UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE AGRONOMIA
PROGRAMA DE PÓS GRADUAÇÃO EM FITOTECNIA

SEQUENCIAMENTO DO GENOMA DE ARROZ VERMELHO (*Oryza sativa* L.) E ANÁLISE DE GENES RELACIONADOS AO CARÁTER DEGRANE

THAÍS RAQUEL HAGEMANN
Engenheira Agrônoma / UTFPR
Mestre em Agronomia/ UFPEL

Tese apresentada como um dos requisitos a
obtenção do grau e Doutor em Fitotecnia
Ênfase em Melhoramento e Biotecnologia Vegetal

Porto Alegre (RS), Brasil
Agosto de 2015

CIP - Catalogação na Publicação

Hagemann, Thais Raquel
Sequenciamento do Genoma de Arroz Vermelho (Oryza Sativa L.) e Análise de Genes Relacionados ao Caráter Degrane / Thais Raquel Hagemann. -- 2015.
71 f.

Orientador: Jose Fernandes Barbosa Neto.

Tese (Doutorado) -- Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Programa de Pós-Graduação em Fitotecnia, Porto Alegre, BR-RS, 2015.

1. Sequenciamento. 2. Arroz Vermelho. 3. Montagem genômica. 4. Degrane. I. Barbosa Neto, Jose
Fernandes, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

THAÍS RAQUEL HAGEMANN
Engenheira Agrônoma - UTFPR
Mestre em Agronomia - UFPel

TESE

Submetida como parte dos requisitos

para obtenção do Grau de

DOUTOR EM FITOTECNIA

Programa de Pós-Graduação em Fitotecnia

Faculdade de Agronomia

Universidade Federal do Rio Grande do Sul

Porto Alegre (RS), Brasil

Aprovado em: 11.08.2015
Pela Banca Examinadora

Homologado em: 07.10.2016
Por

JOSÉ FERNANDES BARBOSA NETO
Orientador - PPG Fitotecnia

SIMONE MUNDSTOCK JAHNKE
Coordenadora do Programa de
Pós-Graduação em Fitotecnia

FERNANDO IRAJÁ FÉLIX DE CARVALHO
Aposentado/UFPel

FERNANDA BERED
PPG em Genética e Biologia
Molecular/UFRGS

ANTONIO COSTA DE OLIVEIRA
Agronomia – UFPel

PEDRO ALBERTO SELBACH
Diretor da Faculdade
de Agronomia

AGRADECIMENTOS

À Deus, pela vida, pela luz e proteção sempre.

Ao meu orientador, Dr. José Barbosa, pelos ensinamentos repassados, dedicação, conversas, incentivo, por acreditar em mim, pela amizade e por muitas vezes ser um pai. Foi muito bom esse tempo contigo, serei eternamente grata e levarei para sempre comigo na minha vida.

As pessoas mais importantes da minha vida, minha família, que sempre esteve presente em todos os momentos, dando apoio, palavras de conforto e incentivo e alegrias. Meus queridos pais, vocês são meu exemplo de vida, meu porto seguro e fortaleza. Obrigada por ter a casa e os braços sempre abertos, esperando por nós, para rir ou para chorar em conjunto. Também para a razão do meu viver, meus irmãos, vocês sabem o tamanho do meu amor por vocês!

Ao meu marido Cristiano, que nesses anos de vida em conjunto sempre me apoiou em todas as aventuras e nunca tremeu nas atribulações. Obrigada por sempre dizer: nós vamos passar por isso juntos! Pelo companheirismo, por acreditar em mim, pela confiança e pelo carinho e amor.

Ao professor Antonio Costa de Oliveira, que desde o mestrado sempre acreditou em mim, sempre servindo de exemplo na minha carreira profissional.

