.; Blackshields, G.; Brown, N.; Chenna, R.; McGettigan, P.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947–2948. [\[CrossRef\]](#)
52. Saitou, N.; Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **1987**, *4*, 406–425.
53. Kumar, S.; Stecher, G.; Tamura, K. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **2016**, *33*, 1870–1874. [\[CrossRef\]](#) [\[PubMed\]](#)

- 
54. Drummond, A.J.; Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **2007**, *7*, 214. [[CrossRef](#)] [[PubMed](#)]
  55. Weaver, S.; Shank, S.D.; Spielman, S.J.; Li, M.; Muse, S.V.; Pond, S.L.K. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol. Biol. Evol.* **2018**, *35*, 773–777. [[CrossRef](#)] [[PubMed](#)]
  56. Stanke, M.; Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **2005**, *33*, W465–W467. [[CrossRef](#)]
  57. Hu, B.; Jin, J.; Guo, A.Y.; Zhang, H.; Luo, J.; Gao, G. GSDS 2.0: An upgraded gene feature visualization server. *Bioinformatics* **2015**, *31*, 1296–1297. [[CrossRef](#)]
  58. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [[CrossRef](#)]
  59. Letunic, I.; Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **2018**, *46*, D493–D496. [[CrossRef](#)] [[PubMed](#)]
  60. Darling, A.E.; Mau, B.; Perna, N.T. ProgressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **2010**, *5*, e11147. [[CrossRef](#)] [[PubMed](#)]
  61. Copetti, D.; Zhang, J.; El Baidouri, M.; Gao, D.; Wang, J.; Barghini, E.; Cossu, R.M.; Angelova, A.; Maldonado, C.E.; Roffler, S.; et al. RiTE database: A resource database for genus-wide rice genomics and evolutionary biology. *BMC Genom.* **2015**, *16*, 538. [[CrossRef](#)]



## Supplementary materials:

**Table S1.** Full information about each of the Starch Synthesis-Related Genes (SSRGs) analyzed.

Locus name	# of Transcripts	RAP-DB/Ensembl ID	Location	Start	End	Size (pb)	ID	E-value
<i>AGPL1</i>	1	<i>Os05g0580000</i>	Chr. 5	28871811	28877272	5462		
<i>AGPL3</i>	2	<i>Os03g0735000</i>	Chr. 3	30099369	30104572	5204		
<i>AGPL4</i>	1	<i>Os07g0243200</i>	Chr. 7	7983125	7989283	6159		
<i>AGPS2a</i>	4	<i>Os08g0345800</i>	Chr. 8	15666389	15672583	6248		
<i>ALK</i>	1	<i>Os06g0229800</i>	Chr. 6	6748398	6753302	4916		
<i>DPEI</i>	1	<i>Os07g0627000</i>	Chr. 7	25980423	25984853	4431		
<i>GBSSII</i>	2	<i>Os07g0412100</i>	Chr. 7	12916883	12924202	7320		
<i>ISA</i>	1	<i>Os08g0520900</i>	Chr. 8	25893657	25900576	6920		
<i>PUL</i>	1	<i>Os04g0164900</i>	Chr. 4	4408357	4418889	10533		
<i>SBE1</i>	4	<i>Os06g0726400</i>	Chr. 6	30897378	30905803	8426		
<i>SBE3</i>	1	<i>Os02g0528200</i>	Chr. 2	19355790	19367127	11338		
<i>SSI</i>	1	<i>Os06g0160700</i>	Chr. 6	3079296	3086808	7513		
<i>SSII 1</i>	1	<i>Os10t0437600</i>	Chr. 10	15673243	15681075	7833		
<i>SSII 2</i>	2	<i>Os02g0744700</i>	Chr. 2	31233292	31238210	4929		
<i>SSIII 1</i>	2	<i>Os04g0624600</i>	Chr. 4	31751600	31759420	7821		
<i>SSIII 2</i>	1	<i>Os08g0191433</i>	Chr. 8	5353697	5363276	9580		
<i>SSIV 1</i>	4	<i>Os01g0720600</i>	Chr. 1	30032428	30041425	8998		
<i>SSIV 2</i>	2	<i>Os05g0533600</i>	Chr. 5	26485770	26493983	8214		
<i>Wx</i>	2	<i>Os06g0133000</i>	Chr. 6	1765622	1770653	5032		
<i>AGPL1</i>	1	<i>BGIOSGA017490</i>	Chr. 5	30220701	30222552	3388	99.9	0.0
<i>AGPL3</i>	1	<i>BGIOSGA009855</i>	Chr. 3	34318562	34322802	4241	99.6	0.0
<i>AGPL4</i>	1	<i>BGIOSGA024540</i>	Chr. 7	7995391	7999961	4571	99.6	0.0
<i>AGPS2a</i>	1	<i>BGIOSGA027135</i>	Chr. 8	16750552	16746477	4076	99.5	0.0
<i>ALK</i>	1	<i>BGIOSGA022586</i>	Chr. 6	7778365	7782784	4420	99.9	0.0
<i>DPEI</i>	1	<i>BGIOSGA026185</i>	Chr. 7	24042478	24046564	4087	99.7	0.0
<i>GBSSII</i>	1	<i>BGIOSGA024424</i>	Chr. 7	12148263	12152181	3919	99.2	0.0
<i>ISA</i>	1	<i>BGIOSGA026650</i>	Chr. 8	27633755	27640589	6835	99.9	0.0
<i>PUL</i>	1	<i>BGIOSGA015875</i>	Chr. 4	3476296	3488775	12480	98.2	0.0
<i>SBE1</i>	1	<i>BGIOSGA020506</i>	Chr. 6	32596001	32603277	7277	99.8	0.0
<i>SBE3</i>	1	<i>BGIOSGA006344</i>	Chr. 2	20764778	20775734	10957	99.7	0.0
<i>SSI</i>	1	<i>BGIOSGA021860</i>	Chr. 6	3609454	3616267	6814	100.0	0.0
<i>SSII 1</i>	1	<i>BGIOSGA033011</i>	Chr. 10	14120504	14127623	7120	99.6	0.0
<i>SSII 2</i>	1	<i>BGIOSGA005631</i>	Chr. 2	33274118	33278322	4215	99.8	0.0
<i>SSIII 1</i>	1	<i>BGIOSGA014316</i>	Chr. 4	30784815	30792005	7186	99.9	0.0
<i>SSIII 2</i>	1	<i>BGIOSGA028122</i>	Chr. 8	5614172	5624994	10823	99.7	0.0
<i>SSIV 1</i>	1	<i>BGIOSGA000900</i>	Chr. 1	33261071	33269378	8308	99.7	0.0
<i>SSIV 2</i>	1	<i>BGIOSGA020250</i>	Chr. 5	27869304	27876940	7637	99.7	0.0
<i>Wx</i>	1	<i>BGIOSGA022241</i>	Chr. 6	1931535	1935014	3480	100.0	0.0
<i>AGPL1</i>	1	<i>ORUF105G29020</i>	Chr. 5	25558424	25561810	3387	100.0	0.0
<i>AGPL3</i>	1	<i>ORUF103G34620</i>	Chr. 3	28543622	28547924	4303	99.8	0.0
<i>AGPL4</i>	1	<i>ORUF107G08660</i>	Chr. 7	7250572	7256218	5647	99.9	0.0
<i>AGPS2a</i>	2	<i>ORUF108G13010</i>	Chr. 8	13811400	13817275	5876	100.0	0.0
<i>ALK</i>	1	<i>ORUF106G08580</i>	Chr. 6	6231488	6236538	5051	99.5	0.0
<i>DPEI</i>	2	<i>ORUF107G24050</i>	Chr. 7	22751548	22756102	4555	100	0.0
<i>GBSSII</i>	1	<i>ORUF107G11630</i>	Chr. 7	11317247	11324432	7186	99.1	0.0
<i>ISA</i>	1	<i>ORUF108G23320</i>	Chr. 8	23528272	23535101	6830	99.9	0.0
<i>PUL</i>	1	<i>ORUF104G02940</i>	Chr. 4	3519334	3532236	12903	98.3	0.0
<i>SBE1</i>	5	<i>ORUF106G29940</i>	Chr. 6	27526184	27533602	7419	99.8	0.0

<i>SBE3</i>	1	<i>ORUF102G19870</i>	Chr. 2	18185004	18195898	10895	100.0	0.0
<i>SSI</i>	1	<i>ORUF106G04040</i>	Chr. 6	2791984	2798799	6816	99.9	0.0
<i>SSII 1</i>	3	<i>ORUF110G11670</i>	Chr. 10	13668717	13679236	10520	99.7	0.0
<i>SSII 2</i>	2	<i>ORUF102G33990</i>	Chr. 2	29295244	29299474	4231	99.8	0.0
<i>SSIII 1</i>	1	<i>ORUF104G27570</i>	Chr. 4	27123710	27130905	7196	99.9	0.0
<i>SSIII 2</i>	3	<i>ORUF108G05900</i>	Chr. 8	4762126	4775482	13357	99.4	0.0
<i>SSIV 1</i>	3	<i>ORUF101G32260</i>	Chr. 1	27539416	27547732	8317	99.9	0.0
<i>SSIV 2</i>	2	<i>ORUF105G25430</i>	Chr. 5	23281640	23290019	8380	99.8	0.0
<i>Wx</i>	2	<i>ORUF106G02030</i>	Chr. 6	1559398	1564292	4895	99.9	0.0
<i>AGPL1</i>	2	<i>ONIVA05G29110</i>	Chr. 5	26996830	26999189	5239	100.0	0.0
<i>AGPL3</i>	1	<i>ONIVA03G34930</i>	Chr. 3	30015137	30019384	4248	99.6	0.0
<i>AGPL4</i>	1	<i>ONIVA07G07390</i>	Chr. 7	6096533	6102180	5648	99.7	0.0
<i>AGPS2a</i>	2	<i>ONIVA04G29180</i>	Chr. 4	27995462	28001390	5929	99.5	0.0
<i>ALK</i>	1	<i>ONIVA06G09520</i>	Chr. 6	6979112	6984349	5238	99.4	0.0
<i>DPEI</i>	<b>UNKNOWN</b>							
<i>GBSSII</i>	2	<i>ONIVA07G09560</i>	Chr. 7	9331482	9338655	7174	99.8	0.0
<i>ISA</i>	2	<i>ONIVA01G02600</i>	Chr. 1	1822103	1830976	8874	99.7	0.0
<i>PUL</i>	3	<i>ONIVA08G18600</i>	Chr. 8	20035949	20048691	12743	99.8	0.0
<i>SBE1</i>	6	<i>ONIVA06G30960</i>	Chr. 6	28851472	28859303	7832	99.8	0.0
<i>SBE3</i>	1	<i>ONIVA02G20920</i>	Chr. 2	19160312	19171267	10956	99.8	0.0
<i>SSI</i>	1	<i>ONIVA06G20230</i>	Chr. 6	18799660	18806464	6805	99.4	0.0
<i>SSII 1</i>	3	<i>ONIVA10G10320</i>	Chr. 10	11413747	11424234	10488	99.7	0.0
<i>SSII 2</i>	2	<i>ONIVA06G04720</i>	Chr. 6	3161289	3166670	5391	99.5	0.0
<i>SSIII 1</i>	2	<i>ONIVA04G24880</i>	Chr. 4	24741542	24749237	7696	99.9	0.0
<i>SSIII 2</i>	2	<i>ONIVA08G05260</i>	Chr. 8	4445515	4458770	13256	99.7	0.0
<i>SSIV 1</i>	3	<i>ONIVA01G33370</i>	Chr. 1	29128063	29136375	8313	99.9	0.0
<i>SSIV 2</i>	1	<i>ONIVA05G24550</i>	Chr. 5	24014997	24023064	8040	99.7	0.0
<i>Wx</i>	1	<i>ONIVA06G02500</i>	Chr. 6	1715144	1719189	4046	99.9	0.0
<i>AGPL1</i>	1	<i>OGLUM05G28600</i>	Chr. 5	29490471	29492029	3389	99.7	0.0
<i>AGPL3</i>	1	<i>OGLUM03G32900</i>	Chr. 3	31084033	31088270	4238	99.6	0.0
<i>AGPL4</i>	1	<i>OGLUM07G07800</i>	Chr. 7	7504668	7510317	5650	99.6	0.0
<i>AGPS2a</i>	1	<i>OGLUM08G12560</i>	Chr. 8	13673152	13677206	4055	99.8	0.0
<i>ALK</i>	1	<i>OGLUM06G08830</i>	Chr. 6	6894230	6899271	5050	99.5	0.0
<i>DPEI</i>	2	<i>OGLUM07G22950</i>	Chr. 7	24895289	24899871	4583	99.7	0.0
<i>GBSSII</i>	2	<i>OGLUM07G10960</i>	Chr. 7	12170159	12177340	7182	99.7	0.0
<i>ISA</i>	1	<i>OGLUM08G22090</i>	Chr. 8	24554177	24560987	6811	99.9	0.0
<i>PUL</i>	3	<i>OGLUM04G01860</i>	Chr. 4	2886872	2900509	13638	98.5	0.0
<i>SBE1</i>	7	<i>OGLUM06G29370</i>	Chr. 6	31128403	31137352	8950	99.8	0.0
<i>SBE3</i>	1	<i>OGLUM02G19180</i>	Chr. 2	20189625	20200539	10915	99.7	0.0
<i>SSI</i>	1	<i>OGLUM06G04130</i>	Chr. 6	2892170	2898982	6821	99.1	0.0
<i>SSII 1</i>	2	<i>OGLUM10G10900</i>	Chr. 10	14772075	14780343	8269	97.4	0.0
<i>SSII 2</i>	2	<i>OGLUM02G32940</i>	Chr. 2	32627141	32631778	4638	99.5	0.0
<i>SSIII 1</i>	3	<i>OGLUM04G25800</i>	Chr. 4	28951767	28958958	7192	98.0	0.0
<i>SSIII 2</i>	2	<i>OGLUM08G05570</i>	Chr. 8	4670325	4683755	13441	99.4	0.0
<i>SSIV 1</i>	2	<i>OGLUM01G33230</i>	Chr. 1	32640551	32648862	8312	99.5	0.0
<i>SSIV 2</i>	2	<i>OGLUM05G25180</i>	Chr. 5	26994912	27002552	7641	98.2	0.0
<i>Wx</i>	2	<i>OGLUM06G02020</i>	Chr. 6	1469264	1474035	4772	99.9	0.0
<i>AGPL1</i>	1	<i>ORGLA05G0234700</i>	Chr. 5	22419555	22421766	3384	98.9	0.0
<i>AGPL3</i>	1	<i>ORGLA03G0301700</i>	Chr. 3	27225984	27230182	4199	99.7	0.0
<i>AGPL4</i>	1	<i>ORGLA07G0073000</i>	Chr. 7	6872607	6877182	4516	99.8	0.0
<i>AGPS2a</i>	1	<i>ORGLA08G0104300</i>	Chr. 8	11519446	11523524	5891	99.5	0.0
<i>ALK</i>	1	<i>ORGLA06G0078600</i>	Chr. 6	6002051	6006473	4423	99.6	0.0
<i>DPEI</i>	1	<i>ORGLA07G0179000</i>	Chr. 7	18641028	18645164	4137	98.2	0.0
<i>GBSSII</i>	1	<i>ORGLA07G0097800</i>	Chr. 7	10557492	10561434	3943	99.5	0.0

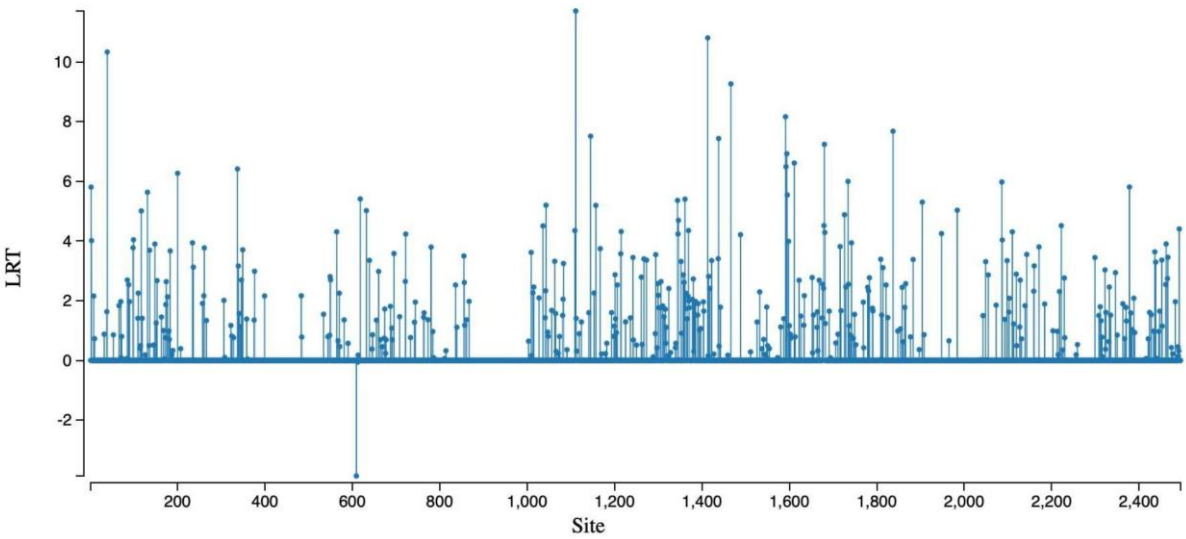
ISA	1	ORGLA08G0184300	Chr. 8	19624422	19631228	6807	99.7	0.0
PUL	1	ORGLA04G0016300	Chr. 4	2098830	2109127	10308	99.6	0.0
	1	ORGLA06G0243400	Chr. 6	23629078	23642103	11836	99.4	0.0
SBE1	1	ORGLA06G0237000	Chr. 6	23121600	23127808	6209	99.6	0.0
SBE3	1	ORGLA02G0162800	Chr. 2	15767001	15777902	10902	99.8	0.0
SSI	1	ORGLA06G0037300	Chr. 6	2672463	2679274	6812	99.5	0.0
SSII 1	1	ORGLA10G0098400	Chr. 10	12560982	12568108	7127	99.8	0.0
SSII 2	1	ORGLA02G0278400	Chr. 2	25119935	25124140	4206	99.5	0.0
SSIII 1	1	ORGLA04G0229100	Chr. 4	23494796	23501988	7193	99.7	0.0
SSIII 2	1	ORGLA08G0045500	Chr. 8	3902326	3913414	11089	99.7	0.0
SSIV 1	1	ORGLA01G0239700	Chr. 1	21752449	21760765	8317	99.6	0.0
SSIV 2		UNKNOWN						
Wx	1	ORGLA06G0020500	Chr. 6	1582446	1585771	3326	99.6	0.0
AGPL1	1	OBART05G27050	Chr. 5	23324094	23326305	3384	99.8	0.0
AGPL3	1	OBART03G33230	Chr. 3	27626346	27630535	4190	99.7	0.0
AGPL4	1	OBART07G08590	Chr. 7	7013571	7017866	4296	99.8	0.0
AGPS2a	2	OBART08G11760	Chr. 8	11861898	11867788	5891	99.3	0.0
ALK	1	OBART06G08250	Chr. 6	6050736	6055714	4979	99.6	0.0
DPEI	2	OBART07G22910	Chr. 7	20754611	20758640	4030	98.2	0.0
GBSSII	2	OBART07G11410	Chr. 7	10578751	10585992	7142	99.5	0.0
ISA	1	OBART08G20830	Chr. 8	20404034	20410832	6699	99.7	0.0
PUL	1	OBART04G02210	Chr. 4	2180683	2193763	13081	99.6	0.0
SBE1	2	OBART06G27990	Chr. 6	25397160	25403370	6211	99.6	0.0
SBE3	1	OBART02G18870	Chr. 2	17066878	17077468	10591	99.7	0.0
SSI	1	OBART06G04010	Chr. 6	2754467	2761274	6808	99.4	0.0
SSII 1	2	OBART10G11180	Chr. 10	12596913	12604730	7818	99.8	0.0
SSII 2	1	OBART02G32300	Chr. 2	27493261	27497466	4206	99.5	0.0
SSIII 1	1	OBART04G26440	Chr. 4	24614479	24621672	7194	99.6	0.0
SSIII 2	3	OBART08G05320	Chr. 8	4222533	4234438	11906	99.6	0.0
SSIV 1	2	OBART01G29190	Chr. 1	25014283	25022599	8317	99.6	0.0
SSIV 2	4	OBART05G23810	Chr. 5	21216867	21225276	8310	99.5	0.0
Wx	1	OBART06G02000	Chr. 6	1521439	1525491	4094	99.6	0.0
AGPL1	4	OMERI05G23130	Chr. 5	25687807	25689281	7452	98.6	0.0
AGPL3	3	OMERI03G29830	Chr. 3	31404907	31420237	15331	98.7	0.0
AGPL4	1	OMERI07G06990	Chr. 7	7300131	7308247	8117	99.8	0.0
AGPS2a	1	OMERI09G03640	Chr. 9	5815979	5826226	10248	87.4	4.3E-71
ALK	1	OMERI06G09850	Chr. 6	8591311	8596588	5278	98.6	0.0
DPEI	1	OMERI07G19310	Chr. 7	22508514	22522342	13829	99.0	0.0
GBSSII		UNKNOWN						
ISA	1	OMERI02G01830	Chr. 2	1481268	1491073	9806	99.6	0.0
PUL	3	OMERI04G02160	Chr. 4	3050487	3064210	13418	98.5	0.0
SBE1	6	OMERI06G27930	Chr. 6	31435504	31444602	9099	99.4	0.0
SBE3	1	OMERI02G18870	Chr. 2	20742387	20758950	16564	99.0	0.0
SSI	1	OMERI06G04610	Chr. 6	3672873	3680742	7870	97.1	0.0
SSII 1	2	OMERI04G07060	Chr. 4	11369715	11377827	8113	98.7	0.0
SSII 2	1	OMERI06G09850	Chr. 6	8591311	8596588	5278	88.7	1.3E-142
SSIII 1	1	OMERI04G21470	Chr. 4	26281268	26296875	15151	97.9	0.0
	1	OMERI04G21490	Chr. 4	26300044	26307350	7307	97.9	0.0
SSIII 2	1	OMERI08G05080	Chr. 8	4840603	4852759	12157	99.5	0.0
SSIV 1	1	OMERI01G26630	Chr. 1	27952587	27955762	3176	97.1	0.0
	4	OMERI01G26980	Chr. 1	28230264	28240799	10536	97.1	0.0
SSIV 2	1	OMERI05G21310	Chr. 5	24160650	24175895	14194	99.6	0.0
Wx	3	OMERI06G01920	Chr. 6	1676732	1682574	5235	97.9	0.0
AGPL1	1	OPUNC05G24670	Chr. 5	30120865	30121813	3375	96.7	0.0