A todos os professores da UFRGS com os quais cursei disciplinas ou convivi, pelos ensinamentos repassados, discussões e idéias trocas, em especial aos professores Carla Delatorre, Itamar Nava, André Thomas, Aldo Merotto que entenderam os

infortúnios da vida de doutorando e sempre estiveram presentes dando apoio e conversas tranquilizadoras.

Ao técnico/amigo/compadre Fábio e a Carol Tessele que me auxiliaram no laboratório, não tenho palavras para agradecer! Sem vocês esse trabalho não existiria.

Ao Daniel Farias, pelo auxilio nas análises, por tirar as dúvidas e me socorrer em alguns momentos. Não tenho palavras pra te agradecer, mas tu sabes que sem você esse trabalho seria inexistente. Muito, muito obrigada!

As minhas amigas/os de longe Taciane Finatto, Daniela Priori, Eduardo Beche e a minha família portoalegrense Renata, Juliano, Jamile e em especial a Ciba e a Thanise, muito obrigada, vocês passaram por esse período comigo e sempre entenderam as ausências, o estresse e também comemoraram as alegrias. Than, obrigada por todo o Skype e pelo carinho! Vocês moram no meu coração.

A Alice Weber, Karina Pieretti e em especial a Marisa Bello, pela torcida, auxilio, e colaboração, vocês são indispensáveis e incomparáveis.

Aos amigos que fiz na França Emmanuelle, Samuel, Emilie e Emilie! Amenizaram a dor de estar longe, em especial a Dany e o Rafa, pessoas incríveis que foram minha família e meu suporte nesse tempo longe e que levarei pra sempre comigo!

Por todos os amigos e colegas que estiveram comigo nesses quatro anos, pelos momentos de descontração, de amizade e de conversas.

A sociedade brasileira e ao Cnpq e Capes pelas bolsas concedidas para a realização deste sonho e minha formação.

A todos aqueles, que contribuíram para esta formação, obrigada.

SEQUENCIAMENTO DO GENOMA DE ARROZ VERMELHO (*Oryza sativa* L.) E ANÁLISE DE GENES RELACIONADOS AO CARÁTER DEGRANE[1]

Autora: Thaís Raquel Hagemann
Orientador: José Fernandes Barbosa Neto

RESUMO

Até o momento nenhuma espécie daninha de importância agrícola tinha sido sequenciada no Brasil, dessa forma o arroz vermelho oferece uma oportunidade ímpar para isso, podendo levar a um melhor entendimento do seu genoma e facilitar o desenvolvimento de estratégias mais eficientes de controle destas plantas em condições de campo. Sendo assim, o objetivo deste trabalho foi sequenciar o genoma de dois genótipos de arroz vermelho coletados no estado do Rio Grande do Sul, e que possuem degrane elevado (AV60) intermediário (AV53), bem como realizar uma análise estrutural dos genes relacionados a este caráter. Vários programas montadores foram comparados e o Abbys foi o que gerou o maior número de *scaffolds*, totalizado um tamanho de genoma de 391,7 e 390,2 megabases, respectivamente para os genótipos AV53 e AV60, o que representa cerca de 90% do genoma referência do arroz. A cobertura do sequenciamento foi de 36.6 X e 32.1 X respectivamente para os mesmos genótipos. A análise estrutural de seis principais genes relacionados ao degrane relevou que a composição gênica é similar entre os genótipos analisados, corroborando com resultados de estudos anteriores. Um grande número de variantes genômicos foi descoberto (SNPs e INDELs). O alinhamento dos genótipos com o genoma de referência*Oryza sativa ssp. indica* revelou relativamente menor número de variantes, com frequência de um SNP a cada 264 pb. Por outro lado, o alinhamento contra o genoma referência de *Oryza sativa ssp. japonica* gerou uma frequência média de 1 SNP a cada 154pb, sendo que a mesma tendência foi observada para os INDELs. A menor frequência de SNPs e INDELs no primeiro alinhamento sugere maior similaridade entre as espécies alinhadas, indicando que o arroz vermelho do sul do Brasil é provavelmente originário da espécie*Oryza sativa spp indica*.