AGPL3	1	OPUNC03G30430	Chr. 3	32831814	32836265	4452	95.3	0.0
AGLP4	1	OPUNC07G08120	Chr. 7	8462242	8467841	5600	94.3	0.0
AGPS2a	2	OPUNC08G10720	Chr. 8	15575237	15580527	5291	97.6	0.0
ALK	1	OPUNC06G08000	Chr. 6	6492378	6497173	4796	96.9	0.0
DPEI	1	OPUNC07G21530	Chr. 7	27721637	27726637	5001	89.0	2.0E-135
GBSSII	1	OPUNC07G10360	Chr. 7	13869752	13873662	3911	96.1	0.0
ISA	1	OPUNC08G18890	Chr. 8	27058634	27065585	6952	95.9	0.0
PUL	1	OPUNC04G02180	Chr. 4	3369443	3383422	13980	95.6	0.0
SBE1	4	OPUNC06G25520	Chr. 6	34223277	34229751	6475	97.5	0.0
SBE3	2	OPUNC02G16910	Chr. 2	21616213	21627157	10945	94.1	0.0
SSI	2	OPUNC06G03820	Chr. 6	2602107	2608648	6542	97.9	0.0
SSII 1	3	OPUNC10G09440	Chr. 10	18391785	18399027	7243	96.2	0.0
SSII 2	1	OPUNC02G29850	Chr. 2	34702141	34710291	8151	98.5	0.0
SSIII 1	1	OPUNC04G23440	Chr. 4	30084234	30092699	8466	95.1	0.0
SSIII 2	1	OPUNC08G05170	Chr. 8	4642721	4654350	11630	97.1	0.0
SSIV 1	1	OPUNC01G29020	Chr. 1	32623724	32631889	8166	94.4	0.0
SSIV 2	1	OPUNC05G21230	Chr. 5	27633570	27640939	7370	96.8	0.0
Wx	1	OPUNC06G01860	Chr. 6	1254773	1258390	3618	96.1	5.4E-170
AGPL1	1	OB05G34620	Chr. 5	19404719	19405022	3789	93.4	2.1E-122
AGPL3	1	OB03G40320	Chr. 3	24403484	24407693	4210	89.1	1.1E-74
AGPL4	1	OB07G16320	Chr. 7	4817190	4823741	6552	94.0	1.3E-71
AGPS2a	1	OB08G20190	Chr. 8	9379642	9384572	6348	92.4	1.6E-126
ALK	1	OB06G17800	Chr. 6	5218731	5224167	5437	94.2	0.0
DPEI		UNKNOWN						
GBSSII	1	OB0037G10230	Chr. 7			4776	90.5	2.9E-159
ISA	1	OB08G27630	Chr. 8	16176164	16184129	7966	89.5	1.2E-93
PUL	1	OB04G11820	Chr. 4	1712825	1725559	12735	90.1	3.1E-98
SBE1	1	OB06G35740	Chr. 6	21520777	21526573	5797	94.4	0.0
SBE3	1	OB02G26660	Chr. 2	13948443	13958806	10364	88.7	1.6E-78
SSI	1	OB06G13760	Chr. 6	2200850	2208525	7462	93.7	0.0
SSII 1	1	OB10G18500	Chr. 10	8851531	8863293	8050	93.6	0.0
SSII 2	1	OB02G39280	Chr. 2	23179404	23191306	11218	95.1	0.0
SSIII 1	1	OB04G32800	Chr. 4	18435317	18443066	7750	91.6	0.0
SSIII 2	1	OB08G15430	Chr. 8	3676339	3688697	12359	92.7	0.0
SSIV 1	1	OB01G38130	Chr. 1	22754355	22763315	8961	90.3	0.0
SSIV 2	1	OB05G31380	Chr. 5	17372953	17379376	6424	92.5	0.0
Wx	1	OB06G11980	Chr. 6	1175992	1178897	2915	94.9	1.1E-108
AGPL1	1	KN538814.1_FG015	Chr. 5	24862240	24865615	3376	99.4	0.0
AGPL3	1	KN539195.1_FG019	Chr. 3	37651614	37655808	6268	99.2	0.0
AGPL4	1	KN538938.1_FG006	Chr. 3	24152577	24156527	4281	99.1	0.0
AGPS2a	1	KN542860.1_FG001	Chr. 8	12526401	12529162	5564	99.3	0.0
ALK	1	KN538722.1_FG043	Chr. 6	8195027	8198855	3964	99.6	0.0
DPEI	1	KN538920.1_FG009	Chr. 7	22602297	22605871	3575	98.2	0.0
GBSSII	1	KN539544.1_FG005	Chr. 7	12091388	12095326	3939	99.0	0.0
ISA	1	KN538852.1_FG008	Chr. 8	21633299	21637060	6624	98.4	0.0
PUL	1	KN541174.1_FG002	Chr. 4	1902073	1904954	8157	99.0	0.0
SBE1	1	KN538785.1_FG052	Chr. 6	30095026	30100130	6120	99.5	0.0
SBE3	1	KN539270.1_FG007	Chr. 2	18668529	18679912	10831	99.3	0.0
SSI	1	KN538819.1_FG032	Chr. 6	4298110	4304233	6798	99.4	0.0
SSII 1	1	KN538870.1_FG026	Chr. 10	11500196	11503685	9557	97.8	0.0
SSII 2	1	KN539373.1_FG011	Chr. 2	31221350	31225771	11863	99.2	0.0
SSIII 1	1	KN539225.1_FG011	Chr. 4	28535423	28542610	7198	98.0	0.0
SSIII 2	1	KN539976.1_FG005	Chr. 8	4385551	4396060	10509	99.3	0.0
SSIV 1	1	KN538890.1_FG010	Chr. 1	25879105	25889274	11120	97.3	0.0

SSIV 2	1	KN542368.1_FG001	Chr. 9	3065269	3071676	8426	98.6	0.0
Wx	1	KN539033.1_FG002	Chr. 6	2346028	2348767	3075	99.7	0.0
AGPL1	1	LPERR05G22710	Chr. 5	19941516	19941853	3365	90.2	1.2E-111
AGPL3	3	LPERR03G27770	Chr. 3	24280936	24285189	4254	94.8	1.9E-79
AGPL4	1	LPERR07G07430	Chr. 7	6441935	6453460	11426	93.8	8.7E-76
AGPS2a	2	LPERR08G09730	Chr. 8	9674267	9674615	5439	94.6	1.4E-151
ALK	1	LPERR06G07610	Chr. 6	5565368	5569799	4432	92.1	0.0
DPEI	1	LPERR07G19870	Chr. 7	19361114	19361592	4173	93.5	0.0
GBSSII	1	LPERR05G07400	Chr. 5	6809691	6818650	8960	91.0	9.7E-107
ISA	2	LPERR08G17590	Chr. 8	17935360	17943243	7884	89.6	2.1E-98
PUL	5	LPERR04G01450	Chr. 4	1637441	1651731	14299	92.5	2.4E-151
SBE1	1	LPERR06G23440	Chr. 6	20752621	20758851	6231	94.6	0.0
SBE3	1	LPERR02G15110	Chr. 2	13299407	13312765	13359	91.3	3.8E-144
SSI	1	LPERR06G03680	Chr. 6	2514855	2524553	9699	93.2	8.9E-98
SSII 1	3	LPERR10G07960	Chr. 10	9684239	9694403	10165	95.9	0.0
SSII 2	1	LPERR02G26450	Chr. 2	22543587	22548109	4523	94.0	0.0
	1	LPERR06G07610	Chr. 6	5565368	5569799	4432	89.1	5.4E-166
SSIII 1	1	LPERR04G21480	Chr. 4	20045750	20053049	7300	90.6	0.0
SSIII 2	1	LPERR08G05100	Chr. 8	4147149	4157275	10127	91.3	0.0
SSIV 1	5	LPERR01G25000	Chr. 1	21819891	21835575	15685	89.9	0.0
SSIV 2	2	LPERR05G19650	Chr. 5	17931037	17938653	7610	91.7	0.0
Wx	2	LPERR06G01820	Chr. 6	1223031	1227394	4364	92.5	8.9e-94
	1	LPERR06G01810	Chr. 7	1218230	1221309	3080	92.0	1.3E-89

**Figure S1:** Positive selection and Mixed Effects Model of Evolution (MEME) in AGPS-AGPL genes. A total of 34 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT).

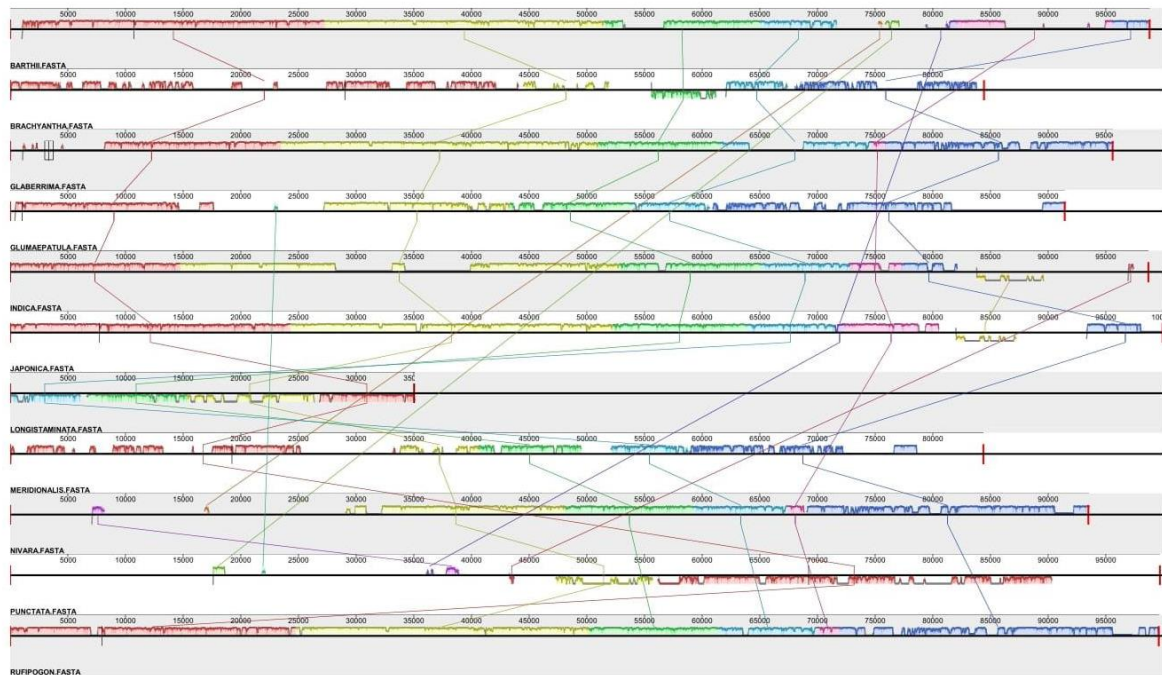


**Figure S2:** Alignment of the promoter region (1,500 bps upstream) of AGPS2a genes, showing possible translocation events between ONIVAGPS2a and OMERAGPS2a.

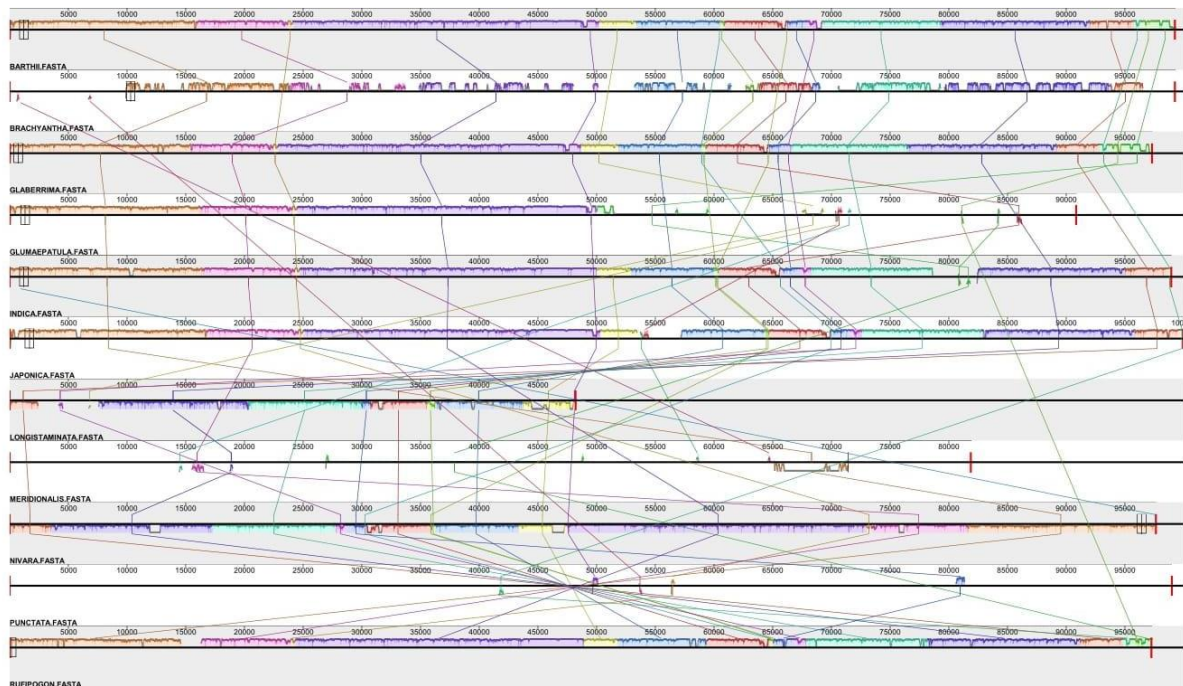




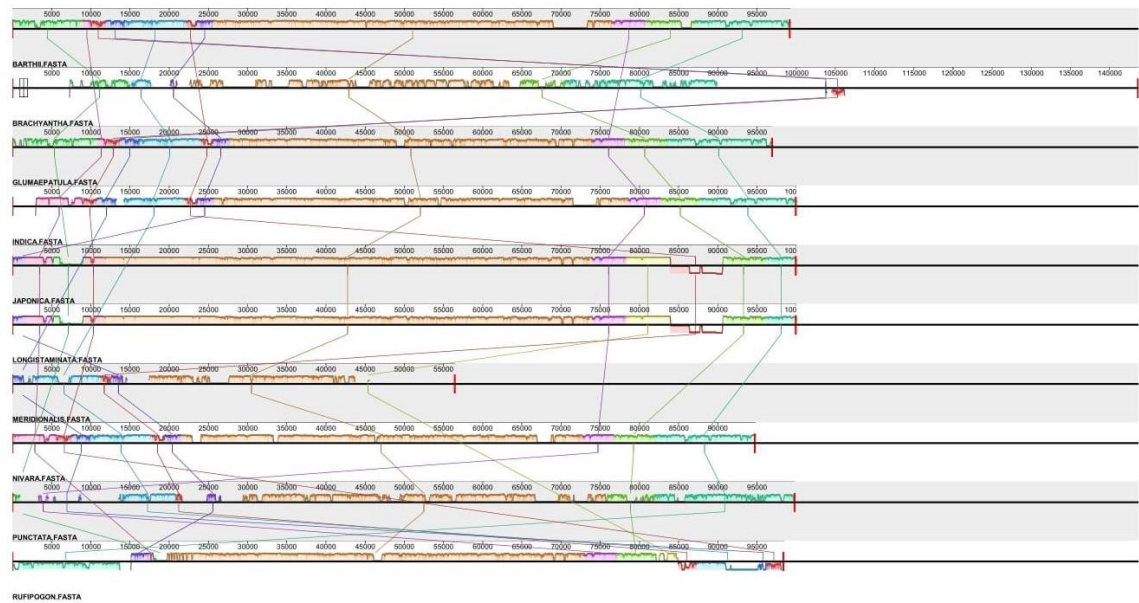
**Figure S3:** Alignment of the promoter region (1,500 bps upstream) of SSII1 genes, showing possible translocation events between OMERSSII1 and other *Oryza* species.



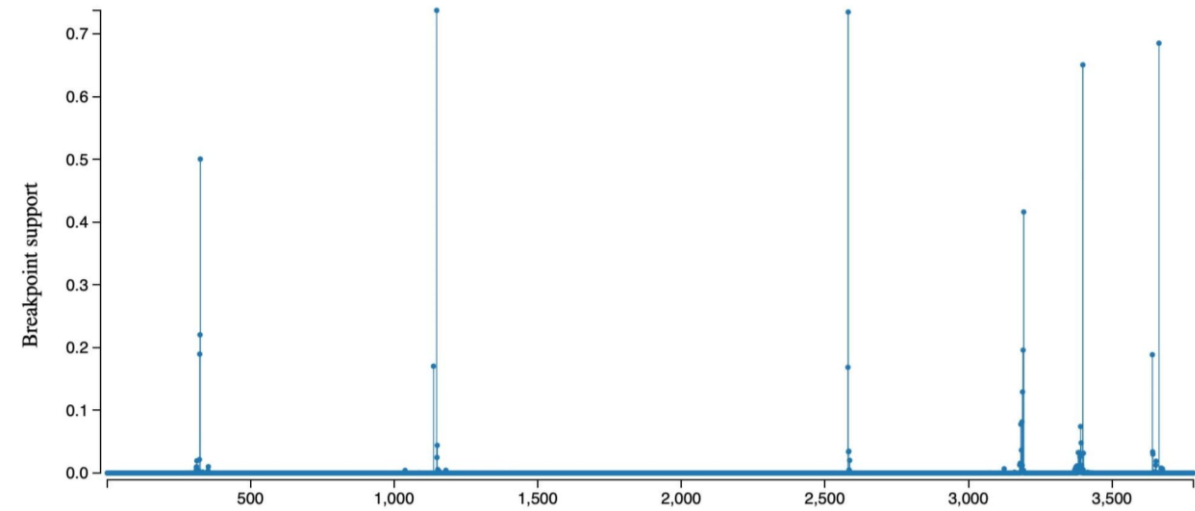
**Figure S4:** Alignment of the promoter region (1,500 bps upstream) of SSII2 genes, showing possible translocation events between OMERSSII2 and ONIVSSII2 and other *Oryza* species.



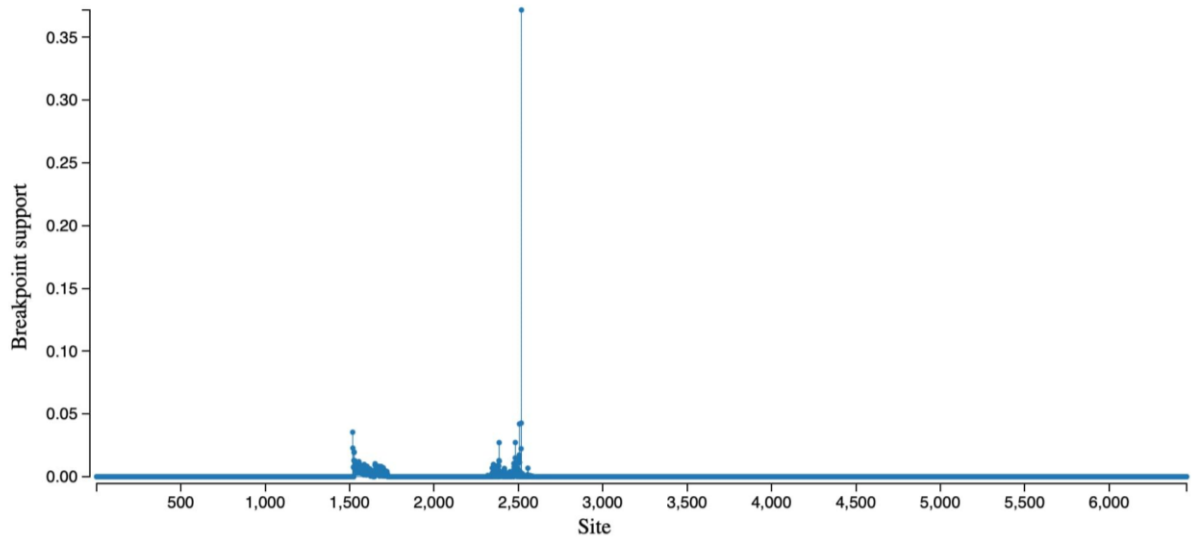
**Figure S5:** Alignment of the promoter region (1,500 bps upstream) of SSIV2 genes, showing possible translocation events in OLONSSIV2 and other *Oryza* species.



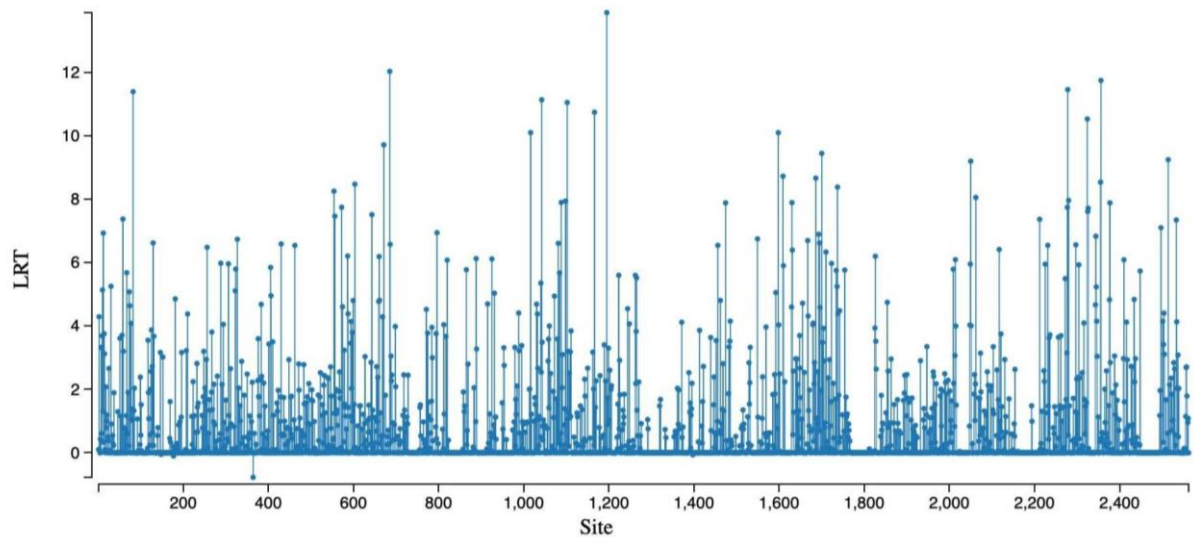
**Figure S6:** Genetic Algorithm for Recombination Detection (GARD) for Waxy genes. Comparing the AICc score of the best fitting GARD model, that allows for different topologies between segments (21,426.8), and that of the model that assumes the same tree for all the partitions inferred by GARD the same tree, but allows different branch lengths between partitions (21,430.9) suggests that because the multiple tree model cannot be preferred over the single tree model by an evidence ratio of 100 or greater, some or all of the breakpoints may reflect rate variation instead of topological incongruence.



**Figure S7:** Genetic Algorithm for Recombination Detection (GARD) for ALK genes. Comparing the AICc score of the best fitting GARD model, that allows for different topologies between segments (32,200.2), and that of the model that assumes the same tree for all the partitions inferred by GARD the same tree, but allows different branch lengths between partitions (32,211.2) suggests that because the multiple tree model can be preferred over the single tree model by an evidence ratio of 100 or greater, at least of one of the breakpoints reflects a true topological incongruence.



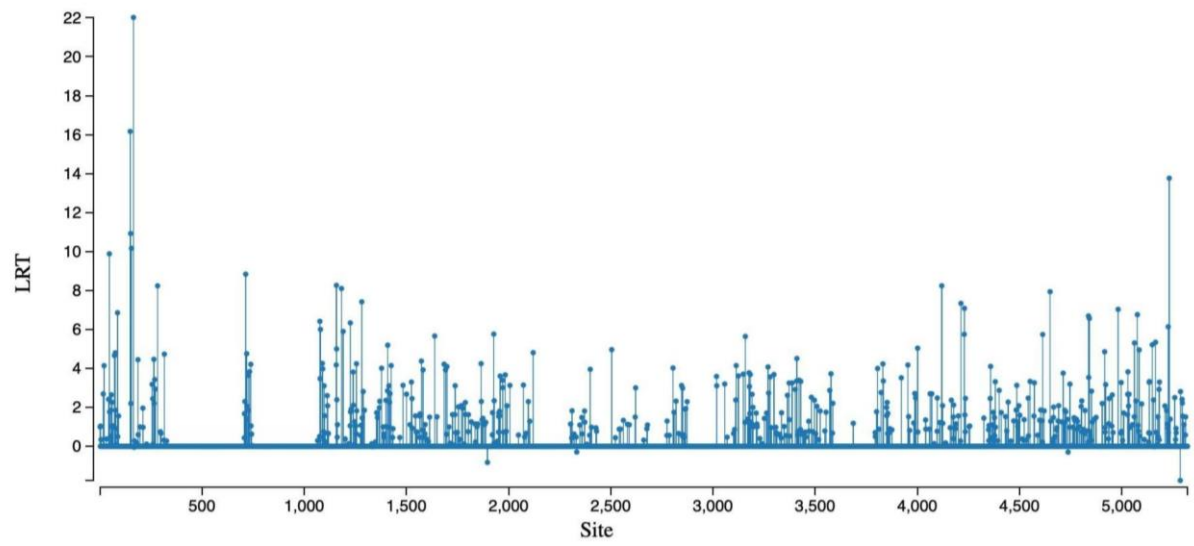
**Figure S8:** Positive selection and Mixed Effects Model of Evolution (MEME) in SS genes. A total of 117 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT).







**Figure S11:** Positive selection and Mixed Effects Model of Evolution (MEME) in SBE genes. A total of 46 sites were found under positive/diversifying selection based in the Likelihood ratio test statistic for episodic diversification (LRT).



## 4 CAPÍTULOS

4.3 Artigo 3 – *In silico* analysis of *Oryza* genomes: Occurrence of cis-regulatory elements of starch-related genes provide new breeding possibilities for increased grain quality in rice under stress

*Artigo submetido na revista Plant Stress (Elsevier)*



## ***In silico* analysis of *Oryza* genomes: Occurrence of cis-regulatory elements of starch-related genes provide new breeding possibilities for increased grain quality in rice under stress**

**Karine Elise Janner de Freitas<sup>1</sup>; Railson Schreinert dos Santos<sup>2</sup>; Vivian Ebeling Viana<sup>1</sup>; Filipe de Carvalho Victoria<sup>3</sup>; Camila Pegoraro<sup>1</sup>; Carlos Busanello<sup>1</sup>; Antonio Costa de Oliveira<sup>1</sup>**

<sup>1</sup> Universidade Federal de Pelotas, UFPel/FAEM, Centro de Genômica e Fitomelhoramento, C.P. 354, CEP 96010-900, Pelotas-RS, Brazil.

<sup>2</sup> Instituto Federal de Educação, Ciência e Tecnologia Catarinense (IFC), Rodovia SC 283, s/n Fragosos, SC, 89703-720, Concórdia, SC, Brazil.

<sup>3</sup> Universidade Federal do Pampa (UNIPAMPA), Núcleo de Estudos da Vegetação Antártica—NEVA, São Gabriel-RS, Brazil

**Corresponding author:** Antonio Costa de Oliveira, Universidade Federal de Pelotas, UFPel/FAEM, Centro de Genômica e Fitomelhoramento, C.P. 354, CEP 96010-900, Pelotas-RS, Brasil. Fone: 55 (53) 3275 7263. E-mail: acostol@terra.com.br.

### **Abstract**

Rice (*Oryza sativa* L.) is a crop of great importance and widely consumed worldwide. Although grain quality is something of extreme importance, the concept of quality varies from population to population, according to local culinary and culture. The comparison of quality components among different cultivars and environmental cultivation conditions reveals significant variation. Considering that genetic differences and environmental influences on gene regulation can alter grain quality, it is important to analyze DNA variations among the different species of the *Oryza* genus, paying special attention to their regulatory regions. In this sense, the study of loci that encode starch synthetases and seed-storage proteins, as well as their *cis*- and *trans*-acting regulators, is something of substantial relevance. In this study DNA motifs that are probably related with modification of the expression of genes involved in starch-related pathways are highlighted, helping the understanding of Starch Synthesis-Related Genes (SSRGs) regulation in plants under stress. The role of *cis*-acting regulators (CREs) of SSRGs as well as the signaling cascade involving these genes are discussed indicating possible uses for quality-related breeding.

**Key words:** cold, dehydration, salt, starch synthetases, seed-storage proteins.

## List of abbreviations

SSRGs: Starch Synthesis-Related Genes  
TFs: Transcription Factors  
CREs: *Cis* Regulatory Elements  
RAP-DB: The Rice Annotation Project Database  
MEME: Multiple Em for Motif Elicitation  
TFBSs: Transcription Factor Binding Sites  
SNP: Single Nucleotide Polymorphism  
AGPL: ADP-glucose Pyrophosphorylase Large Subunit  
SS: Starch Synthase  
GBSSII: Granule-Bound Starch Synthase II  
WAXY: Granule-Bound Starch Synthase 1  
ALK: Starch Synthase III  
SBE: Starch Branching Enzyme  
PUL: Pullulanase  
ISA: Isoamylase  
DPE1: Disproportionating Enzyme 1  
DBE: Debranching Enzyme  
ABA: Absciscic Acid  
SA: Salicylic Acid  
SEBF: Silencing Element Binding Factor  
GT: Gelatinization Temperature  
AC: Amylose Content  
GA: Giberellin  
IAA: Indole-3-Acetic Acid  
JA: Jasmonate  
ET: Ethylene

## 1. Introduction

Asian rice (*Oryza sativa* L.) is a worldwide staple food, although still predominant in Asian countries, where it is part of local culture. Being widely consumed and having different forms of preparation, the idea of rice grain quality can be different, according to each population. The concept of quality is subjective especially when we talk about grains, in which it covers physical, biochemical and physiological properties. It is also important to notice that starch and proteins are two of the main components of rice endosperm and therefore constitute keys for quality achievement (Ahmed *et al.* 2020).

Comparison of quality components reveal significant variation among cultivars grown in the same environment, which indicates that some of the decisive factors controlling starch/protein content, and thus grain quality, lie in rice genome itself. Besides that, grain quality is also affected by environmental conditions like water availability,

temperature, fertility, and salinity (Cameron *et al.* 2008; Sharifi *et al.* 2009; Bao *et al.* 2000). In this sense, the study of loci that encode starch synthetases and seed-storage proteins, as well as their *cis*- and *trans*-acting regulators is something of substantial relevance. These elements highly affect the activity of these genes under different conditions, (Adu-Kwarteng *et al.* 2003; Cameron and Wang 2005; Kang *et al.* 2006; Vidal *et al.* 2007). Signal transduction pathways controlling grain quality remain largely unclear. Understanding the transcriptional regulation of starch synthesis genes in *Oryza* species can allow one to effectively change the expression pattern of these genes in specific ways, which can further provide new possibilities for plant genetic engineering to control Starch Synthesis-Related Genes (SSRGs) and grain quality in plants grown under environmental stresses (de Freitas *et al.*, 2021; Zhu *et al.*, 2011).

All seed-storage proteins and starch synthetase genes are strongly expressed during grain development, indicating that they might share similar regulation mechanisms on the transcriptional level (Chen *et al.*, 2012). Transcription regulation involves association between transcription factors (TFs) and *cis* regulatory elements (CREs) of specific genes. CREs are short regulatory motifs (5±20 bp) present in the promoter regions of target genes. These promoters play important roles in controlling gene expression and multiple CREs, such as TATA box, GC box, and CAAT box contain coupling sites for TFs, as well as other important elements required for proper spatiotemporal expression of genes (Hasan *et al.*, 2017; Lenka & Bansal, 2019).

CREs are essential regulatory units for genetic stress responses and its study in SSRGs promoters can lead to better comprehension of the transcriptional expression of these genes. Understanding CREs role is one of the first steps that we must focus on in order to find the main points of regulation and to design strategies to change gene expression. Here we aim to find CREs responsible for altering the expression of SSRGs and to better understand their possible impact in rice plants that are under stress.

## **2. Materials and methods**

### **2.1 Selecting promoters of SSRGs**

The SSRGs used in this study were selected according to Zeng *et al.*, (2018). Initially, the SSRGs of *O. sativa* ssp. *japonica* were obtained through The Rice Annotation Project Database (RAP-DB) (<https://rapdb.dna.affrc.go.jp/index.html>).

The similarity of *O. sativa* ssp. *japonica* SSRGs to other *Oryza* sequences was evaluated through BLAST (Altschul *et al.*, 1990), using sequences already available in ENSEMBL PLANTS database (<http://plants.ensembl.org/index.html>). Only SSRGs with high coverage (100%) and low e-value (zero) were selected for further analysis (Table S1).

The promoter sequences (1.5 kbp upstream of translation start site) of each *Oryza* SSRG were extracted from ENSEMBL PLANTS database and analyzed for CRE identification using Plant *cis*-acting regulatory DNA elements (PLACE) database (<http://www.dna.affrc.go.jp/htdocs/PLACE/>). Also, MEME suite (Bailey and Elkan, 1994) was used in motif identification and PlantPAN 3.0 (Chow *et al.*, 2019) in TF binding sites (TFBSs) detection.

### **2.2 Ontology**



The ontology enrichment for SSRGs selected for this study was obtained from ShinyGo (Ge et al., 2020) and adjusted for the 10 most highly represented GO terms that included biological process gProfiler (Rauduvere et al., 2019) and visualized with RStudio script from REVIGO (Supek et al., 2011). A total of 100 GO terms were significantly enriched among with adjusted p-value of  $< 0.05$ . Significantly over-represented biological processes based on GO terms were visualized in REVIGO.

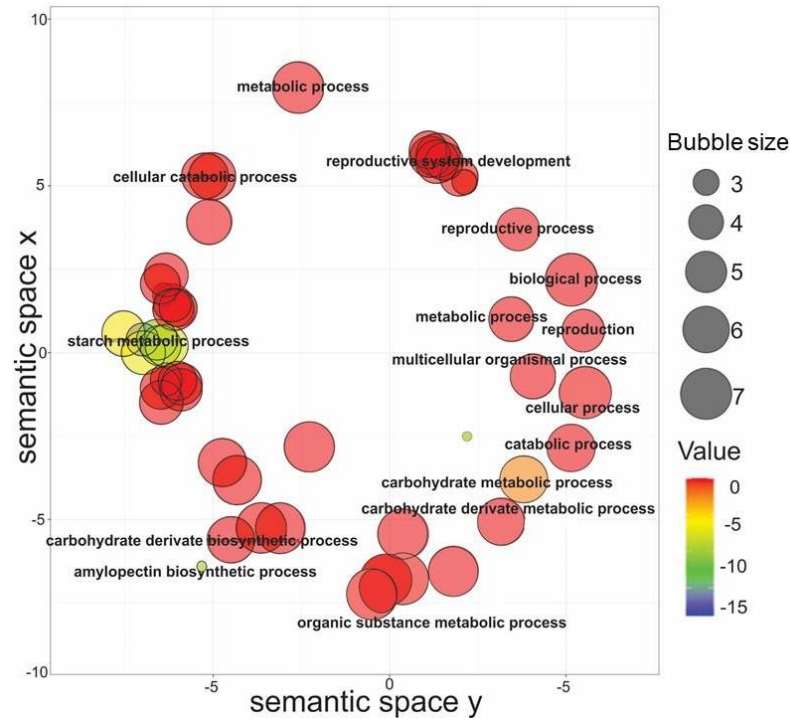
### 2.3 Transcriptional analysis

Genevestigator (<https://genevestigator.com/gv/>) (Zimmermann *et al.*, 2004) was used for analyzing SSRGs transcriptional expression in *Oryza sativa*. The “perturbations” tool was used to determine in which experimental conditions the selected genes were differentially expressed using plants under cold, dehydration and salt stresses, according to the database obtained from Jain *et al.*, (2007) (Experiment ID: OS-00008 GSE6801; MicroArray data) *Oryza sativa* ssp. *indica* var IR64. Detailed treatment conditions: cold (IR64 seedlings, kept at  $4 \pm 1^\circ\text{C}$  for 3h/Untreated IR64 seedlings, kept in beaker containing water for 3h at  $28 \pm 1^\circ\text{C}$ ), dehydration (IR64 seedlings, dried between folds of tissue paper at  $28 \pm 1^\circ\text{C}$  for 3 hours/ Untreated IR64 seedlings, kept in beaker containing water for 3h at  $28 \pm 1^\circ\text{C}$ ), salt (IR64 seedlings, transferred to a beaker containing 200 mM NaCl solution for 3h at  $28 \pm 1^\circ\text{C}$ /Untreated IR64 seedlings, kept in beaker containing water for 3h at  $28 \pm 1^\circ\text{C}$ ). Transcriptional expression was analyzed during the milk stage of endosperm formation. This data was used to establish a relationship between previous described CREs and modification of the expression in rice under different conditions.

## 3. Results and discussion

### 3.1 Ontology Enrichment and Network Pathway Analysis for SSRGs in *Oryza*

Functional analysis of SSRGs copies was conducted to find in which pathways or processes these genes participate. It is possible to observe six genes related with Starch metabolic processes and one related with amylopectin biosynthesis (Figure 1). These roles in stress responses and CREs activation are discussed below.



**Figure 1.** The scatterplot shows the cluster representatives (i.e. terms remaining after the redundancy reduction) in a two dimensional space derived by applying multidimensional scaling to a matrix of the GO terms in Biological Process for SSRGs genes copies selected for this study. Bubble color indicates the p-value (in  $\log^{-10}$ ), according to “value” legend seen at the bottom right, while its size indicates the frequency of the GO term according to “bubble size” legend seen at the bottom right.

### 3.2 Presence of CREs in SSRGs promoters

The search performed using PLACE reveals a total of 5,907 CREs in SSRGs (*AGPL1*, *AGPL3*, *AGPL4*, *AGPS2a*, *ALK*, *SBE1*, *SBE3*, *PUL*, *ISA*, *DPE1*, *SSI*, *SSII1*, *SSII2*, *GBSSII*, *SSIII1*, *SSIII2*, *SSIV1*, *SSIV2*, *Waxy*) promoters, considering all the occurrences in the 11 analyzed *Oryza* species. CREs are grouped in different functional categories as shown in Supplementary Table S2. “Cellular development” presents the highest number of different types of CREs, followed by “abiotic and biotic stress”, “hormonal regulation” and “amylose related”. The categories of CREs are organized from those with higher frequency to those with lower: “dehydration responsive”, “common cis regulatory element”, “light responsive” and “root specific responses”. The total types of CREs in each gene are shown in Figure 2.





### 3.3.1 Dehydration responses

Among the most abundant CREs, those involved in dehydration responses, as MYCCONSENSUSAT, MYBCORE and ACGTATERD1, are the most frequent in the SSRGs. From these, MYCCONSENSUSAT is present in every species except for *OMERSSII2* and *OsJAPALK*, evidencing its importance in abiotic stress response, since it is also involved with other stresses and hormonal signaling (Abe *et al.*, 2003). Other frequent CRE, ACGTATERD1, is related to early response to dehydration and etiolation. Whereas the promoters contained multiple dehydration regulatory elements in some groups of SSRGs in *Oryza* species, between them CRE related to MYC site (MYCATRD22 and MYCATERD1) composed by bHLH proteins that are involved in the response to injury, drought and salinity stress, regulation of seed germination, trichome and fetal development (Feller *et al.*, 2011) and CRE related to DREBs binding site (DRE1COREZMRAB17, DREDR1ATRD29AB, DRECRTCOREAT and CBFHV). A CRE MYB related binding sites (MYB2CONSENSUSAT, MYBCOREATCYCB1, MYBPZM, MYBCORE, MYBPLANT, MYB26PS, MYBATRD22, MYB1LEPR, BOXLCOREDCPAL, MYBST1, MYB1AT, MYB2AT) presented the most diversity.