[1]Tese de Doutorado em Fitotecnia, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brasil. (71p.) Agosto, 2015.

GENOME SEEQUENCING OF WEEDY RICE (*Oryza sativa L.*) AND ANALYSIS OF GENES RELATED TO SEED SHATTERING[1]

Author: Thaís Raquel Hagemann
Advisor: José Fernandes Barbosa Neto

ABSTRACT

So far no weed species of agricultural importance had been sequenced in Brazil, thus weedy red rice provided a unique opportunity for it, leading to a better understanding of genomic and facilitating the development of more effective strategies to control these plants under field conditions. The objective of this study was sequencing the genome of two weedy red rice genotypes found in the Rio Grande do Sul state, which present high (AV60) and intermediate (AV53) degree of shattering,and perform a structural analysis of the genes related to this trait. Several genome assemblers programs were compared, and Abbys was the one that generated the highest number of scaffolds, totalizing a genome size of 391.7 and 390.2 Mb, respectively for AV53 and AV60 genotypes, which represent about 90% of the rice reference genome. The sequencing coverage was 32.1 and 36.6X respectively for the same genotypes. The structural analysis of 6 key genes related to seed shattering revealed that its genetic composition is similar among the genotypes analyzed, confirming results from previous studies. A large number of genomic variants (SNPs and INDELs) were discovered. The alignment of AV53, and AV60 with the *Oryza sativa ssp. indica* reference genome showed relatively lower number of variants, with a frequency of 1 SNP every 220 bp, whereas the alignment with the *Oryza sativa spp. japonica* yielded an average frequency of 1 SNP every 154pb; the same trend was observed for INDELS. The lower frequency of SNPs and INDELS in the first alignment suggests greater similarity between the aligned species, indicating that the origin of weedy red rice in southern Brazil is most likely from the *Oryza sativa spp indica*.

[1]Doctoral thesis in Agronomy, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil. (71 p.) August, 2015.

SUMÁRIO

Página

1. INTRODUÇÃO 1

2. REVISÃO BIBLIOGRÁFICA 3
2.1 Aspectos gerais e evolutivos do arroz cultivado 3
2.2 A cultura do arroz no Brasil 4
2.3 O arroz vermelho 6
2.4 O carater degrane 7
2.5 Sequenciamento de nova geração e montagem de genomas 10

3. MATERIAL E MÉTODOS 19
3.1 Material vegetal e extração de DNA 19
3.2 Sequenciamento 20
3.3 Montagem e alinhamento dos genomas 21
3.4 Análise estrutural de genes relacionados ao degrane 21
3.5 Detecção de SNPs e INDELs 22

4. RESULTADOS E DISCUSSÃO 23
4.1 Montagem e comparação dos genomas 23
4.2 Análise estrutural de genes relacionados ao degrane 29
4.3 Detecção de SNPs e INDELs 37

5. CONCLUSÕES 43

6. REFERÊNCIAS BIBLIOGRÁFICAS 44

7. APÊNDICES 53

RELAÇÃODE TABELAS

Página

1. Diferenças morfológicas entre as subespécies de *indica, japonica* e *javanica*. Porto alegre, 2015..... 4

2. Tamanho de leituras e rendimento das plataformas de sequenciamento sanger e de nova geração de sequenciamento (ngs). Porto alegre, 2015..... 12

3. Tamanho dos genomas e número aproximado de genes das pricipais de espécies cultivadas. Porto alegre, 2015. 15

4. Qualidade das leituras do sequenciamento dos genótipos de arroz vermelho AV53 e AV60 após limpeza pelo trimm. Porto alegre, 2015. 26

5. Resumo da montagem do genoma do acesso de arroz vermelho AV53, com detalhes das sequências, *contigs* e *scaffolds*. Porto alegre, 2015. 27