When plant cells sense the loss of water, many different genes downstream a signaling cascade are transcriptionally activated, which will display a role in drought tolerance (Kuromori *et al.*, 2014). The plant hormone abscisic acid (ABA) is produced under water-deficit and regulates the expression of many drought-responsive genes. At this step, regulatory elements such as MYB, MYC and ERD, play important role in adapting endosperm to the water deficit and are activated upon ABA signaling in an ABA-dependent manner. ERD respond immediately to water deficit, but the activation of MYB and MYC occurs through a DNA fragment found in promoters that contains two putative recognition sites for a basic helix-loop-helix protein. Here the MYC and MYB proteins cooperate with each other for the transcriptional activation of ABA-responsive gene expression (Abe *et al.*, 1997), thus forming a protection for plants and consequent normal SSRG expression for endosperm composition.

Water deficit in rice can reduce starch accumulation up to 40%, leading to changes in starch composition, structure, and functionality (Thitisaksakul *et al.*, 2012). The reduced starch levels lead to a significantly lower amylose content, specifically increasing flour swelling power, peak viscosity, cohesiveness, gel hardness, and granular breakdown. Possibly, this effect is due to changes in *Waxy* and *GBSSII* gene regulation, which presents higher abundance of CREs in *O. sativa* ssp. *indica* but higher diversity in wild species. Also, it was possible to notice that most of the major shared CREs (MYBCORE, MYCCONSENSUSAT, ACGTATERD1) related to dehydration was present in all genes indicating that the SSRGs share a remarkably similar regulatory network.

Knowing that starch pathway works in a fine regulation network (Qu *et al.*, 2018), any alteration in correct temporal expression of SSRGs due to stress, can affect the molecular network controlling starch biosynthesis. Here, we find a great frequency in sequences associated with dehydration, mainly in *O. barthi* and *O. brachyantha* promoters of *AGPase*, *SS*, *DBE*, and *SBE* (Supplementary table 2, 3, 4, 5 and 6). Less diversity can be found in *O. sativa* ssp. *japonica*, once again highlighting the importance of the study of wild species aiming drought tolerance.

### 3.3.2 Light responses

The second group of most abundant stress related CREs are those associated with light responses (Figure 2), these are important since, as suggested by some authors, disease resistance can be related to light responses (Zhang *et al.*, 2016; Brodersen *et al.*, 2002; Asai *et al.*, 2000). GT1CONSENSUS, IBOXCORE, and GATABOX, which are motifs that interact and stabilize the TFIIA-TBP-TATA complex, are frequent in all the *Oryza* species analyzed here and are shared by all SSRGs. INRNTPSADB is a TATA box independent motif, able to quickly perform gene activation. SORLIPs are motifs found in every SSRG, while some other light motifs are less frequent, like GT1CORE, BOXIIPCCHS, GT1MOTIFPSRBCS that bind to box II; and IBOX, IBOXCORENT, LREBOXIPCCHS1 that can bind to box I (Supplementary table 2). HEXAT is a bZIP and G-box binding site that is frequent only in promoters of *SSI* genes (Berendzen *et al.*, 2012). In addition to motifs that activate genes in response to light, there are also CREs (GT2OSPHYA and BOXCPSAS1) that have been identified to inactivate genes in this same situation. Light is a predominant factor which controls the circadian rhythm of various life processes such as growth, development, nitrate uptake, and stress responses. Both low and intense light can influence rice endosperm formation, affecting starch deposition (Beckles & Thitisaksakul, 2012). These light-associated motifs are essential for light-controlled transcriptional activity. The high frequency of these elements, like what was observed for dehydration-responsive-elements in promoters of rice SSRGs, suggests that the expression of these genes may have a fine regulation especially dependent on light and dehydration. Some other gene families in rice have been shown to be regulated by light due to the presence of GT1 box and TGACG motifs in their promoters, setting advantage against light stresses (Fukunaga *et al.*, 2010).

### 3.3.3 Anaerobic responses

Few anaerobic responsive CREs were identified in SSRGs promoters of *Oryza* species (Supplementary Table 2). Among these is CURECORECR, that coordinates gene expression in response to oxygen deficiency and in response to copper. ANAEROCONSENSUS 1, 2 and 3, common motifs in the fermentative pathway, and the GCBP2ZMGAPC4, which also responds directly to anaerobic conditions, could also be found (Niemeyer *et al.*, 2010). It is interesting to notice that, in *O. sativa* ssp. *japonica*, the presence of GCBP2ZMGAPC4 motif is found only in *AGPL1*, which, being one of the first genes in starch biosynthetic pathway, acts in the cytosol.

It has been reported that under elevated CO<sub>2</sub> concentrations, rice normally has a clear trend of quality deterioration and therefore lower commercial value (Wang *et al.*, 2011). The low number of CREs related to anaerobiosis may be a limiting component, which further suggests the importance of adapting cultivars for production in atmospheric environments with high concentrations of CO<sub>2</sub>, which is a future scenario, since the content of CO<sub>2</sub> in the environment tends to increase. On the other hand, the stage that is more tolerant to anaerobiosis in rice is the germination, being rice seeds more tolerant to anaerobiosis than oilseeds, due to their ability to maintain a higher energy metabolism under oxygen deficiency (Illangakoon *et al.*, 2018; Raymond *et al.*, 1985).

### 3.3.4 Cold and heat stress

Other important CREs from abiotic stress that can affect rice germination in extreme temperatures are available in Supplementary Table 2. Generally, there is an increase in the ratio of amylose/amylopectin in grains of rice grown

under controlled low temperature (25°C), and, besides that, it generally causes a decrease in gelatinization and pasting temperatures in most of the cooking analyzes (Beckles and Thitisaksakul, 2014; Asaoka *et al.*, 1984). Normally, *SSIII*, *SSII2*, *SSI* and *SBEs*, form an important complex for the biosynthesis of short and intermediate amylopectin chains. However, these are the most affected genes by cold, therefore directly affecting the gelatinization temperature. In general, *SSI*, *SSII* and *SBE* present low frequency of cold-responsive CREs, like LTRE1HVBLT49, LTRECOREATCOR15 and LTREATLTI78 that are frequent in the other SSRGs. Wild species present the CRTDREHVCBF2 motif that can upregulate genes at temperatures below 25°C and which binding activity gradually increases as the temperature decreases up to 0°C (Xue, 2003). This CRE has an advantage when compared to LTRs, since these act in cold adaptation only when temperatures are below 12°C, when damage can already be observed in the endosperm (Xue, 2003). On the other hand, regarding starch formation, it is known that *Waxy* expression increases in response to low temperatures (18°C), resulting in greater amylose accumulation and demonstrating that temperature does not alter negatively starch formation to a certain extent (Ahmed *et al.*, 2008; Hirano & Sano, 1998).

The CCAATBOX1, a heat shock element, could be found in almost every SSRG, being the unique heat responsive CRE in the analyzed promoters. When high temperature peaks occur during growth and endosperm filling, alterations in starch granule size, shape, and structure, can also occur, causing problems as pitting and fissures. These problems can vary according to the period of exposure to the stress and to genotype-specific responses (Zhen-zhen *et al.*, 2015).

### 3.3.5 Biotic stress responses (herbivores and pathogens)

Plants respond to biotic stress through a well-regulated defense system and resistance to biotic stress can be induced through specific chemical compounds and plant hormones as salicylic acid (SA), jasmonic acid (JA), and ethylene (ET) which play central roles in cell signaling (Llorens *et al.*, 2017). The different hormones are possibly activated by a signaling cascade through some CREs as BIHD1OS, which is an ethylene-responsive factor (ERF) that interact with BELL homeodomain and activate all SSRGs in infected plants (Kaur *et al.*, 2017; Luo *et al.*, 2005). ABA-related CREs that can participate in the responses to infections were also identified (Supplementary table 2). Among them, MYB1LEPR, GCCCORE and CACGTGMOTIF that regulate defense-related gene expression via GCC box, non-GCC box and G-box. WBOXNTERF3 is the unique element that activates genes in wounding response, usually due to insect attack.

Some CREs were present in low quantity in SSRGs: GT1GMSCAM4, which is associated with pathogen and salt; WBOXNTCHN48, an elicitor associated with virus infection; and ELRECOREPCR1 other elicitor associated with fungal infection. In contrast, the presence of the motif SEBFCONSSTPR10A in *AGPL4*, *SSI*, *SI11*, *SI12*, *SI13*, *SIV2* in *O. sativa* ssp. *japonica* and some wild species can repress these genes through a Silencing Element Binding Factor (SEBF) between residues -52 and -27 in response to pathogen infection or elicitor treatment (Barsain *et al.*, 2019; Boyle *et al.*, 2001) directly influencing starch synthesis. *O. sativa* ssp. *indica* presents more frequency of defense related CREs as pathogen and wounding than other species.

### 3.3.6 Mineral responses



CREs related to mineral responses were less frequent in SSRGs (Supplementary table 2). However, it is interesting to note that mineral fertilization can cause pronounced effects in starch properties, altering pasting and thermal properties. Also, it has already been described that mineral content in rice grains is related to cooking quality traits (Singh *et al.*, 2011; Li *et al.*, 2013; Beckles and Thitisaksakul, 2014).

On the other hand, some wild species have a higher frequency of motifs related to nitrogen, salt and phosphate responses. Variation in soil nitrogen (N) usually has direct consequences in grain starch/protein ratio, suggesting ramifications that can affect starch functionality, possibly altering properties related to gelatinization and pasting (Beckles *et al.*, 2012; Beckles and Thitisaksakul, 2014). NODCON2GM acts regulating genes involved in nitrogen fixation in soybean and it has already been reported that rice grown in soils with low N-concentration presents higher amylose content in its grains (Singh *et al.*, 2011). In agreement, here we identified this CRE in all the genes of almost every wild species, while for *O. sativa ssp. japonica*, it was identified in only *AGPL1* (involved in ADP glucose biosynthesis), *Waxy* and *GBSSII* (responsible for amylose production).

SSRGs present two quite common CREs that are associated to salt stress, including GT1GMSCAM4, which also responds to pathogen attack. However, the most frequent motif in *Oryza* SSRGs is ARR1AT. In agreement, it was reported that grains of both salt tolerant and susceptible cultivars grown on saline soils have higher storage protein contents, but less translucent grain, and lower starch and amylose content than those grown on normal soil (Rao *et al.*, 2013; Siscar-Lee *et al.*, 1990).

High phosphorus fertilization is also known to alter starch functionality. However, available phosphate (Pi) is a major limiting factor for plant growth, development, and productivity. PIBS, a motif that was identified in all the analyzed species is an important CRE reported to be present in genes that collaborate to improve soybean tolerance to low phosphorus conditions (Ni *et al.*, 2012; Li *et al.*, 2015).

In sulfur fertilization (S) both pasting and thermal properties of starch can be influenced (Li *et al.*, 2013). Some SSRGs present the SURECOREATSULTR11, an element associated with sulfur deficiency. It is suggested that SURE core sequences may commonly regulate the expression of a gene set required for adaptation to soils with a high sulfur content (Maruyama-Nakashita *et al.*, 2005). S is a vital element for every organism due to its important role in methionine and cysteine biosynthesis. S deficiency in rice is increasing each year since its availability depends on soil temperature and moisture, which both influence organic matter decomposition rates (Lucheta *et al.*, 2012), therefore, presence of a CRE that promote rapid SSRGs response when the plant needs to face sulfur deficiency is something advantageous.

The regulation of gene expression by intracellular calcium is crucial for plant defense against biotic and abiotic stresses. The regulation of some SSRGs is also calcium dependent, but the number of genes known to respond to these specific transient signals is limited (Kaplan *et al.*, 2006). The calmodulin-binding/CGCG box element is involved in multiple signaling pathways that are rapidly and differentially induced by environmental signals such as extreme temperatures, UVB, wounding, hormones (ET and ABA), and signaling molecules (methyl jasmonate, H<sub>2</sub>O<sub>2</sub>, and salicylic acid). These CREs, which are found in high amount mainly in *SSIII*, *SSIV* and *ALK* genes of *O. sativa japonica*, can cause important responses when plants are facing different stresses.

IRO2OS is the unique motif associated with iron deficiency identified in some wild species, but motifs related to iron excess, could not be identified. Even though Fe is an essential micronutrient for human nutrition, the content of metal ions in rice grain is usually poor and an increase in its amount would be beneficial. (Heinemann *et al.* 2005).

### 3.4 AMYLOSE-RELATED CREs

Starch-related cres are less frequent than those associated with stresses but as expected, can also be found in every srrg across *Oryza* species (supplementary table 2). These elements are important for high transcriptional expression of genes related to starch and of others that can mediate cellular development and hormonal regulation (Agarwal *et al.*, 2011; Hwang *et al.*, 2004; Mitsui and Itoh *et al.*, 1997).

AMYBOX1 and AMYBOX2 are found in rice  $\alpha$ -amylase multigene family (Mitsui and Itoh *et al.*, 1997). Although several enzymes are involved in the germination process, these  $\alpha$ -amylases are primarily responsible for the endoglycolytic cleavage of amylose and amylopectin (Damaris *et al.*, 2019). Similarly, CGACGOSAMY3 motif is also responsible for inducing expression upon sugar starvation and was relatively frequent in SSRGs promoters. The presence of these CREs in SSRGs is important to increase gene expression when the plant has low sugar concentrations levels in the cells, promoting starch breakdown and providing glucose. TATCCAOSAMY mediates the response of genes that catalyze amylose biosynthesis in rice endosperm in response to sugar and hormones. This is the same case of SP8BFIBSP8BIB, which can also promote a decrease in gene expression when leaves are treated with sucrose (Ishiguro and Nakamura, 2004; Ishiguro and Nakamura, 1994).

WBBOXPCWRKY1 is a binding motif observed in starch related genes regulated by gibberellic acid (GA). This element was identified in promoters that present AMYBOX1 and AMYBOX2, suggesting that these motifs are also regulated by GA (Himmelbach *et al.*, 2010). It is already expected since cereal grain  $\alpha$ -amylase gene expression is stimulated by endogenous GA in germinating seeds (Damaris *et al.*, 2019).

Interestingly, the BP5OSWX motif, which acts as a transcriptional activator of *Waxy*, is also found in *AGPL3*, *SSIII* and *SSIII* promoters, suggesting that this CRE is necessary not only for *Waxy* activation but also for the coordinated action of other genes involved in the amylose/amylopectin pathway. Similarly, TGACGTVMAMY, which interacts with bZIP TFs, is a common motif in promoters of rice storage proteins (glutelin, globulin, prolamin and albumin), and is required for high expression of some SSRGs in cotyledons of germinating seeds (Yamauchi, 2001; Nakase *et al.*, 1997).

### 3.5 HORMONE RESPONSIVE CREs

Stress responses in plants are also associated with hormone signaling and cell division/developmental processes. Stresses can activate signaling molecules (such as ROS and  $Ca^{+2}$ ), affect hormones and mitogen-activated protein kinase (MAPK) cascades which will depend on TFs that will ultimately modify expression when plants are facing a stress (Sewelam *et al.*, 2012). In rice endosperm, GA-responsive CREs were the most frequent in *Oryza* species (Supplementary table S2). Gene expression in the cereal endosperm can be stimulated by endogenous GA in germinating seeds. However, the frequent presence of WRKY71OS CREs in all SSRGs can repress gibberellin signaling. Showing that OsWRKY71, besides being involved in regulation of  $\alpha$ -amylase genes in rice (Zhang *et al.*, 2004), can also cause GA repression in all SSRGs, possibly through interaction with GAMyb motif (MYBGAHV CREs) as detailed by Gubler & Jacobsen (1992). This repression may be associated with rice seed dormancy, which is interesting since WRKY TFs are involved in a wide range of processes, including GA repression also in  $\alpha$ -amylase-related genes.

GARE is another important element for high level of GA induction, still it is less frequently found in SSRGs. As previously described by Heidari *et al.* (2015), in some promoters a single GARE motif can stimulate a high level of transcription in response to hormones due to its cooperation with other cis-acting elements. Another CRE found in SSRGs that deserves special attention is VIVIPAROUS 1 (VP1), but this plays an opposite role to that of GA in the regulation of seed dormancy and germination in rice, suppressing gibberellin-induced expression (Chen *et al.*, 2020). Those associated with auxin and salicylic acid (SA) are part the second most frequent group of CREs in SSRGs promoters. Elements as CATATGGMSAUR, ARFAT, NTBBF1ARROLB are involved with auxin responsiveness, while WBOXATNPR1 responds to SA. On the other hand, the ASF1MOTIFCAMV is involved in transcriptional activation of genes by both, auxin and SA. SA also has an important role inducing WRKY DNA binding proteins in response to pathogen infection (Yu *et al.*, 2001). It is known that indole-3-acetic acid (IAA) can increase spikelet growth and promote the development of distal branches, however, it can also suppress proximal branches (Patel and Mohapatra 1992).

CPBCSPOR is the unique motif found in some SSRGs that exhibits cytokinin-dependent protein binding. Despite that, this element can interact with other cytokinin-related elements that can lead to an up-regulation of several genes that contribute with plastid development and lipid production in thylakoid (Fusada *et al.*, 2005). During seed development, levels of cytokinin are transiently elevated and constitute an important factor for determining grain size (Jameson & Song, 2016).

Jasmonates (JA) are important regulators of plant defense responses that activate the expression of many wound-induced genes (Wasternack and Song, 2017). In some SSRGs, T/GBOXATPIN2 is the unique JA-response CRE that, through to the interaction with MYC TF, can have a role in the expression of genes in wound responses. JA can interact with ET (*ethylene*) through the *ERF1* in response to pathogen attack and plays a role in repression of genes differentially regulated by JA (Li *et al.*, 2019; Lorenzo *et al.*, 2004). This interaction with JA can occur through ERF TF family, also via GCC box of GCCCORE that is most frequent in *AGPL4*, *SBE3*, *SSIII*, and *SSIV1*. These interactions increase our knowledge related to the signaling cascade during plant defense responses. On the other hand, ET evolution rate is negatively correlated with grain-filling rate (Liu *et al.*, 2008), resulting in an inverse correlation with chalky kernel percentage.

Changes in ET, GA and ABA levels in rice is often related to water deficit stress which affects grain-filling rate and eventually causes yield and quality problems (Chen *et al.*, 2012). ABA is important mainly because, like other hormones mentioned above, regulates the activation of MAPK cascades in response to stress. ABRECE1HVA22 contains ABRE element that promotes SSRGs regulation through ABA, as well as DPBFCOREDCDC3, which also interacts with bZIP TF for stress-responsive gene induction (Finkelstein and Lynch, 2000). ABA has been reported to display antagonistic effect on various plant processes such as seed development, germination, and seedling growth (Shu *et al.*, 2018). Relatively high concentrations of ET and ABA in inferior spikelets suppress the expression of starch synthesis-related genes and their enzyme activities, consequently leading to a low grain-filling rate (Zhu *et al.*, 2011).

### 3.6 Differences in SSRGs CREs motif occurrence between Indica and Japonica subspecies

#### 3.6.1 Abiotic/biotic stress

Differences in gene regulation between Japonica and Indica subspecies can be observed in relation to the frequency and diversity of CREs motifs between the SSRGs genes.

For dehydration stress, Indica rice SSRGs genes present higher frequency and diversity of CREs (Supplementary file-table S7), suggesting greater complexity in regulation of Indica genotypes for grain quality face to this stress, as previously reported (Lenka et al., 2011; Illey & Ludlow, 1996). Furthermore, higher frequency of CREs can also be observed in Indica rice for stress related to light and diseases/pathogens, but not necessarily, diversity was observed. Salinity and submergence anaerobiosis CREs seem to be little more frequent in Japonica genotypes than in Indica, being the first stress, which presents more diversity of GT1GMSCAM4 and AGCBOXNPGLB motifs. For the stress caused by extreme temperatures, the two subspecies have practically the same occurrence between gene promoters, indicating to be similarly regulated.

### 3.6.2 Mineral stress

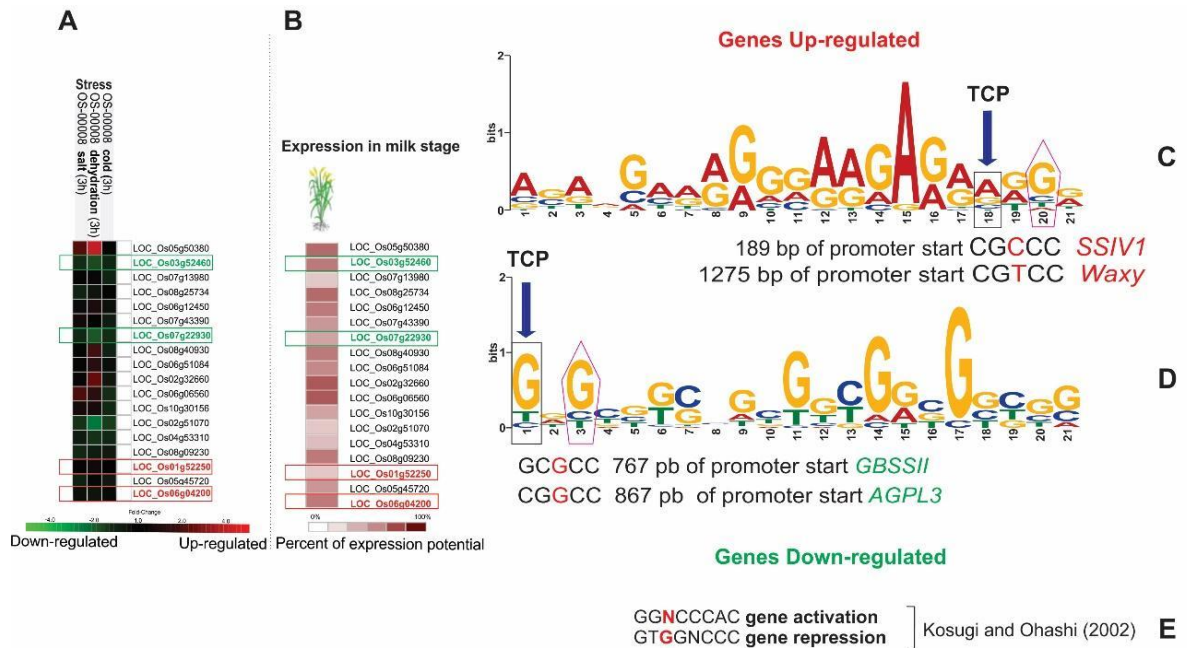
Higher frequency in occurrence of CREs can be observed in the subspecies Indica in response to nitrogen, which may indicate more complexity in the regulation of starch genes, although only the motifs GLMHVCHORD and NODCON2GM are observed. A difference in the efficient use of nitrogen has already been reported between the two subspecies (Gao et al., 2019), related to the allelic variation of OsNR2 being the Indica rice allele with the most activity. Motifs related to phosphate, iron, calcium, sulfur and ammonium do not show differences between the subspecies, as they have little diversity and frequency.

## 3.7 SSRGs expression in dehydration, cold and salinity stresses

Rice varieties are affected by abiotic stresses as dehydration, salinity and cold. These stresses often occur in combination, and stress responsive pathways often show extensive crosstalk (Wang, 2005; Mittler, 2006; Ismail *et al.*, 2013; Cruz *et al.*, 2013; Krannich *et al.*, 2015; Zhou *et al.*, 2016).

These stresses can directly affect grain quality being able to inactivate the expression of certain SSRGs. Our analyzes using RNA-seq and microarray data show that *AGPL3* and *GBSSII* genes are downregulated in dehydration, salt and cold stresses, while *SSIV1* and *Waxy* genes are upregulated in plants facing the same three stresses. *AGPL3*, *GBSSII*, *SSIV1* and *Waxy* are upregulated in *Oryza* grains at milk stage (Figure 4). Knowing this, we tried to identify, in the promoters of these and of other 74 genes that presented the same behavior under the same conditions, a binding region which could be the key to this upregulation. When comparing downregulated with upregulated genes a SNP with potential for use in genetic engineering in a TFBS is found (Figure 4C and D).





**Figure 4. Transcriptional analysis of Starch Synthesis-Related Genes and factors that influence it. A** - Transcriptional expression analysis of Starch Synthesis-Related Genes (except *Pul*, data not available) under dehydration, cold and salt stresses. The green rectangles represent *AGPL3* and *GBSSII* (downregulated), and red rectangles *SSIV1* and *Waxy* (upregulated). **B** - Expression analysis of SSRGs in grains (milk stage) of *Oryza sativa ssp. japonica*. Rectangles indicate the expression of *AGPL3*, *GBSSII*, *SSIV1* and *Waxy*, as described previously. **C** - Motif analysis for 74 upregulated *O. sativa japonica* genes, including *SSIV1* and *Waxy*. The arrow with rectangle indicates the TCP transcription factor binding site, while the pentagon indicates the nucleotide that seems to determine the down- or upregulation of the gene. Under this motif analysis it is possible to see the transcription start site and the promoter binding sequence for *SSIV1* and *Waxy*. **D** - Motif analysis for 74 downregulated *O. sativa japonica* genes, including *AGPL3* and *GBSSII*. The arrow with rectangle indicates the TCP transcription factor binding site, while the pentagon indicates the nucleotide that seems to determine the down- or upregulation of the gene. Under this motif analysis it is possible to see the transcription start site and the promoter binding sequence for *AGPL3* and *GBSSII*. **E** - Sequence and base that determine the gene up or downregulation according to Kosugi and Ohashi (2002).

Among the 74 promoters of upregulated genes analyzed in plants under abiotic stresses, we found TCP (Teosinte branched1/Cycloidea/Proliferating cell factor) TFBS. This TFBS is similar to what was found in the promoter of other 74 downregulated genes. In promoters of downregulated genes, the TFBS sequence presents a guanine base in the third position, which is not found in the TFBS of upregulated genes (Figure 3C and D). This result has also been found by Kosugi and Ohashi (2002) and Sharma *et al.*, (2010), that show that the key to up- or downregulation of growth/development associated genes can be observed in this TCP binding site, where a guanine appears to be responsible for gene downregulation. TCP transcription factors are involved in three sub-stages of early panicle development and responses to dehydration, salt and cold stresses (Sharma *et al.*, 2010). In our analysis, we observe that some *Oryza* genes present a key SNP that is probably associated with the alteration of the transcriptional expression in genes of rice plants that are under stressful conditions like dehydration, cold or high salt concentrations.

#### **4. Conclusions**

Here we raise new possibilities for plant breeding, highlighting sequence motifs probably related with the modification of the expression of genes involved in starch related pathways. The information provided here helps us understanding SSRGs regulation in rice plants under stressful conditions, something that influences grain quality of *Oryza* genotypes. The role of CREs as well as the signaling cascade involving SSRGs are discussed. SSRGs share common stress related CREs, indicating coordinated regulation of these genes, mainly for dehydration, light, anoxia, pathogen, and salt response. Also, other motifs particularly related to other less common stresses are present. These can also bind to many other TFs, subjecting the gene to multiple regulatory controls. The information provided here can help breeders in better understanding the transcriptional expression of SSRGs and further analysis on these elements will help us increasing and maintaining grain quality in stress-tolerant rice genotypes.

#### **Conflict of Interest**

The authors declare that they have no conflict of interest.

#### **Acknowledgement**

This work was supported by the Brazilian Ministry of Science and Technology, National Counsel of Technological and Scientific Development (CNPq); Coordination for the Improvement of Higher Education Personnel (CAPES) and RS State Foundation for Research Support (FAPERGS).

#### **Supplementary materials**

Supplementary material associated with this article can be found online: Table S1: Features of SSRGs and their function and expression in starch biosynthesis. Table S2: Types of CREs identified in SSRGs. Table S3: Biotic and Abiotic CREs identified in AGPase genes. Table S4: Biotic and Abiotic CREs identified in DBE genes. Table S5: Biotic and Abiotic stress CREs identified in SBE genes. Table S6: Biotic and Abiotic stress CREs identified in Starch synthase genes (SS).

#### **References**

- Abe H, Urao T, Ito T et al (2003) Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Proc Natl Acad Sci USA* 100(1):63-68.
- Abe H, Yamaguchi-Shinozaki K, Urao T et al (1997). Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *Proc Natl Acad Sci USA* 94(10): 1859-1868.

- Adu-Kwarteng E, Ellis WO, Oduro I et al (2003). Rice grain quality: a comparison of local varieties with new varieties under study in Ghana. *Food Control* 14: 507– 514.
- Agarwal P, Reddy MP, Chikara J. (2011) WRKY: its structure, evolutionary relationship, DNA-binding selectivity, role in stress tolerance and development of plants. *Mol Biol Rep.* 38(6):3883-96.
- Ahmed N, Maekawa M, Tetlow IJ (2008) Effects of low temperature on grain filling, amylose content, and activity of starch biosynthesis enzymes in endosperm of basmati rice. *Aust J Ag Res* **59**:599-604.
- Ahmed F, Abro TF, Kabir MS et al (2020) Rice Quality: Biochemical Composition, Eating Quality, and Cooking Quality. In: Costa de Oliveira A., Pegoraro C., Ebeling Viana V. (eds) *The Future of Rice Demand: Quality Beyond Productivity*. Sprin, Cham.
- Altschul SF, Gish W, Miller W, et al (1990) Basic Local Alignment Search Tool. *J. Mol. Biol.* 215:403-410.
- Asai T, Stone JM, Heard E et al (2000). Fumonisin B1-induced cell death in *Arabidopsis* protoplasts requires jasmonate-, ethylene-, and salicylate-dependent signaling pathways. *P Cell* 12: 1823-1836.
- Asaoka M, Okuno K, Sugimoto Y et al (1985) Developmental changes in the structure of endosperm starch of rice (*Oryza sativa* L.). *Agric. Biol. Chem* ,49:1973–1978.
- Bao J, Zheng SXW, Xia YW, et al (2000) QTL mapping for the paste viscosity characteristics in rice (*Oryza sativa* L.). *Theor. Appl. Genet.* 100: 280– 284.
- Barsain BL, Yadav SK & Hallan V (2019) Promoter and methylation status analysis revealed the importance of *PkGES* gene in picroside biosynthesis in *Picrorhiza kurrooa*. *J. Plant Biochem. Biotechnol.* 28, 424–436.
- Beckles DM, Tananuwong K, Shoemaker CF et al (2012) Starch characteristics of transgenic wheat (*Triticum aestivum* L.) overexpressing the Dx5 high molecular weight glutenin subunit are substantially equivalent to characteristics of transgenic wheat (*Triticum* those in nonmodified wheat. *J. Food Sci.* 77:C437–C442.
- Beckles DM and Thitisaksakul M (2014). How environmental stress affects starch composition and functionality in cereal endosperm. *Starch - Stärke*, 66: 58-71.
- Berendzen KW, Weiste C, Wanke D et al (2012). Bioinformatic *cis*-element analyses performed in *Arabidopsis* and rice disclose bZIP- and MYB-related binding sites as potential AuxRE-coupling elements in auxin-mediated transcription. *BMC Plant Biol* 12:125.
- Boyle B, Brisson N. (2001) Repression of the defense gene PR-10a by the single-stranded DNA binding protein SEBF. *P. Cell* 13(11):2525-37.

- Brodersen P, Petersen M, Pike HM et al (2002). Knockout of Arabidopsis ACCELERATED-CELL-DEATH1 encoding a sphingosine transfer protein causes activation of programmed cell death and defense. *Genes Dev.* 16: 490-502.
- Cameron DK, and Wang Y. (2005) A better understanding of factors that affect the hardness and stickiness of long-grain rice. *Cereal Chem.* 82: 113– 119.
- Cameron DK, Wang YJ, Moldenhauer KA (2008) Comparison of physical and chemical properties of medium-grain rice cultivars grown in California and Arkansas. *J. Food Sci.* 73: C72– C78.
- Chen Y, Wang M, & Ouwerkerk PB (2012). Molecular and environmental factors determining grain quality in rice. *Food Energy Sec.* 1(2):111-132.
- Chen W, Wang W, Lyu Y et al (2020). OsVP1 activates Sdr4 expression to control rice seed dormancy via the ABA signaling pathway. *Crop J.*
- Cruz RP, Sperotto RA, Cargnelutti D, et al (2013) Avoiding damage and achieving cold tolerance in rice plants. *Food Energy Sec.* 2:96-119.
- Damaris RN, Lin Z, Yang, P et al (2019). The Rice Alpha-Amylase, Conserved Regulator of Seed Maturation and Germination. *Int. Jf Mol Sci.* 20(2):450.
- De Freitas, KEJ.; dos Santos, RS.; Busanello, C; de Carvalho Victoria, F; Lopes, JL.; Wing, RA.; de Oliveira, AC. *Starch Synthesis-Related Genes (SSRG) Evolution in the Genus Oryza. Plants* 2021, 10, 1057.
- Feller A, Machemer K, Braun EL, Grotewold E. (2011) Evolutionary and comparative analysis of MYB and bHLH plant transcription factors. *Plant J.* 66(1):94–116.
- Finkelstein RR, Lynch TJ. (2000) The Arabidopsis abscisic acid response gene ABI5 encodes a basic leucine zipper transcription factor. *Plant Cell* 12(4):599-609.
- Fukunaga K, Fujikawa Y & Esaka M (2010). Light regulation of ascorbic acid biosynthesis in rice via light responsive cis-elements in genes encoding ascorbic acid biosynthetic enzymes. *Biosc. Biotec. Bioch.* 74(4): 888-891.
- Fusada N, Masuda T, Kuroda H. et al (2005) Identification of a Novel *Cis*-Element Exhibiting Cytokinin-Dependent Protein Binding *in Vitro* in the 5'-region of NADPH-Protochlorophyllide Oxidoreductase Gene in Cucumber. *Plant Mol Biol* 59:631–645.



- Gao Z, Wang Y, Chen G. *et al* (2019) The *indica* nitrate reductase gene *OsNR2* allele enhances rice yield potential and nitrogen use efficiency. *Nat Commun* 10, 5207.
- Ge S X, Jung D, Yao R (2020) ShinyGO: a graphical gene-set enrichment tool for animals and plants, *Bioinformatics*, Volume 36, Issue 8, 15 April 2020, Pages 2628–2629.
- Gubler F, and Jacobsen JV (1992). Gibberellin-responsive elements in the promoter of a barley high-pl  $\alpha$ -amylase gene. *P Cell* 4:1435-1441.
- Hasan Khan Z, Kumar B, Dhattewal P, Mehrotra, S. *et al* (2017). Transcriptional regulatory network of cis-regulatory elements (CREs) and transcription factors (TFs) in plants during abiotic stress. *Int J Plant Biol Res*.
- Hwang Y, Karrer E, Thomas B (2004). Three cis-elements required for rice  $\alpha$ -amylase Amy3D expression during sugar starvation. *P. Mol. Biol*, 36:331-341.
- Heidari P, Ahmadizadeh M, & Najafi-Zarrini H. (2015). In Silico Analysis of Cis-Regulatory Elements on Co-Expressed Genes. *J. Biol. Environ. Sci*, 9(25):1-9.
- Heinemann RJB, Fagundes PL, Pinto EA, *et al* (2005). Comparative study of nutrient composition of commercial brown, parboiled and milled rice from Brazil. *J. Food Compos. Anal.* 18:287–296.
- Himmelbach A, Liu L, Zierold U, Altschmied L, Maucher H, Beier F, *et al*. Promoters of the barley germin-like *GER4* gene cluster enable strong transgene expression in response to pathogen attack. *Plant Cell*. 2010;22:937–52.
- Hirano HY, and Sano Y (1998). Enhancement of Wx gene expression and the accumulation of amylose in response to cool temperatures during seed development in rice. *P Cell Physiol.* 39:807–812.
- Illangakoon TK, Marambe B, Keerthisena R *et al* (2018). Performance of anaerobic germination-tolerant rice varieties in direct seeding: effects on stand establishment, weed growth and yield under **different** seeding rates. *Trop. Agric. Res.* 29:276-287.
- Ishiguro S, Nakamura K. (1994) Characterization of a cDNA encoding a novel DNA-binding protein, SPF1, that recognizes SP8 sequences in the 5' upstream regions of genes coding for sporamin and beta-amylase from sweet potato. *Mol Gen Genet.* 244(6):563-71.
- Ismail AM, Singh US, Singh S *et al* (2013) The contribution of Submergence-Tolerant (Sub1) rice varieties to food security in flood-prone rainfed lowland areas in Asia. *Field Crops Res.* 152:83-93.

- Ishiguro S, and Nakamura K. (2004) Characterization of a cDNA encoding a novel DNA-binding protein, SPF1, that recognizes SP8 sequences in the 5' upstream regions of genes coding for sporamin and  $\beta$ -amylase from sweet potato. *Mol Gen Genet MGG* 244:563-571.
- Jain M, Nijhawan A, Arora R et al (2007) F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiol.* 143(4):1467-83.
- Jameson PE, and Song J (2016) Cytokinin: a key driver of seed yield, *J Exp Bot* 67(3):593–606.
- Kaur A, Pati PK, Pati AM et al (2017). In-silico analysis of cis-acting regulatory elements of pathogenesis-related proteins of *Arabidopsis thaliana* and *Oryza sativa*. *PLoS One* 12(9):e0184523.
- Kuromori T., Mizoi J., Umezawa T et al (2014) Drought Stress Signaling Network. In: Howell S. (eds) *Molecular Biology. The P. Sci.* 2 Springer, New York, NY.
- Kang HJ, Hwang IK, Kim KS et al (2006) Comparison of the physicochemical properties and ultrastructure of japonica and indica rice grains. *J Agric Food Chem* 54(13): 4833-4838.
- Kaplan B, Davydov O, Knight H, et al (2006). Rapid transcriptome changes induced by cytosolic  $Ca^{2+}$  transients reveal ABRE-related sequences as  $Ca^{2+}$ -responsive cis elements in *Arabidopsis*. *P. cell*, 18(10):2733–2748.
- Kosugi S, Ohashi Y, (2002). DNA binding and dimerization specificity and potential targets for the TCP protein family. *Plant J.* 30: 337-348.
- Krannich CT, Maletzki L, Kurowsky C et al (2015) Network candidate genes in breeding for drought tolerant crops. *Int. J. Mol. Sci.* 16:16378-16400.
- Lenka S, & Bansal KC (2019). Abiotic stress responsive cis-regulatory elements (CREs) in rice (*Oryza sativa* L.) and other plants. *New Phytol* 224:134–143.
- Lenka SK, Katiyar A, Chinnusamy V, Bansal KC (2011) Comparative analysis of drought-responsive transcriptome in Indica rice genotypes with contrasting drought tolerance. *Plant Biotechnol J*, 9: 315–327.
- Li, W. H., Shan, Y. L., Xiao, X. L., Zheng, J. M., et al., Effect of nitrogen and sulfur fertilization on accumulation characteristics and physicochemical properties of A- and B-wheat starch. *J. Agric. Food Chem.* 2013, 61, 2418–2425.

- Li L, Guo H, Wu N, et al (2015). P1BS, a conserved motif involved in tolerance to phosphate starvation in soybean. *Genet. Mol. Res.* 14(3): 9384-9394.
- Li N, Han X, Feng D (2019). Signaling Crosstalk between Salicylic Acid and Ethylene/Jasmonate in Plant Defense: Do We Understand What They Are Whispering? *Int. J. Mol. Sci.* 20:671.
- Liu K, Ye Y, Tang C et al (2008). Responses of ethylene and ACC in rice grains to soil moisture and their relations to grain filling. *Front. Agric. China* 2:172–180.
- Lilley JM, & Ludlow MM. (1996) Expression of osmotic adjustment and dehydration tolerance in diverse rice lines. *Field Crops Research*, 48 (2-3): 185-197.
- Llorens E, García-Agustín P, & Lapeña L (2017). Advances in induced resistance by natural compounds: towards new options for woody crop protection. *Scie. Agric.* 74(1):90-100.
- Lorenzo O, Chico JM, Sánchez-Serrano JJ et al (2004). JASMONATE-INSENSITIVE1 encodes a MYC transcription factor essential to discriminate between different jasmonate-regulated defense responses in *Arabidopsis*. *Plant Cell*, 16(7):1938-1950.
- Lucheta A, Lambais M. (2012) Sulfur in agriculture. *Rev. Bras. Ciência do Solo* 36: 1369–1379.
- Luo H, Song F, Goodman RM et al (2005) Up-regulation of OsBIHD1, a rice gene encoding BELL homeodomain transcriptional factor, in disease resistance responses. *Plant Biol (Stuttg)* 7(5):459-68.
- Maruyama-Nakashita A, Nakamura Y, Watanabe-Takahashi A, et al (2005). Identification of a novel cis-acting element conferring sulfur deficiency response in *Arabidopsis* roots. *Plant J.* 42(3):305-314.
- Mittler R (2006) Abiotic stress, the field environment and stress combination. *Trends Plant Sci.* 11:15-19.
- Mitsui T and Kimiko I (1997) The  $\alpha$ -amylase multigene family, *Trends P. Sci.* 2 (7):255-261.
- Nakase M, Aoki N, Matsuda T. et al (1997). Characterization of a novel rice bZIP protein which binds to the  $\alpha$ -globulin promoter. *P. Mol Biol* 33, 513–522.
- Ni Y, Wang Z, Yin Y et al (2012) Starch granule size distribution in wheat grain in relation to phosphorus fertilization. *J. Agr. Sci.* 150:45–52.
- Niemeyer J, Machens F, Fornfeld E et al (2011) Factors required for the high CO<sub>2</sub> specificity of the anaerobically induced maize GapC4 promoter in transgenic tobacco. *P. Cell Environ.* 34(2):220-9.