6. Resumo da montagem do genoma do acesso de arroz vermelho AV60, com detalhes das sequências, *contigs* e *scaffolds*. Porto alegre, 2015. 27

7. Estatísticas da detecção de snps e indels entre os genomas de arroz vermelho (AV53 e AV60) alinhados com os genomas refêrencia *O. japonica* e *O. Indica*. Porto alegre, 2015..... 38

RELAÇÃO DE FIGURAS

Página

1. Progresso do rendimento de grão do arroz cultivado no Brasil e no estado do Rio Grande do Sul de 1976 a 2015. Fonte: Conab (2015). Porto Alegre, 2015. 5

2. Ilustração do método de sequenciamento Illumina. Fonte: Tieppo (2014). Porto Alegre, 2015. 13

3. Ilustração das leituras geradas pelos sequenciamentos de nova geração. (A) Leituras simples, chamada de Single-end, a leitura é feita em apenas um lado (indicado pela seta em azul). (B) Nas leituras pareadas, paired-end, o sequenciamento é realizado em ambos os lados do fragmento, em sentidos contrários indicado pelas setas em azul). (C) Nas leituras mate-pair, são ligados nucleotídeos com biotina em ambos os lados dos fragmentos, e em seguida o fragmento é circularizado e cortado em sequências menores que são selecionadas com base na biotina para seguir com a leitura das sequências (Simples pela plataforma 454 e pareada pelas Illumina e SOLID. (HAMILTON; BUELL, 2012). Porto Alegre, 2015. 16

4. Boxplot mostrando a distribuição da qualidade das bases nas leituras para as amostras provenientes do sequenciamento dos genótipos AV53 e AV60 antes do refinamento. As áreas em verde, amarelo e vermelho nos gráficos indicam respectivamente elevada (≥28), intermediária (20 a 28) e baixa (<20) qualidade das leituras. Porto Alegre, 2015. 24

5. Boxplot mostrando a distribuição da qualidade das bases nas leituras para as amostras provenientes do sequenciamento dos genótipos AV53 e AV60 após o refinamento. As áreas em verde, amarelo e vermelho nos gráficos indicam respectivamente elevada (≥28), intermediária (20 a 28) e baixa (<20) qualidade das leituras. Porto Alegre, 2015. 25

6. Dotplot do locus do gene *Sh4*, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015. 30

7. Distância evolutiva inferida pelo método de agrupamento de vizinho mais próximo entre as sequências de proteína do gene de degrane *sh4* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1026 aminoácidos. Porto Alegre, 2015. 30

Página

8. Dotplot do locus do gene *OsCPL1*, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015. 31

9. Distância evolutiva inferida pelo método de agrupamento de vizinho mais próximo entre as sequências de proteína do gene de degrane *OsCPL1* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1246 aminoácidos. Porto Alegre, 2015. 32

10. Dotplot do locus do gene *Os01g0849100*, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015. 32

11. Distância evolutiva inferida pelo método de agrupamento vizinho mais próximo entre as sequências de proteína do gene de degrane *Os01g0849100* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1637 aminoácidos. Porto Alegre, 2015. 33

12. Dotplot do locus do gene *qsh1*, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015. 34

13. Distância evolutiva inferida pelo método de agrupamento de vizinho mais próximoentre as sequências de proteína do gene de degrane *qsh1* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1831 aminoácidos. Porto Alegre, 2015. 34

14. Dotplot do locus do gene *OsCel9D*, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015. 35

15. Distância evolutiva inferida pelo método de agrupamento vizinho mais próximo entre as sequências de proteína do gene de degrane *OsCeI9D* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1783 aminoácidos. Porto Alegre, 2015. 35

16. Dotplot do locus do gene *OsXTH8*, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015. 36

Página

17. Distância evolutiva inferida pelo método de agrupamento vizinho mais próximo entre as sequências de proteína do gene de degrane *OsxTH8* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 845 aminoácidos. Porto Alegre, 2015. 36