- Patel R, and Mohapatra PK. (1992). Regulation of spikelet development in rice by hormones. *J. Exp. Bot.* 43:257–262.
- Qu J, Xu S, Zhang Z et al (2018) Evolutionary, structural and expression analysis of core genes involved in starch synthesis. *Sci Rep* 8:12736.
- Rao P, Surekha MB, Gupta SR (2013) Effects of soil salinity and alkalinity on grain quality of tolerant, semi-tolerant and sensitive rice genotypes. *Rice Sci.* 20(4):284-291.
- Raymond P, Alani A, Pradet A (1985) ATP production by respiration and fermentation, and energy-charge during aerobiosis and anaerobiosis in 12 fatty and starchy germinating-seeds. *P. Phys.* 79:879—884.
- Raudvere U, Kolberg L, Kuzmin I et al (2019) Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update) *Nucleic Acids Research* .
- Sewelam N, Kazan K, & Schenk PM (2016). Global plant stress signaling: reactive oxygen species at the cross-road. *Front Plant Sci* 7:187.
- Singh N, Pal N, Mahajan G et al (2011). Rice grain and starch properties: Effects of nitrogen fertilizer application. *Carb. Poly.* 86(1), 219-225.
- Siscar-Lee JJ, Juliano JO, Qureshi RH (1990). Effect of saline soil on grain quality of rices differing in salinity tolerance. *Plant Foods Hum. Nutr.* 40:31–36.
- Sharifi P, Dehghani H, Mumeni A et al (2009) Genetic and genotype × environment interaction effects for appearance quality of rice. *Agric. Sci. China* 8: 891– 901.
- Sharma R, Kapoor MK, Tyagi A et al (2010). Comparative transcript profiling of TCP family genes provide insight into gene functions and diversification in rice and *Arabidopsis*. *J. Plant Mol. Biol. Biotechnol.* 1:24–38.
- Shu K, Zhou W, Chen F, (2018). Absciscic acid and gibberellins antagonistically mediate plant development and abiotic stress responses. *Front. Plant Sci.* 9:416.
- Supek F, Bošnjak M, Škunca N et al (2011) REVIGO summarizes and visualizes long lists of Gene Ontology terms. *PLoS ONE*.
- Thitisaksakul M, Jimenez RC, Arias MC et al (2012) Effects of environmental factors on cereal starch biosynthesis and composition. *J. Cereal Sci.* 56: 67– 80.



- Xue GP. (2003) The DNA-binding activity of an AP2 transcriptional activator HvCBF2 involved in regulation of low-temperature responsive genes in barley is modulated by temperature. *Plant J.* 33(2):373-83.
- Yamauchi D. (2001) A TGACGT motif in the 5'-upstream region of alpha-amylase gene from *Vigna mungo* is a cis-element for expression in cotyledons of germinated seeds. *P. Cell Physiol.* 42(6):635-41.
- Yu D, Chen C, & Chen Z. (2001). Evidence for an important role of WRKY DNA binding proteins in the regulation of NPR1 gene expression. *Plant cell*, 13(7):1527–1540.
- Zhang ZL, Xie Z, Zou X, et al (2004). A rice WRKY gene encodes a transcriptional repressor of the gibberellin signaling pathway in aleurone cells. *P Phys*, 134(4):1500–1513.
- Zhang BJ, Zhang HP, Chen QZ et al (2016) Molecular cloning and analysis of a receptor-like promoter of Gbvdr3 gene in sea island cotton. *Genet Mol Res: GMR.* 15(2).
- Zeng D, Tian Z, Rao Y et al (2017) Rational design of high-yield and superior-quality rice. *Nat. Plants* 3: 17031.
- Zhen-zhen C, Gang P, Fu-biao W et al. (2015) Effect of high temperature on the expressions of genes encoding starch synthesis enzymes in developing rice endosperms. *J Int Agric*, 14(4):642-659.
- Zhou Y, Yang P, Cui F, et al (2016) Transcriptome analysis of salt stress responsiveness in the seedlings of Dongxiang wild rice (*Oryza rufipogon* Griff.). *PLoS ONE* 11:e0146242.
- Zhu G, Ye N, Yang J, et al (2011). Regulation of expression of starch synthesis genes by ethylene and ABA in relation to the development of rice inferior and superior spikelets. *J. Exp. Bot.* 62(11):3907-16.
- Wang G (2005) Agricultural drought in a future climate: Results from 15 global climate models participating in the IPCC 4<sup>th</sup> assessment. *Clim Dynam* 25:739-753.
- Wang Y, Frei M, Song Q et al (2011) The impact of atmospheric CO2 concentration enrichment on rice quality—A research review. *Acta Ecol. Sin.* 31: 277–282.
- Wasternack C, Song S (2017) Jasmonates: biosynthesis, metabolism, and signaling by proteins activating and repressing transcription. *J Exp Bot* 68(1):1303–1321.

## Supplementary materials:

**Table S1.** Features of SSRGs and your function and expression in starch biosyntheses.

Group	Genes	Referencia para RAP-DB/Ensembl ID	Location	Start	End	Size (pb)	Function (Tian et al., 2009)	Temporal endosperm expression (Ohdan et al., 2005; Hirose & Terao, 2004).
AGPase	<i>AGPL1</i>	Os05g0580000	Chro 5	28871811	28877272	5462	GC, AC	early expressers
AGPase	<i>AGPL3</i>	Os03g0735000	Chro 3	30099369	30104572	5204	GC, AC	early expressers
AGPase	<i>AGPL4</i>	Os07g0243200	Chro 7	7983125	7989283	6159	GC, AC	early expressers
AGPase	<i>AGPS2a</i>	Os08g0345800	Chro 8	15666389	15672583	6248	AC	early expressers
SS	<i>ALK</i>	Os06g0229800	Chro 6	6748398	6753302	4916	GT, AC, GC	late expressers
DBE	<i>DPEI</i>	Os07g0627000	Chro 7	25980423	25984853	4431	AC, GT?	early expressers
SS	<i>GBSSH</i>	Os07g0412100	Chro 7	12916883	12924202	7320	AC	early expressers
DBE	<i>ISA</i>	Os08g0520900	Chro 8	25893657	25900576	6920	AC, GT	steady expressers
DBE	<i>PUL</i>	Os04g0164900	Chro 4	4408357	4418889	10533	AC, GT	steady expressers
SBE	<i>SBE1</i>	Os06g0726400	Chro 6	30897378	30905803	8426	AC, GT	early and middle expressers
SBE	<i>SBE2</i>	Os04g0409200	Chro 4	20240211	20243460	3249	GC, GT	early and middle expressers
SBE	<i>SBE3</i>	Os02g0528200	Chro 2	19355790	19367127	11338	GC, GT	early and middle expressers
SS	<i>SSI</i>	Os06g0160700	Chro 6	3079296	3086808	7513	AC	steady expressers
SS	<i>SSH 1</i>	Os10g0437600	Chro 10	15673243	15681075	7833	GT	steady expressers
SS	<i>SSH 2</i>	Os02g0744700	Chro 2	31233292	31238210	4929	AC,GT	early expressers
SS	<i>SSH 1</i>	Os04g0624600	Chro 4	31751600	31759420	7821	AC,GT	early expressers
SS	<i>SSH 2</i>	Os08g0191433	Chro 8	5353697	5363276	9580	AC	late expressers
SS	<i>SSIV 1</i>	Os01g0720600	Chro 1	30032428	30041425	8998	AC,GT	steady expressers
SS	<i>SSIV 2</i>	Os05g0533600	Chro 5	26485770	26493983	8214	GT	steady expressers
SS	<i>Wx</i>	Os06g0133000	Chro 6	1765622	1770653	5032	AC, GC, GT	late expressers

**Table S2.** Types of CREs identified in SSRGs.

Arquivo disponibilizado de forma separada disponível em formato .xlsx.

**Table S3.** Biotic and Abiotic CREs identified in AGPase genes from *O. sativa japonica* specie.

		Motif	Sequence	Response	AGPL1	AGPL3	AGPL4	AGPS2a
Abiotic stress	Dehydration	MYB2AT	TAAGTG	Involved in regulation of genes that are responsive to water stress	0	0	0	1
Abiotic stress	Dehydration	DRBCRTCOREAT	RCCGAC	Core motif of DRE/CRT (dehydration-responsive element/C-repeat)	0	0	3	1
Abiotic stress	Dehydration	MYCATRD22	CACATG	Binding site for MYC	0	0	3	1
Abiotic stress	Dehydration	MYCATERD1	CATGTG	Necessary for expression of erd1 (early responsive to dehydration) in dehydrated	0	0	2	0
Abiotic stress	Dehydration	MYCONSENSUSAT	CANNTG	MYC recognition site found in the promoters of the dehydration-responsive	13	11	9	10
Abiotic stress	Dehydration	MYBIAT	WAACCA	MYB recognition site found in the promoters of the dehydration-responsive	1	2	3	4
Abiotic stress	Dehydration	MYBCORE	CNGTTR	Myb homolog is induced by dehydration stress	5	2	2	4
Abiotic stress	Dehydration	MYB2CONSENSUSAT	YAACKG	MYB recognition site found in the promoters of the dehydration-responsive	2	0	0	0
Abiotic stress	Dehydration	ACGTATERD1	ACGT	Required for etiolation-induced expression of erd1 (early responsive to dehydration)	5	14	7	3
Abiotic stress	Dehydration	ABRELATERD1	ACGTG	Induction by dehydration stress and dark-induced senescence	4	6	3	0
Abiotic stress	Dehydration	UPRMOTIFIAT	CCNNNCCACG	MYB recognition site found in the promoters of the dehydration-responsive	1	0	0	0
Abiotic stress	Dehydration	MYBST1	GGATA	Potato MYB homolog	0	1	1	3
Abiotic stress	Dehydration	CBFHV	RYCGAC	CBFs are also known as dehydration-responsive element	0	1	3	1
Abiotic stress	Dehydration	MYBP LANT	MACCWAAC	Plant MYB binding site	0	1	1	1
Abiotic stress	Dehydration	MYBPZM	CCWACC	MYB homolog	0	2	1	4
Abiotic stress	Light	SORLREP3AT	TGTATATAT	"Sequences Over-Represented in Light-Repressed Promoters	0	0	0	1
Abiotic stress	Light	INRNTFSADB	YTCANTYY	Responds to light through an initiator	5	0	2	2
Abiotic stress	Light	GT1CONSENSUS	GRWAAW	Consensus GT-1 binding site in many light-regulated genes	7	12	7	10
Abiotic stress	Light	SORLREP4AT	CTCCTAATT	Sequences Over-Represented in Light-Repressed Promoters	1	0	0	0
Abiotic stress	Light	IBOXCORE	GATAA	Conserved sequence upstream of light-regulated genes	1	1	2	3
Abiotic stress	Light	SORLIP2AT	GGGCC	Sequences Over-Represented in Light-Repressed Promoters	6	2	0	0
Abiotic stress	Light	SORLIP1AT	GCCAC	Sequences Over-Represented in Light-Repressed Promoters	4	2	13	2
Abiotic stress	Light	GT1CORE	GTTTAA	Nuclear protein factor GT-1 correlate with sequences required for light-dependent	1	0	0	2
Abiotic stress	Light	SV40CORENHAN	GTGGWWHG	Sequence-specific interactions of a pea nuclear factor with light-responsive elements upstream of the rbcS3A gene	1	0	0	0
Abiotic stress	Light	GATABOX	GATA	Light regulated	5	5	9	11
Abiotic stress	Light	IBOXCORENT	GATAAGR	light-responsive	0	1	0	1
Abiotic stress	Light	IBOX	GATAA	light-responsive	0	1	0	1
Abiotic stress	Light	TBOXATGAPB	ACTTTG	Mutations in the "Tbox" resulted in reductions of light-activated gene transcription	0	2	1	0
Abiotic stress	Light	PALBOXAPC	CCGTCC	light responsiveness	0	3	1	0
Abiotic stress	Light	BOXIIPCHS	ACGTGGC	Essential for light regulation	0	0	1	0
Abiotic stress	Light	LRENPCABE	ACGTGGCA	A positive light regulatory element in tobacco	0	0	1	0
Abiotic stress	Anaerobic	OCBP2ZMGAPC4	GTGGGCCCG	Anaerobiosis-specific	1	0	0	0
Abiotic stress	Anaerobic	ANAERO2CONSENSUS	AGCAGC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	2	2	2	0
Abiotic stress	Anaerobic	CURECORECR	GTAC	Involved in oxygen-response	4	14	8	2
Abiotic stress	Anaerobic	ANAERO1CONSENSUS	AAACAAA	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	0	3	0	0
Abiotic stress	Anaerobic	ANAERO3CONSENSUS	TCATCAC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	0	1	0	0
Abiotic stress	Low temperature	CCAATBOX1	CCAAT	promoter of heat shock protein genes	1	5	6	2
Abiotic stress	Low temperature	LTR1HVBTL49	CCGAAA	"LTRE-1" (low-temperature-responsive element)	0	2	2	1
Abiotic stress	Low temperature	LTRCOREATCOR15	CCGAC	Core of low temperature responsive element (LTRE)	0	1	4	0
Abiotic stress	Low temperature	LTRATLIT78	ACCGACA	low-temperature-responsive element	0	0	0	1
Abiotic stress	Pathogen	BOXLCOREDCPAL	ACCWWCC	Transcriptional activator of the carrot phenylalanine ammonia-lyase gene (DcPAL1) in response to elicitor treatment.	0	1	3	1
Abiotic stress	insects	WBOXINTERF3	CTGACY	NtWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	0	8	6	5
Abiotic stress	Pathogen	BIHDIOS	TGTCA	Disease resistance responses	2	1	1	4
Abiotic stress	Pathogen	MYBI LEPR	GTTAGTT	Tomato Pti4(ERF) regulates defence-related gene expression via GCC box	1	0	0	0
Abiotic stress	Pathogen	CACGTGMOTIF	CACGTG	Tomato Pti4(ERF) regulates defense-related gene expression via GCC box	1	4	1	0
Abiotic stress	Pathogen	GT1GMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced SCaM4 gene expression	0	4	1	4
Abiotic stress	Pathogen	SEBFCONSSTPR10A	YTGTCWC	(SEBF) gene found in promoter of pathogenesis-related gene	0	0	2	0
Abiotic stress	Pathogen	CTRMCAMV35S	TCTCTCTCT	The cauliflower mosaic virus 35S promoter extends into the transcribed region.	0	0	1	0
Abiotic stress	Pathogen	WBOXNTHN48	CTGACY	NtWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	0	0	1	0
Abiotic stress	Pathogen	ELRECOREPCRP1	TTGACC	EIRE (Elicitor Responsive Element)	0	0	0	1
Mineral stress	Sulfur deficiency	SURECOREATSULTR11	GAGAC	Core of sulfur-responsive element (SURE)	3	0	2	1
Mineral stress	Nitrogen	NODCON2GM	CTCTT	One of two putative nodulin consensus sequences	3	0	0	0
Mineral stress	Calcium	ABRERATCAL	MACGYGB	Ca(2+)-responsive unregulated genes	2	5	2	0
Mineral stress	Calcium	CGCBOXAT	VCGCGB	A calmodulin-binding	0	14	8	0
Mineral stress	Salt	ARR1AT	NGATT	ARR1 and ARR2 response regulators	21	15	15	9
Mineral stress	Salt	GT1GMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced	0	4	1	4



**Table S4.** Biotic and Abiotic CREs identified in DBE genes in *O. sativa japonica* specie.

	Motif	Sequence	Response	DPE1	ISA	PUL
Dehydration	MYB2AT	TAACGT	Involved in regulation of genes that are responsive to water stress	1	0	0
Dehydration	DRE/CRT/COREAT	RCCGAC	Core motif of DRE/CRT (dehydration-responsive element/C-repeat)	0	0	0
Dehydration	MYCATRD22	CACATG	Binding site for MYC	0	0	0
Dehydration	MYCATERD1	CATGTG	Necessary for expression of erd1 (early responsive to dehydration) in dehydrated	0	0	0
Dehydration	MYCCONSENSUSAT	CANNTG	MYC recognition site found in the promoters of the dehydration-responsive	11	10	22
Dehydration	MYB1AT	WAACCA	MYB recognition site found in the promoters of the dehydration-responsive	1	5	0
Dehydration	MYBCORE	CNGTTR	Myb homolog is induced by dehydration stress	3	4	6
Dehydration	MYB2CONSENSUSAT	YAACKG	MYB recognition site found in the promoters of the dehydration-responsive	2	1	21
Dehydration	ACGTATERD1	ACGT	Required for etiolation-induced expression of erd1 (early responsive to dehydration)	4	2	6
Dehydration	ABRELATERD1	ACGTG	Induction by dehydration stress and dark-induced senescence	3	0	1
Dehydration	UPRMOTIFIAT	CCNNCCACG	MYB recognition site found in the promoters of the dehydration-responsive	1	0	0
Dehydration	MYBST1	GGATA	Potato MYB homolog	5	2	0
Dehydration	CBFHV	RYCGAC	CBFs are also known as dehydration-responsive element	0	0	2
Dehydration	MYBPLANT	MACCWAMC	Plant MYB binding site	0	2	2
Dehydration	MYBPZM	CCWACC	MYB homolog	0	3	2
Light	INRNTPSADB	YTCANTYY	Responds to light through an initiator	1	3	2
Light	GT1CONSENSUS	GRWAAW	Consensus GT-1 binding site in many light-regulated genes	11	4	10
Light	IBOXCORE	GATAA	Conserved sequence upstream of light-regulated genes	3	2	4
Light	SORLIP2AT	GGGCC	Sequences Over-Represented in Light-Repressed Promoters	6	2	1
Light	SORLIP1AT	GCCAC	Sequences Over-Represented in Light-Repressed Promoters	1	1	4
Light	GT1CORE	GGTTAA	Nuclear protein factor GT-1 correlate with sequences required for light-dependent	1	0	0
Light	SV40CORENHAN	GTGGWWHG	Sequence-specific interactions of a pea nuclear factor with light-responsive elements upstream of the rbc3A gene	0	1	0
Light	GATABOX	GATA	Light regulated	8	5	10
Light	IBOXCORENT	GATAAGR	light-responsive	1	0	0
Light	IBOX	GATAA	light-responsive	1	2	0
Light	TBOXATGAPB	ACTTTG	Mutations in the "Tbox" resulted in reductions of light-activated gene transcription	0	1	0
Light	PALBOXAPC	CCGTCC	light responsiveness	1	0	0
Anaerobic	ANAERO2CONSENSUS	AGCAGC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	3	2	1
Anaerobic	CURECORECR	GTAC	Involved in oxygen-response	10	12	12
Anaerobic	ANAERO3CONSENSUS	TCATCAC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	0	2	0
Low temperature	CCAATBOX1	CCAAT	promoter of heat shock protein genes	2	4	1
Low temperature	LTRE1HVBLT49	CCGAAA	"LTRE-1" (low-temperature-responsive element)	0	0	1
Low temperature	LTRECOREATCOR15	CCGAC	Core of low temperature responsive element (LTRE)	2	2	0
Low temperature	CRTDREHVCBF2	GTGCAC	regulated by temperature	0	0	2
Pathogen	BOXLCOREDPCAL	ACCWWCC	Transcriptional activator of the carrot phenylalanine ammonia-lyase gene (DcPAL1) in response to elicitor treatment.	1	2	2
Insects	WBOXNTERF3	CTGACY	NWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	4	5	1
Pathogen	BIHD1OS	TGTCA	Disease resistance responses	2	2	6
Pathogen	CACGTGMOTIF	CACGTG	Tomato Pti4 (ERF) regulates defense-related gene expression via GCC box	2	0	0
Pathogen	GT1GMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced ScaM-4 gene expression	5	2	3
Pathogen	WBOXNTHCN48	CTGACY	NWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	1	0	0
Sulfur deficiency	SURECOREATSULTR11	GAGAC	Core of sulfur-responsive element (SURE)	3	0	0
Calcium	ABRERATCAL	MACGYGB	Ca(2+)-responsive upregulated genes	2	0	1
Ammonium	AMMORESIVDCRNIA1	CGAACTT	ammonium response	0	2	0
Iron	IRO2OS	CACGTGG	induced exclusively by Fe deficiency	1	0	0
Salt	ARRIAT	NGATT	ARR1 and ARR2 response regulators	17	13	13
Salt	GT1GMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced	5	2	3
Phosphate	P1BS	GNATATNC	Found in the upstream regions of phosphate starvation responsive genes	2	4	2

**Table S5.** Biotic and Abiotic stress CREs identified in SBE genes in *O. sativa japonica*.

	Motif	Sequence	Response	SBE1	SBE3
Dehydration	MYB2AT	TAACTG	Involved in regulation of genes that are responsive to water stress	1	0
Dehydration	DRECRTCOREAT	RCCGAC	Core motif of DRE/CRT (dehydration-responsive element/C-repeat)	0	1
Dehydration	MYCATRD22	CACATG	Binding site for MYC	1	5
Dehydration	MYCATERD1	CATGTG	Necessary for expression of erd1 (early responsive to dehydration) in dehydrated	1	1
Dehydration	MYCCONSENSUSAT	CANNTG	MYC recognition site found in the promoters of the dehydration-responsive	10	1
Dehydration	MYB1AT	WAACCA	MYB recognition site found in the promoters of the dehydration-responsive	4	0
Dehydration	MYBCORE	CNGTTR	Myb homolog is induced by dehydration stress	5	1
Dehydration	MYB2CONSENSUSAT	YAACKG	MYB recognition site found in the promoters of the dehydration-responsive	1	0
Dehydration	ACGTATERD1	ACGT	Required for etiolation-induced expression of erd1 (early responsive to dehydration)	5	2
Dehydration	ABRELATERD1	ACGTG	Induction by dehydration stress and dark-induced senescence	1	2
Dehydration	MYBST1	GGATA	Potato MYB homolog	1	3
Dehydration	CBFHV	RYCGAC	CBFs are also known as dehydration-responsive element	0	2
Dehydration	MYBPLANT	MACCWAMC	Plant MYB binding site	0	1
Dehydration	MYBPZM	CCWACC	MYB homolog	0	1
Light	INRNTPSADB	YTCANTYY	Responds to light through an initiator	6	3
Light	GT1CONSENSUS	GRWAAW	Consensus GT-1 binding site in many light-regulated genes	16	5
Light	IBOXCORE	GATAA	Conserved sequence upstream of light-regulated genes	2	4
Light	SORLIP1AT	GCCAC	Sequences Over-Represented in Light-Repressed Promoters	0	1
Light	GT1CORE	GGTTAA	Nuclear protein factor GT-1 correlate with sequences required for light-dependent	2	1
Light	GATABOX	GATA	Light regulated	15	13
Light	IBOX	GATAA	light-responsive	1	1
Light	TBOXATGAPB	ACTTTG	Mutations in the "Tbox" resulted in reductions of light-activated gene transcription	3	1
Anaerobic	ANAERO2CONSENSUS	AGCAGC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	1	5
Anaerobic	CURECORECR	GTAC	Involved in oxygen-response	8	6
Anaerobic	ANAERO3CONSENSUS	TCATCAC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	1	1
Low temperature	CCAATBOX1	CCAAT	promoter of heat shock protein genes	3	2
Low temperature	LTRECOREATCOR15	CCGAC	Core of low temperature responsive element (LTRE)	0	4
Pathogen	BOXLCOREDCPAL	ACCWWCC	Transcriptional activator of the carrot phenylalanine ammonia-lyase gene (DcPAL1) in response to elicitor treatment.	1	2
Insects	WBOXINTERF3	CTGACY	NtWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	5	6
Pathogen	BIHD1OS	TGTCA	Disease resistance responses	3	4
Pathogen	GT1GMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced SCaM-4 gene expression	4	1
Pathogen	WBOXNTCHN48	CTGACY	NtWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	1	3
Sulfur deficiency	SURECOREATSULTR11	GAGAC	Core of sulfur-responsive element (SURE)	1	5
Calcium	CGCGBOXAT	VCGCGB	A calmodulin-binding	0	2
Salt	ARR1AT	NGATT	ARR1 and ARR2 response regulators	1	9
Salt	AGCBOXNPGLE	AGCCGCC	NaCl-responsive	0	1
Salt	GT1GMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced	4	1
Phosphate	P1BS	GNATATNC	Found in the upstream regions of phosphate starvation responsive genes	2	0

**Table S6.** Biotic and Abiotic stress CREs identified in *Starch synthase genes* (SS) in *O. sativa japonica* specie.

	Motif	Sequence	Response	SSI	SSI1	SSI2	SSI3	SSI4	SSI5	SSI6	Wav	ALK	GBSSI
Dehydration	MYE2AT	TAACTG	Involved in regulation of genes that are responsive to water stress	1	2	0	5	0	0	1	0	1	0
Dehydration	DRECR/COREAT	ECGAC	Core motif of DRE/CRT/dehydration-responsive element/C-repeat	0	0	2	0	1	6	0	2	3	0
Dehydration	MYCATRD22	CACATG	Binding site for MYC	2	3	1	4	0	1	3	0	0	6
Dehydration	MYCATRED1	CATGTG	Necessary for expression of erd1 (early responsive to dehydration) in dehydrated	0	0	1	0	0	1	2	0	0	3
Dehydration	MYCCONSUSAT	CANNTG	MYC recognition site found in the promoters of the dehydration-responsive	14	8	12	21	11	6	13	8	10	24
Dehydration	MYE1AT	WAACCA	MYB recognition site found in the promoters of the dehydration-responsive	1	3	4	0	3	0	2	3	5	2
Dehydration	MYBCORE	CNGTTR	Myb homolog is induced by dehydration stress	5	6	2	1	6	2	4	5	5	6
Dehydration	MYECCONSUSAT	YAACRG	MYB recognition site found in the promoters of the dehydration-responsive	3	3	0	0	3	2	1	3	15	4
Dehydration	ACGTATERD1	ACGT	Required for etiolation-induced expression of erd1 (early responsive to dehydration)	12	4	4	11	4	2	7	20	16	3
Dehydration	ABREATERD1	ACGTG	Induction by dehydration stress and dark-induced defense	5	1	2	8	0	0	0	8	2	1
Dehydration	UPFMO TFIAT	CCNNCCACG	MYB recognition site found in the promoters of the dehydration-responsive	1	1	0	0	0	0	0	0	0	0
Dehydration	MYESTI	GGATA	Potato MYB homolog	4	1	1	1	2	2	0	0	3	4
Dehydration	CEPHV	RYCGAC	CEB are also known as dehydration-responsive element	1	2	2	9	1	8	1	6	3	0
Dehydration	MYEPLANT	MACCWAMC	Plant MYB binding site	0	2	0	0	0	1	0	0	0	0
Dehydration	MYEPM	CCWACC	MYB homolog	1	2	2	0	0	2	0	0	0	1
Light	SORLIP3AT	TGTATATAT	"Sequences Over-Represented in Light-Responsive Promoters"	0	0	0	0	0	0	0	0	0	0
Light	IRAYD SADB	YTCANTTY	Responds to light through an initiator	2	0	0	1	5	0	2	0	2	5
Light	GTCI CONSENSUS	GRWAAW	Consensus GT1 binding site in many light-regulated genes	14	9	4	1	14	6	14	9	10	7
Light	SORLIP4AT	CTCTAATT	Sequences Over-Represented in Light-Responsive Promoters	0	0	0	0	0	0	0	0	0	0
Light	IBOXCORE	GATGA	Conserved sequence upstream of light-activated genes	5	2	2	1	1	1	3	2	2	1
Light	SORLIP2AT	CGGCC	Sequences Over-Represented in Light-Responsive Promoters	0	2	1	2	0	6	3	1	4	1
Light	SORLIP1AT	CGCAC	Sequences Over-Represented in Light-Responsive Promoters	1	2	2	6	0	3	3	6	4	0
Light	GTCI CORE	GGTAA	Nuclear protein factor GT1 correlate with sequences required for light-dependent	4	1	2	0	1	0	1	0	0	0
Light	SV40COREENHAN	GTGGWWHG	Sequence-specific interactions of a peptide artifact with light-responsive elements upstream of the rbc3A gene	0	2	0	0	0	0	0	0	1	0
Light	GATBOX	GATA	Light-activated	5	4	12	5	12	2	12	6	5	5
Light	IBOXCORENT	GATAAGR	Light-responsive	0	0	0	0	0	0	0	0	0	1
Light	IBOX	GATAA	Light-responsive	0	0	0	0	0	0	1	0	0	1
Light	IBOXATGAPB	ACTTGG	Mutations in the "Box" result in reductions of light-activated gene transcription	0	0	0	1	0	0	0	0	1	0
Light	PALBOXAPC	CCGTCC	light responsiveness	1	0	0	4	1	1	1	2	0	2
Light	PIATGAPB	CAGCTCCATG	Result in reductions of light-activated gene	2	0	0	0	0	0	0	0	0	2
Light	BOXCPSAS1	CTCCAC	Light-induced transcriptional repression	0	1	0	0	0	0	0	1	0	0
Light	CDALATCAE2	CAAAACGC	dark response	0	0	0	0	0	0	0	0	0	1
Light	HEXAT	CCGTGG	light-responsive	1	0	0	0	0	0	0	0	0	0
Anaerobic	ANAERO2 CONSENSUS	AGCAGC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	1	2	9	0	2	1	0	3	0	4
Anaerobic	CURECOREJR	GTAC	Involved in oxygen response	6	4	18	16	4	8	14	12	4	6
Anaerobic	ANAERO1 CONSENSUS	AAACAAA	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	3	1	0	0	0	1	0	1	1	2
Anaerobic	ANAERO3 CONSENSUS	TCAATAC	One of 16 motifs found in silico in promoters of 13 anaerobic genes involved in the fermentative pathway	0	2	0	0	2	0	0	0	0	0
Low temperature	CCAAIBOX1	CCAAAT	promoter of heat shock protein genes	3	6	2	2	0	2	3	1	3	0
Low temperature	LTRH1HVL179	CCGAAA	"L TRE-1" (low temperature-responsive element)	0	1	0	1	1	0	0	0	1	1
Low temperature	LTRC1OREATCOR15	CCGAC	Core of low temperature-responsive element (L TRE)	0	0	2	7	1	8	1	2	3	0
Low temperature	LTRCA1L178	ACCGACA	low temperature-responsive element	0	0	1	1	1	0	0	0	1	0
Low temperature	CRD1REHVCBF2	GTGAC	regulated by temperature	0	0	0	2	0	2	2	2	0	0
Pathogen	BOXLCORED CPAL	ACWWCC	Transcriptional activator of the carotophenylalanine ammonia-lyase gene (DcPAL1) in response to elicitor treatment	1	4	0	0	0	1	0	0	0	0
Insects	WBOXNTERF3	CTGACY	NWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	4	5	9	3	3	2	3	4	4	5
Pathogen	EBHD10S	TGTCA	Disease resistance responses	2	2	4	2	3	1	9	4	1	7
Pathogen	MYH1EPF	GTTAGTT	Tomato Pti4 (ERF) regulates defense-related gene expression via GCC box	0	0	0	0	0	0	0	1	0	0
Pathogen	CACGTGMOIF	CACGTG	Tomato Pti4 (ERF) regulates defense-related gene expression via GCC box	2	0	0	0	0	0	0	2	2	2
Pathogen	GTLGMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced SCAM4 gene expression	7	0	1	0	4	0	3	1	2	0
Pathogen	SEEPCONSIPR10A	YTGTCWC	(SEEP) gene found in promoter of pathogenesis-related gene	8	3	1	0	1	0	4	0	0	0
Pathogen	CTRMCAV35S	TCTCTCTCT	The cauliflower mosaic virus 35S promoter extends into the transcribed region	1	0	0	0	0	0	0	1	0	0
Pathogen	WBOXNTERH48	CTGACY	NWRKYs possibly involved in elicitor-responsive transcription of defense genes in tobacco	1	0	3	0	0	0	1	0	1	0
Pathogen	ELRECOREPCRP1	TTGACC	ELRE (Elicitor Responsive Element)	0	0	2	0	0	0	2	0	1	0
Sulfur deficiency	SURECOREAT SUL TR11	GAGAC	Core of sulfur-responsive element (SURE)	0	2	2	2	0	1	1	3	2	0
Nitrogen	NODCON2GM	CTCTT	One of two putative nodulin consensus sequences	0	0	0	0	0	0	0	1	0	2
Calcium	ABREERATCAL	MACGYGB	C(2+) responsive upstream-activated genes	3	2	8	1	0	2	0	4	5	1
Calcium	CGCGBOXAT	VCGCGB	A calmodulin-binding	0	4	2	24	0	15	9	1	18	0
Ammonium	AMMOCORESDCRNTA1	CGAAGTT	ammonium response	0	0	0	1	0	1	0	0	0	0
Iron	IRO20S	CACGTGG	induced exclusively by Fe deficiency	0	0	0	0	0	0	0	0	1	1
Salt	ARE1 AT	NGATT	ARE1 and ARE2 response regulators	14	14	8	8	19	7	11	7	15	17
Salt	AGCBOXNPGLE	AGCCGCC	NaCl-responsive	0	0	0	0	0	1	0	0	0	0
Salt	GTLGMSCAM4	GAAAAA	Plays a role in pathogen- and salt-induced	7	0	1	0	4	0	4	1	2	0
Phosphate	P1ES	GNATATNC	Found in the upstream regions of phosphate starvation-responsive genes	2	0	2	0	2	0	0	0	0	0

## **4 CAPÍTULOS**

### **4.4 Artigo 4 – Mapping and analysis of plastid Simple Sequence Repeats in genus *Avena***

*A ser submetido na revista Crop Breeding and Applied Biotechnology*



# Mapping and analysis of plastid Simple Sequence Repeats in genus *Avena*

Karine E. Janner de Freitas<sup>1</sup>, Tatieli Silva Silveira<sup>1</sup>, Railson Schreinert dos Santos<sup>2</sup>, Antonio Costa de Oliveira<sup>1\*</sup>

1 Universidade Federal de Pelotas, Campus Universitário, S/N, 96160-000, Capão do Leão, RS, Brazil.

2 Instituto Federal de Educação, Ciência e Tecnologia Catarinense (IFC), Rodovia SC 283, s/n Fragosos, SC, 89703-720, Concórdia, SC, Brazil.

## Abstract

Oat is an important crop that grain offer nutritional components. Molecular characterization of wild relative germplasm using plastid (cp) markers, such as SNPs (Single Nucleotide Polymorphism) (cpSNPs) or SSRs (Simple Sequence Repeats) (cpSSRs), may be more informative than those using nuclear genomic tools and can assist breeders in separating and distinguishing between haplome groups. In this study we provide characterization, quantification and markers based in cpSSRs in 26 *Avena* species. This could be used in further research and characterization of species and populations inside the genus *Avena*.

**Keywords:** cpSSRs, chloroplast, breeding, oat.

## INTRODUCTION

Oat (*Avena* spp.) is one of the most important cultivated crops. Oat grains offer a great nutritional content, with high protein levels and healthy lipids (EFSA, 2010). Hence, there is a growing need to study diverse variety of oats in terms of their economically valuable traits and to search for new genotypes that can serve as the base for development of new varieties with high productivity and resistance to diseases (Gagkaeva et al., 2018).

Simple Sequence Repeats (SSRs or microsatellites) are a class of molecular markers based on tandem repeats of short (1 to 6 nucleotide) DNA sequences and are found in large quantities in both coding and non-coding regions of genomes (Zane et al., 2002). As they are highly polymorphic, they are particularly useful for cultivar fingerprinting, assessing genetic diversity of germplasm, and aiding in molecular breeding to improve crop traits characteristics (Zane et al., 2002).

The use of chloroplast SSR (cpSSR) markers can boost *Avena* breeding programs since SSRs can be transferred between genotypes within or between species (Liu et al., 2020). Earlier, cross-species transferability of SSRs was detected via PCR amplification in different related species (Shukla et al., 2018). Also, it can be especially useful in cultivar fingerprinting, in assessing germplasm diversity, and aiding in molecular breeding to improve crop characteristics (Shukla et al., 2018).

The objective of this work is to quantify and characterize the cpSSRs of 25 *Avena* species, in addition to providing primers that could be used in further research and characterization of species and populations inside the genus *Avena*.

## MATERIAL AND METHODS

### *In silico* identification of SSRs

The search of plastid SSRs was based in a previous study performed by Fu (2018). The plastid genome from the 26 *Avena* species (Table 1) are available in the National Center for Biotechnology Information (NCBI). The complete cpDNA was processed in SSRLocator (da Maia et al., 2008) to identify the microsatellite regions (cpSSR) as mono-, di-, tri-, tetra-, penta-, or hexanucleotide. We considered only those repeats in which the motifs repeated as follows: mononucleotide repeats with a repeat length  $\geq 8$ ; dinucleotide with a repeat length  $\geq 6$ ; and tri-, tetra-, penta-, and hexanucleotide with a repeat length  $\geq 3$ . After identification, we assessed microsatellites for coding (genic) and non-coding (intergenic) regions according to the information available in the NCBI database for each species.

Table 1: Plastid genomes of different *Avena* species with its size and NCBI accession number.

Plant species	Plastome size (pb)	Accession number
<i>Avena abyssinica</i>	135942 bp	NC 044158.1
<i>Avena agadiriana</i>	135945 bp	NC 044159.1
<i>Avena atlântica</i>	136006 bp	NC 044163.1
<i>Avena barbata</i>	135946 bp	NC 044173.1
<i>Avena brevis</i>	135939 bp	NC 044172.1
<i>Avena canariensis</i>	135955 bp	NC 044161.1
<i>Avena clauda</i>	135557 bp	NC 044167.1
<i>Avena damascena</i>	135925 bp	NC 044166.1
<i>Avena eriantha</i>	135560 bp	NC 044157.1
<i>Avena fatua</i>	135889 bp	NC 044170.1
<i>Avena hirtula</i>	135937 bp	NC 050395.1
<i>Avena hispanica</i>	135935 bp	NC 044164.1
<i>Avena hybrida</i>	135900 bp	NC 044148.1
<i>Avena insularis</i>	135967 bp	MG674209.1
<i>Avena longiglumis</i>	135728 bp	NC 044169.1
<i>Avena lusitanica</i>	135879 bp	NC 044149.1
<i>Avena maroccana</i>	135887 bp	NC 044162.1
<i>Avena murphyi</i>	135892 bp	NC 044174.1
<i>Avena muda</i>	135934 bp	NC 044147.1
<i>Avena occidentalis</i>	135893 bp	NC 044175.1
<i>Avena sativa</i>	135890 bp	NC 027468.1
<i>Avena sterilis</i>	135887 bp	NC 031650.1
<i>Avena strigosa</i>	135938 bp	NC 044171.1
<i>Avena vaviloviana</i>	135946 bp	NC 044168.1
<i>Avena ventricosa</i>	135681 bp	NC 044165.1
<i>Avena wiestii</i>	135944 bp	NC 044160.1

## RESULTS AND DISCUSSION

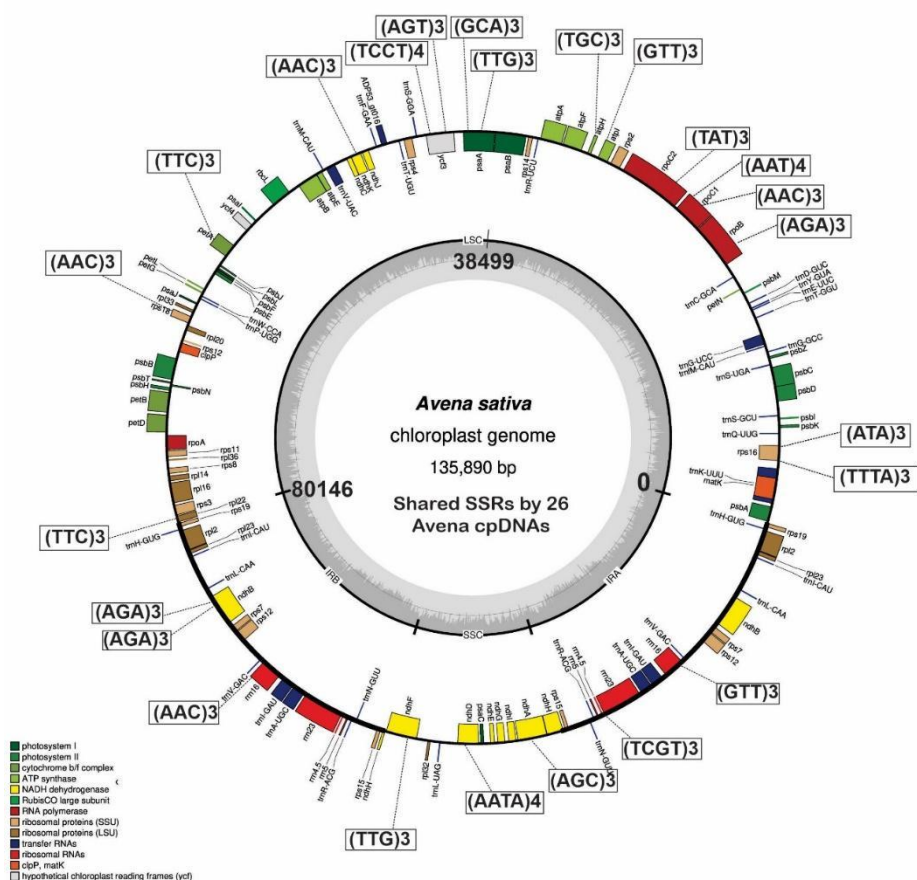
All cpSSRs of the 26 studied *Avena* species were evaluated. Therefore, we provide a general comparison between our results and published studies on other plant species. The total number of microsatellites observed for each species varied from 179 for *A. eriantha* to 187 for *A. atlantica*. While we identified mono-, di-, tri-, and tetranucleotide SSR for all species (Table 2), the majority of *Avena* cpSSRs were classified as mononucleotide followed by tri and tetranucleotide. Fu et al (2019) adopted different parameters as 8, 5, 4, 3, 3, and 3 for mono to hexa, consequently no have found trinucleotides. But unlike from this study, we have found tetra motifs, at least 1 in genic regions. We found predominantly mononucleotide motifs similarly to Liu et al. (2020) in *Avena* that the number of identified SSRs is double compared to what our study identified. Therefore, we provide a general comparison between our results and published studies on other plant species.