18. Diagrama de Venn com o numero de SNPs compartilhados entre os genomas dos acessos AV53, AV60, *Oryza indica* e *Oryza japonica.* Porto Alegre, 2015..... 39

19. Ocorrência de SNPS e INDELs encontrados nos alinhamentos do AV53 e AV60 contra os genomas de referência *O. japonica* e *O. indica*. Porto Alegre, 2015. 40

20. Gráfico mostrando frequências dos tipos de substituição, transversões (TV) e transições (TS). Porto Alegre, 2015. 41

21. Distribuição dos valores da relação TS/TV (substituição transição/transversão), ordenadas pela qualidade do sequenciamento ao longo do genoma. Porto Alegre, 2015. 42

1 INTRODUÇÃO

O arroz (*Oryza sativa* L.) cultivado é amplamente consumido e semeado em todos os continentes, desempenhando papel estratégico tanto no aspecto econômico, quanto no social. O Brasil está entre os dez principais produtores mundiais de arroz, com cerca de 12.4 milhões de toneladas anuais para um consumo de 11,7 milhões de toneladas (CONAB, 2015). No Rio Grande do Sul (RS), este cereal ocupa aproximadamente de 1,12 milhão de hectares, produzindo ao redor de 8,44 milhões de toneladas ao ano, com produtividade media de 7500 kg ha^{-1} (CONAB, 2015).

Apesar de essa produtividade ter aumentado nos últimos anos, ela ainda está abaixo da produtividade alcançada pelas lavouras que adotam alto nível tecnológico e do potencial obtido nas áreas experimentais. Isso ocorre, dentre outro fatores, devido ao controle insatisfatório das plantas daninhas. Entre as espécies daninhas que infestam as lavouras no sul do Brasil, o arroz vermelho (*Oryza sativa* L.) tem sido considerado como aquela que mais limita o potencial de produtividade do arroz. Assim sendo, é importante estudar a biologia desta planta invasora e entender a genética de sua adaptação aos ambientes cultivados com arroz.

A nova geração de tecnologias de sequenciamento (NGS) abre a oportunidade de redesenhar estratégias para a obtenção de maior eficiência na análise genética de caracteres de importância agronômica. Esta maior eficiência poderá refletir diretamente na seleção de genótipos com caracteres desejáveis, possibilitando ao melhorista desenvolver mais facilmente cultivares que atendam as necessidades do mercado. Da mesma forma, estas técnicas aplicadas a espécies daninhas poderão permitir o entendimento mais acurado de sua biologia, facilitando o estabelecimento de novas estratégias de seu controle a campo.

Até recentemente, o sequenciamento de genomas complexos era de difícil execução e de elevado custo. No entanto, após o desenvolvimento da nova geração de tecnologias de sequenciamento (NGS), essa tarefa ficou mais simplificada e com custos acessíveis, possibilitado a realização de técnicas de resequenciamento e sequenciamento

de novo.

Até o momento nenhuma espécie daninha de importância agrícola no Brasil foi sequenciada e o arroz vermelho oferece uma oportunidade ímpar para este trabalho. Primeiramente, por que ela é uma espécie invasora de uma cultura de grande importância econômica nacional e mundial: o arroz cultivado. Além disso, o arroz cultivado e suas espécies relacionadas têm sido intensamente estudadas, tanto do ponto de vista agronômico, como genético e genômico.

Dessa forma, a união de novas técnicas de sequenciamento com o estudo de caracteres adaptativos em arroz vermelho poderá se traduzir em um melhor entendimento da genética destes caracteres, facilitando o desenvolvimento de estratégias mais eficientes de controle destas plantas em condições de campo. Neste sentido, o caráter degrane é emblemático, uma vez que está diretamente relacionado com a manutenção de plantas de arroz vermelho a campo, sendo um fator adaptativo de grande importância. Além disso, futuramente o sequenciamento de variedades de arroz cultivadas no sul do Brasil possibilitará o estabelecimento de um banco de dados nacional. Este banco de dados poderá ser pesquisado para genes de interesse dos melhoristas, permitindo o isolamento e clonagem de genes diretamente a partir da análise *in silico*.

Sendo assim, os objetivos deste trabalho foram sequenciar e montar o genoma de dois genótipos de arroz vermelho coletados a campo no estado do Rio Grande do Sul, bem como analisar a estrutura de alguns genes relacionados ao caráter adaptativo degrane.

2 REVISÃO BIBLIOGRÁFICA

2.1 Aspectos gerais e evolutivos do arroz cultivado

O arroz (*Oryza sativa* L.) é uma espécie anual pertencente à família Poaceae e ao gênero *Oryza*, adaptada ao meio aquático. Esta adaptação se deve à formação do aerênquima no colmo e nas raízes da planta, tecido este que possibilita a passagem de oxigênio do ar para a camada da rizosfera, permitindo o cultivo em ambientes alagados (anaeróbios) (Taiz e Zeiger, 2009).

Esse cereal é originário do continente asiático (Molina et al 2011) sendo que tanto o arroz cultivado quanto o arroz daninho evoluíram a partir de espécies silvestres do gênero *Oryza*. Das 22 espécies silvestres do gênero *Oryza*, 9 são tetraplóides (BBCC, CCDD) e o restante é diploide (Khush, 1997). A diversificação em diferentes grupos desse gênero provavelmente ocorreu na China a cerca de 8.000 anos atrás (Molina et al 2011). As espécies silvestres *O. rufipogon*, *O. nivara*, *O. glumaepatula, O. meridionalis*, *O. breviligulata*, *O. longistaminata* e as espécies cultivadas *O. sativa* e *O. glaberrima* pertencem ao pool gênico diplóide (AA) e podem hibridizar entre si (Jena, 2010). Porém, ainda não está totalmente elucidado de qual espécie *O. sativa* evoluiu, sendo que alguns autores acreditam que essa espécie evoluiu a partir das espécies silvestres *O. nivarae*, *O. rufipogon* (Smith e Dilday, 2003).

Dentre o arroz cultivado *O. sativa* desenvolveu-se três subespécies principais: índica, japônica e javanica que sao detalhadas a seguir (Tabela 1).

TABELA 1. Diferenças morfológicas entre as subespécies de indica, japonica e javanica. Porto Alegre, 2015.

Subespécies	*Indica*	*Japonica Temperado*	*Japonica Tropical*
Folhas	Claras e longas	Escuras e estreitas	Claras e eretas
Perfilhamento	Elevado	Médio	Baixo
Estatura	Alta	Médio	Alta
Sensibilidade ao fotoperíodo	Variada	Ausente a baixa	Baixa
Grãos	Delgados	Curtos e arredondados	Longos e grossos
Aristas	Ausente	Variada	Variada
Degrane natural	Fácil	Difícil	Difícil

A domesticação em diferentes regiões climáticas da Ásia resultou na evolução de dois tipos de arroz japônica (Jena, 2010). O arroz japônica tropical que é cultivado no sul dos Estados Unidos e o arroz japônica temperado é cultivado no Japão e na Califórnia. No Brasil, o arroz cultivado na maioria das áreas pertence a subespécie índica. Já a espécie *O. glaberrima* é largamente cultivada no continente africano, de onde é originária (Delouche *et al.,* 2007), e é considerada daninha em outras regiões do mundo (Smith e Dilday, 2003).