Table 2. Frequency (%) of the genic and intergenic cpSSRs based on motif size for each species.

Species	Mono		Di		Tri		Tetra		Penta		Hexa		Total
	Intergenic %	Genic %	Intergenic %	Genic %	Intergenic %	Genic %	Intergenic %	Genic %	Intergenic %	Genic %	Intergenic %	Genic %	
<i>Avena abyssinica</i>	96	35	1	0	18	24	9	1	0	0	0	0	184
<i>Avena agadiriana</i>	92	34	1	0	18	24	9	1	0	0	0	0	179
<i>Avena atlantica</i>	99	35	1	0	18	24	9	1	0	0	0	0	187
<i>Avena barbata</i>	96	35	1	0	18	24	9	1	0	0	0	0	184
<i>Avena brevis</i>	98	35	1	0	18	24	9	1	0	0	0	0	186
<i>Avena canariensis</i>	95	34	1	0	18	24	9	1	0	0	0	0	182
<i>Avena clauda</i>	90	33	1	0	19	25	10	1	0	0	0	0	179
<i>Avena damascena</i>	96	34	1	0	18	23	9	1	0	0	0	0	182
<i>Avena eriantha</i>	90	33	1	0	19	25	10	1	0	0	0	0	179
<i>Avena fatua</i>	93	34	1	0	18	24	10	1	0	0	0	0	181
<i>Avena hirtula</i>	96	35	1	0	18	24	10	1	0	0	0	0	185
<i>Avena hispanica</i>	98	35	1	0	18	24	9	1	0	0	0	0	186
<i>Avena hybrida</i>	92	34	1	0	18	24	10	1	0	0	0	0	180
<i>Avena insularis</i>	93	34	1	0	18	24	10	1	0	0	0	0	181
<i>Avena longiglumis</i>	93	34	1	0	18	24	9	1	0	0	0	0	180
<i>Avena lusitanica</i>	94	34	1	0	18	23	9	1	0	0	0	0	180
<i>Avena maroccana</i>	93	34	1	0	18	24	10	1	0	0	0	0	181
<i>Avena murphyi</i>	92	34	1	0	18	24	10	1	0	0	0	0	180
<i>Avena nuda</i>	98	35	1	0	18	24	9	1	0	0	0	0	186
<i>Avena occidentalis</i>	92	34	1	0	18	24	10	1	0	0	0	0	180
<i>Avena sativa</i>	92	34	1	0	18	24	10	1	0	0	0	0	180
<i>Avena sterilis</i>	92	34	1	0	18	24	10	1	0	0	0	0	180
<i>Avena strigosa</i>	98	35	1	0	18	24	9	1	0	0	0	0	186
<i>Avena vaviloviana</i>	96	35	1	0	18	24	9	1	0	0	0	0	184
<i>Avena ventricosa</i>	93	34	1	0	18	24	10	1	0	0	0	0	181
<i>Avena wiestii</i>	96	35	1	0	18	24	9	1	0	0	0	0	184

Most cpSSRs occur in the non-coding portion, being in coding portion there are mainly mononucleotides, trinucleotides, and tetranucleotides with only one motif in coding portion. Di-, penta-, and hexanucleotides do not have SSRs in this portion. Liu et al (2020) comments similar results only for mononucleotides and points out the lower polymorphism of coding regions in contrast to non-coding regions the factor for low distribution of SSRs in the coding portion. This is important for the mining and utilization of cpSSR as initiators in genetic analysis.

Primers were elaborated based in grasses parameters for primer designer (Saha et al., 2006) and are available in supplementary material (Table S3). The use of conventional gene cpDNA markers have been widely applied to studies of population history in trees and phylogenetic reconstruction at intraspecific levels because its special nature above (Kelchner, 2000). However, the use of cpSSRs has allowed the examination of speciation events at a finer level of detail than it was previously possible, but specifically for *Avena* species, have been used in separating and distinguishing between haplome (genome) groups (Li et al., 2009). The Figure 1 shows SSR regions that are shared by all *Avena* species that can be used as markers in haplotype study.



**Figure 1.** Chloroplast map with the SSRs found inside the genes and shared by *Avena* species. The motifs of cpSSRs are available in rectangle. Colored rectangles are genes that compose de chloroplast map. SSR motifs positions are identified by black lines.

Several studies have indicated that plastid genomes and cpSSRs have enormous potential (Zhang et al. 2017), being an important tool for plant biologists and breeders in assessing genetic diversity. Although Liu et al., (2020) and Fu (2018) reveal some features about cpSSRs in *Avena*, here we focus specifically on quantification, in addition to providing markers to be used in further studies.

# REFERENCES



Zhang H, Hall N, McElroy JS, Lowe EK and Goertzen LR (2017) Complete plastid genome sequence of goosegrass (*Eleusine indica*) and comparison with other Poaceae. *Gene* 5: 36-43.

Gagkaeva, T. Y., Gavrilova, O. P., Orina, A. S., Blinova, E. V., & Loskutov, I. G. (2018). Diversity of *Avena* species by morphological traits and resistance to *Fusarium* head blight. *Russian Journal of Genetics: Applied Research*, 8(1), 44-51.

Fu, YB. Oat evolution revealed in the maternal lineages of 25 *Avena* species. *Sci Rep* 8, 4252 (2018).

Da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FI, Costa de Oliveira A. SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation. *Int J Plant Genomics*. 2008;2008:412696.

EFSA Panel on Dietetic Products and Nutrition and Allergies (EFSA). Scientific opinion on the substantiation of a health claim related to oat beta-glucan and lowering blood cholesterol and reduced risk of (coronary) heart disease pursuant to article 14 of regulation (EC) no 1924/2006. *EFSA J*. 2010;8(12):1885.

Liu, Q., Li, X., Li, M., Xu, W., Schwarzacher, T., & Heslop-Harrison, J. S. (2020). Comparative chloroplast genome analyses of *Avena*: insights into evolutionary dynamics and phylogeny. *BMC plant biology*, 20(1), 1-20.

Saha, M.C., Cooper, J.D., Mian, M.A.R. *et al*. Tall fescue genomic SSR markers: development and transferability across multiple grass species. *Theor Appl Genet* 113, 1449–1458 (2006).

Li, W.T., Peng, Y.Y., Wei, Y.M., Baum, B.R., Zheng, Y.L. (2009). Relationships among *Avena* species as revealed by consensus chloroplast simple sequence repeat (ccSSR) markers, 56(4), 465-480.

Kelchner SA (2000) The evolution of non-coding chloroplast DNA and its application in plant systematics. *Ann Mo Bot Gard* 87, 482–498.

Zane, L., Bargelloni, L., & Patarnello, T. (2002). Strategies for microsatellite isolation: A review. *Molecular Ecology*, 11, 1-16.

Shukla, N., Kuntal, H., Shanker, A., & Sharma, S. N. (2018). Mining and analysis of simple sequence repeats in the chloroplast genomes of genus *Vigna*. *Biotechnology Research and Innovation*, 2(1), 9-18.

## Supplementary materials:

**Table S3.** *Primer* sequence of all *Avena* specie.

Arquivo disponibilizado de forma separada em formato .xmlx.

## 5 CONCLUSÃO GERAL

Fazendo uso das sequências disponíveis nos bancos de dados públicos, foi possível neste trabalho revelar que os genomas mitocondriais de plantas e algas são altamente abundantes em SSRs, na qual podemos observar um viés para os tipos e distribuição, mostrando as algas como o grupo mais diverso apresentando diferenças no número, abundância e densidade de SSRs entre espécies relacionadas, refletindo a dinâmica desse grupo para sobreviver as flutuações do ambiente durante a evolução. As plantas vasculares são as que apresentam inúmeros e longos motivos refletindo diferenças estruturais do mtDNA desse grupo. Apesar disso, observou-se que a conservação de certos motivos entre plantas e algas reflete um ancestral comum para esses grupos.

O estudo dos genes de síntese de amido (SSRGs) nas espécies selvagens de *Oryza* apontou que a ausência e duplicação de cópias de alguns motivos relatados no estudo podem modificar a atividade metabólica, denotando que o uso de diferentes espécies de arroz podem ser uma fonte rica de variabilidade para o melhoramento direcionado ao amido no arroz.

Os SSRGs compartilham elementos *cis* (CREs) comuns relacionados ao estresse, indicando a regulação coordenada desses genes, principalmente para estresses como desidratação, luz, anoxia, patógenos e resposta a salinidade. Além disso, outros motivos particularmente relacionados a outras tensões menos comuns estão presentes, indicando que esses genes podem responder a diversos estresses, mantendo sua atividade. Também oferecemos uma estratégia, através da identificação de um SNP, para ser testada e utilizada em programas de melhoramento para tolerância aos principais estresses que afetam a qualidade do grão em arroz.

A partir de genomas plastidiais de 26 espécies de Aveia conseguimos propor 52 iniciadores microsatélites para cada espécie que agora podem ser utilizados na distinção de espécies e populações dentro do gênero *Avena*.

## 6 REFERÊNCIAS

- AGARWAL, M., HAO, Y., KAPOOR, A., et al (2006). A R2R3 type MYB transcription factor is involved in the cold regulation of CBF genes and in acquired freezing tolerance. *J Biol Chem*, 281:37636–37645.
- AMBACHEW, D., MEKBIB, F., ASFAW, A., et al. (2015). Trait associations in common bean genotypes grown under drought stress and field infestation by BSM bean fly. *Crop J*, 3(4):305–16.
- AUDIL, G., LONE, A.A. & WANI, N.U.I (2019). Biotic and Abiotic Stresses in Plants, Abiotic and Biotic Stress in Plants. *IntechOpen*.
- BATRA, R., SARIPALLI, G., MOHAN, A., GUPTA, S. et al. (2017) Comparative Analysis of AGPase Genes and Encoded Proteins in Eight Monocots and Three Dicots with Emphasis on Wheat. *Front. Plant Sci*, 8, 19.
- BONETT, L. P. et al. (2006) Divergência genética em germoplasma de feijoeiro comum coletado no estado do Paraná, Brasil. *Semina: Ciências Agrárias*, 27(4), 547-560.
- BROOKS, A., JENNER, C. F., ASPINALL, D., (1982). Effects of water deficit on endosperm starch granules and on grain physiology of wheat and barley. *Aust. J. Plant Physiol*, 9, 423–436.
- BUDI, M., SABOTI, J., MEGLI, V., et al. (2013). Characterization of two novel subtilases from common bean (*Phaseolus vulgaris* L.) and their responses to drought. *Plant Physiol Biochem*, 62:79–87.
- DA-SILVA, P.R., MILACH, S.C.K., TISIAN, L.M. (2011). Transferability and utility of white oat (*Avena sativa*) microsatellite markers for genetic studies in black oat (*Avena strigosa*). *Genet. Mol. Res.* 10(4): gmr1232.
- DEVEY, D.S., CHASE, M.W., CLARKSON, J.J. (2009) A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. *Taxon* 58: 7-15.
- DUBOS, C.; STRACKE, R.; GROTEWOLD, E.; WEISSHAAR, B.; MARTIN, C.; LEPINIEC, L. (2010). MYB transcription factors in Arabidopsis. *Trends Plant Sci.* 15, 573–581.
- EFSA Panel on Dietetic Products and Nutrition and Allergies (EFSA)(2010). Scientific opinion on the substantiation of a health claim related to oat beta-glucan and lowering blood cholesterol and reduced risk of (coronary) heart disease pursuant to article 14 of regulation (EC) no 1924/2006. *EFSA J.*;8(12):1885.
- EMATER (2020). Disponível em <http://www.emater.tcche.br/site/>.



- GEORGE, B., BHATT, B.S., AWASTHI, M., GEORGE, B., SINGH, A.K. (2015). Comparative analysis of microsatellites in chloroplast genomes of lower and higher plants. *Curr Genet*, 61(4):665-77.
- GAGKAEVA, T. Y., GAVRILOVA, O. P., ORINA, A. S., BLINOVA, E. V., & LOSKUTOV, I. G. (2018). Diversity of Avena species by morphological traits and resistance to Fusarium head blight. *Russian Journal of Genetics: Applied Research*, 8(1), 44-51.
- GEORGELIS, N., BRAUN, E.L., HANNAH, L.C. (2008) Duplications and functional divergence of ADP-glucose pyrophosphorylase genes in plants. *BMC Evol. Biol.*, 8, 232.
- GUNARATNE, A., RATNAYAKA, U. K., SIRISENA, N., RATNAYAKA, J., et al. (2011) Erratum: Effect of soil moisture stress from flowering to grain maturity on functional properties of Sri Lankan rice flour. *Starch/Stärke*, 63, 392.
- HASAN S., HENRY R.J. (2020) Wild *Oryza* for Quality Improvement. In: Costa de Oliveira A., Pegoraro C., Ebeling Viana V. (eds) The Future of Rice Demand: Quality Beyond Productivity. *Springer*, Cham.
- JONAH, P., BELLO, L., LUCKY, O., MIDAU, A. & MORUPPA, S. (2011) Review: the importance of molecular markers in plant breeding programmes. *Glob J Sci Front Res*, 11:eV-vers1.
- KALIA, R.K., RAI, M.K., KALIA, S., SINGH, R. & DHAWAN, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica*, 177:309-334.
- KLEMPNAUER, K.H.; GONDA, T.J.; BISHOP, M.J. (1987) Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: The architecture of a transduced oncogene. *Cell*, 31, 453–463.
- OLIVEIRA, E. J., PÁDUA, J.G., ZUCCHI, M.I., VENCovsky, R., VIEIRA, M.L.C. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, 29(2) 294-307.
- LI, W.T., PENG, Y.Y., WEI, Y.M., BAUM, B.R., ZHENG, Y.L. (2009). Relationships among Avena species as revealed by consensus chloroplast simple sequence repeat (ccSSR) markers, 56(4), 465-480.
- LI, Y.C., KOROL, A.B., FAHIMA, T. & NEVO, E. (2004) Microsatellites within genes: structure, function, and evolution. *Mol Biol Evol* 21:991-1007.
- LIMA, R.A.Z., TOMÉ, L.M., ABREU, C.M.P. et al. (2014). Vacuum package: its effect on the hardening and darkening of the beans during storage. *Ciênc Rural*, 44(9):1664–70.

- LEMES, A.C., DE PAULA, L.C., BATISTA, A. et al. (2018). Potencial Antioxidante de Proteínas Extraídas de Feijão Comum (*Phaseolus vulgaris*) cv. BRSMG-Madrepérola. *UNICIÊNCIAS*, 22(3Esp), 38-42.
- LIU, D. H., ZHANG, J. L., CAO, J. H., WANG, Z. H., et al. (2010) The reduction of amylose content in rice grain and decrease of Wx gene expression during endosperm development in response to drought stress. *J. Food Agric. Environ.*, 8, 873–878.
- MA, Q., DAI, X., XU, Y., GUO, J., LIU, Y., et al. (2009) Enhanced tolerance to chilling stress in OsMYB3R-2 transgenic rice is mediated by alteration in cell cycle and ectopic expression of stress genes. *Plant Physiol*, 150:244–256.
- MASON, A.S. (2015) SSR Genotyping. In: Batley J (ed) Plant Genotyping. *Springer*, New York, NY, pp 77-89.
- MESQUITA, F.R., CORRÊA, A.D., ABREU, C.M.P.D., et al. (2007). Linhagens de feijão (*Phaseolus vulgaris* L.): composição química e digestibilidade protéica. *Ciência e Agrotecnologia*, 31, 1114-1121.
- MORGANTE, M., HANAFEY, M. & POWELL, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat Genet* 30:194-200.
- NJOROGE, D.M., KINYANJUI, P.K., et al. (2014). Extraction and characterization of pectic polysaccharides from easy- and hard-to-cook common beans (*Phaseolus vulgaris*). *Food Res Int.*, 64:314-22.
- PAZ-ARES, J., GHOSAL, D. WIENAND, U., PETERSON, P.A., SAEDLER, H. (1987). The regulatory c1 locus of Zea mays encodes a protein with homology to myb proto-oncogene products and with structural similarities to transcriptional activators. *The EMBO journal*, 6(12), 3553–3558.
- PASQUALI G, BIRICOLTI S, LOCATELLI F. (2008): OsMYB4 expression improves adaptive responses to drought and cold stress in transgenic apples. *Plant Cell Rep*, 27:1677–1686.
- PEDO, I., SGARBIERI, V. C., & GUTKOSKI, L. C. (1999). Protein evaluation of four oat (*Avena sativa* L.) cultivars adapted for cultivation in the south of Brazil. *Plant Foods for Human Nutrition*, 53(4), 297-304.
- PU, X., YANG, L., LIU, L. et al. (2020). Genome-wide analysis of the MYB transcription factor superfamily in *Physcomitrella patens*. *International journal of molecular sciences*, 21(3), 975.
- RAJENDRAKUMAR, P., BISWAL, A. K., BALACHANDRAN S. M. (2006) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformation*, 23, 1–4.

- RICACHENEVSKY, F.K., SPEROTTO, R.A.(2016) Into the Wild: *Oryza* Species as Sources for Enhanced Nutrient Accumulation and Metal Tolerance in Rice. *Front Plant Sci*, 29,7:974.
- SCHMUTZ, J., MCCLEAN, P.E., MAMIDI, S., et al. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nat Genet*, 46:707–13.
- SOARES, A.G. Consumo e qualidade nutritiva. (1996). In: Reunião nacional de pesquisa de feijão, 5., Goiânia. Anais... Goiânia: UFGO. v. 2, p. 73-79.
- SEO, P.J, XIANG, F., QIAO, M., et al. (2009). The MYB96 transcription factor mediates abscisic acid signaling during drought stress response in *Arabidopsis*. *Plant Physiol*, 151:275–289.
- STRACKE, R.; WERBER, M.; WEISSHAAR, B. (2001). The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.*, 4, 447–456.
- STEIN, J.C. YU, Y. COPETTI, D., et al. (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat.Genet.* 50, 285–296.
- SWINNEN, G., GOOSSENS, A., PAUWELS, L. (2019). Lessons from Domestication: Targeting *Cis*-Regulatory Elements for Crop Improvement. *Trends in Plant Science*, 24(11),1065.
- SHUFEN, C., YICONG, C., BAOBING, F., GUIAI, J. et al. (2019) Editing of rice isoamylase gene ISA1 provides insights into its function in starch formation. *Rice Sci.* 26, 77–87.
- SUN, Y.; JIAO, G.; LIU, Z.; ZHANG, X. et al. (2010) Generation of high-amylose rice through CRISPR/Cas9-mediated targeted mutagenesis of starch branching enzymes. *Front. Plant Sci*, 8, 298.
- TAUTZ, D. & RENZ, M. (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genomes. *Nucleic Acids Res* 12:4127-4138.
- VIEIRA, M. L. C., SANTINI, L., DINIZ, A. L., & MUNHOZ, C. D. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and molecular biology*, 39, 312-328.
- VENTER, M. & BOTHA, C.F. (2004). Promoter analysis and transcription profiling: Integration of genetic data enhances understanding of gene expression. *Physiologia Plantarum*. 120: 74-83.
- VENTER, M. & BOTHA, F.C. (2010). Synthetic promoter engineering. In: *Plant Developmental Biology - Biotechnological Perspectives*. Volume 2, E-C Pua & M. R. Davey. Berlin, 393- 414.

- YU, G.; OLSEN, K.M.; SCHAAL, B.A.(2011). Molecular evolution of the endosperm starch synthesis pathway genes in rice (*Oryza sativa* L.) and its wild ancestor, *O. rufipogon* L. *Mol. Biol. Evol*, 28, 659–671.
- YANHUI C, XIAOYUAN Y, KUN H, MEIHUA L, et al. (2006) The MYB transcription factor superfamily of *Arabidopsis*: expression analysis and phylogenetic comparison with the rice MYB family. *Plant Mol Biol*, 60:107–124.
- XU, Z., YU, M., YIN, Y. et al. (2020) Generation of selectable marker-free soft transgenic rice with transparent kernels by downregulation of SSSII-2. *Crop J*, 8, 53–61.
- ZHANG, L., YUAN, D., Y.U.S., LI, Z., CAO, Y., MIAO, Z., QIAN, H. & TANG, K. (2004) Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana* *Bioinformatics* 20:1081-1086.
- WALTER, M.; MARCHEZAN, E.; AVILA, L. A. (2008). Arroz: composição e características nutricionais. *Ciência Rural*, 38(4):1184-1192.
- WITTKOPP, P., & KALAY, G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* 13, 59–69.
- WRAY, G.A.; HAHN, M.W.; ABOUHEIF, E.; et al. (2003). The Evolution of Transcriptional Regulation in Eukaryotes. *Molecular Biology and Evolution*. 20: 1377-1419.
- WITTKOPP, P., KALAY, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet* 13, 59–69.