Apesar do arroz vermelho e do arroz cultivado pertencerem a mesma espécie botânica, diferenças relacionadas ao degrane (Li *et al.*, 2006) e dormência fisiológica (Finkelstein *et al.*, 2008) tornam o arroz vermelho indesejável na lavoura pois resulta em prejuízos na produção do arroz cultivado. A dormência das sementes do arroz vermelho permite a sua germinação escalonada no tempo resultando na quase que perpetuação desta planta daninha uma vez estabelecida em uma lavoura.

No entanto, o degrane natural dos grãos de arroz vermelho tem maior importância em relação a características negativas desta planta daninha no sistema de produção do arroz. Este caráter tem importância evolutiva para a perpetuação do gênero *Oryza* (Lin *et al.*, 2007) e resulta em facilidade de disseminação, permanência e perpetuação das sementes de arroz vermelho em lavouras de arroz.

2.2 A cultura do arroz no Brasil

O arroz é um dos cereais mais cultivados no mundo com grande destaque do ponto de vista econômico e social, ocupando atualmente o segundo lugar como o cereal mais cultivado mundialmente. Essa cultura representa um dos alimentos mais importantes para a nutrição humana, servindo de base alimentar para mais de três bilhões de pessoas (SOSBAI, 2012). O Brasil está entre os dez principais países produtores de arroz, cuja

produção anual ficou entre 11 e 13 milhões de toneladas nas últimas safras, que corresponde a 82% da produção do Mercosul (SOSBAI, 2012) e cerca de 65% da produção brasileira na média das últimas 5 safras (CONAB, 2015).

De acordo com a série histórica da Conab (2015), nos útimos 40 anos o rendimento médio da cultura do arroz passou de 1501 para 5453 kg ha^{-1}representendo um ganho médio de 103 kg ha^{-1} por ano (Figura 1). Isso deve-se a melhoria nas tecnolgias de manejo da cultura e ganhos genéticos provenientes do constante lançamento de cultivares mais produtivos. Além de destacar-se como maior produtor nacional, o RS também apresenta historicamente as maiores produtvidades do país, com cerca de 2,5 toneladas acima do rendimento médio brasileiro.

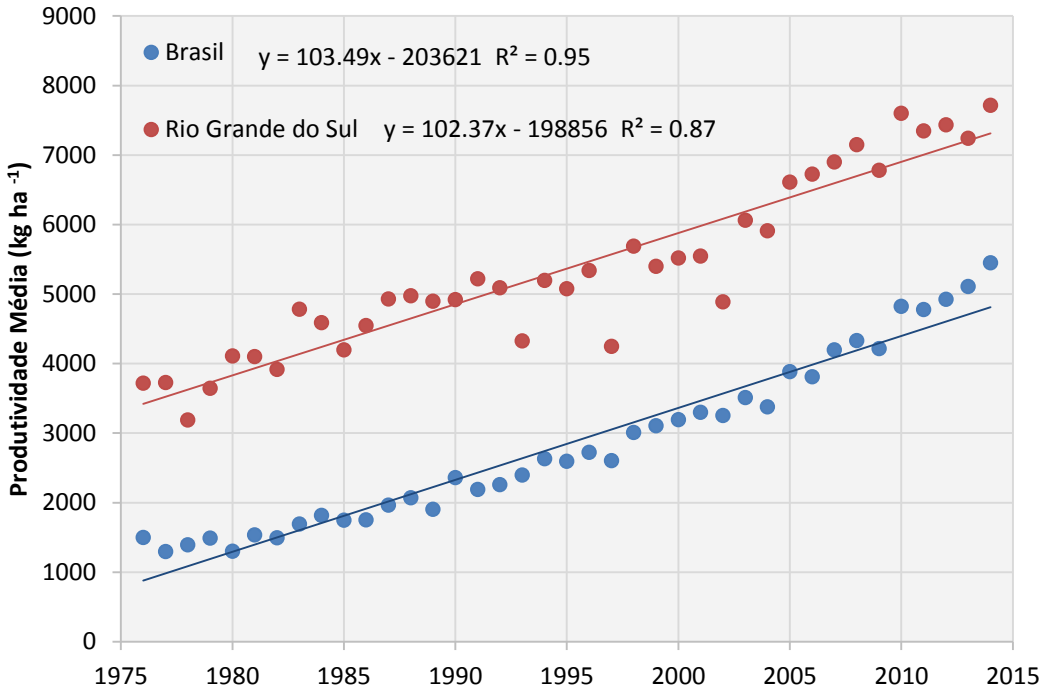


FIGURA 1. Progresso do rendimento de grão do arroz cultivado no Brasil e no estado do Rio Grande do Sul de 1976 a 2015. Fonte: Conab (2015). Porto Alegre, 2015.

Atualmente a área cultivada deste cereal no RS ocupa 1,12 milhões de hectares, com produtividade média de 7716 kg ha^{-1} na última safra (CONAB, 2015). É possível verificar aumento gradativo da produtividade média ao longo dos anos no RS, contudo, ela está aquém da produtividade média obtida em áreas experimentais e em lavouras que adotam alto nível tecnológico (Gomes e Magalhães, 2004). A elevada produção de arroz no Brasil e no mundo deve-se à alta tecnologia empregada pelos agricultores, associada

ao alto potencial genético das cultivares utilizadas, o qual foi alcançado pelo sucesso nos programas de melhoramento genético de arroz, além do manejo correto da cultura, com a utilização de sementes certificadas, época de semeadura e controle de pragas, molestias e plantas invasoras, principalmente o arroz vermelho.

2.3 O arroz vermelho

O arroz vermelho, também denominado arroz silvestre ou daninho, pertence ao mesmo gênero e espécie do arroz cultivado (*Oryza sativa*), porém há várias diferenças morfolológicas e genéticas entre estes (Noldin *et al.,* 1999). Essa invasora originária da Ásia, é considerada a mais importante planta infestante da lavoura orizícola do RS (Noldin *et al.,* 2004), em razão das perdas econômicas causadas pela diminuição da produtividade do arroz cultivado em função da competição entre essas espécies. Além de reduzir o rendimento de grãos, essa daninha também afeta a qualidade dos grãos colhidos, elevando os custos de produção devido à necessidade de controle e aos problemas operacionais na colheita, secagem e beneficiamento.

No sul do Brasil, os prejuízos relacionados à competição com arroz vermelho podem representar até 20% do rendimento de grãos. Dessa forma, aproximadamente 1,3 milhões de toneladas de arroz são perdidas no RS a cada safra, representando um prejuízo anual equivalente à aproximadamente 360 milhões de dólares (IRGA, 2008). Similarmente, na região orizícola do sul dos Estados Unidos, o arroz vermelho também se caracteriza por ser o principal problema em relação à competição com as cultivares de arroz (Gealy *et al*., 2002; Norsworthy *et al.,* 2007). Os prejuízos causados aos produtores do estado do Arkansas, EUA, em 2006 foram de aproximadamente U$ 300,00 por hectare (Burgos et al 2008).

O arroz vermelho corresponde a diversos biótipos silvestres da própria espécie *Oryza sativa* L e a biótipos de *O. nivara* e *O. rufipogon* (Vaughan *et al.,* 2001). Essa planta daninha diferencia-se do arroz cultivado por apresentar maior estatura, folhas decumbentes, elevado vigor e capacidade de afilhamento com emissão de afilhos ontogenicamente atrasados, pericarpo de cor avermelhada, pálea e lema com variação na cor, pilosidade e aderência da pálea e lema ao pericarpo, presença ou não de arista e sementes com dormência (Diarra *et al.,* 1985; Noldin *et al.,* 1999).

Em relação às características morfológicas, as plantas de arroz vermelho geralmente têm colmo com coloração levemente avermelhado, o que permite distinguí-las na lavoura (Agostinetto *et al.,* 2001). Além disso, o arroz vermelho apresenta elevada

debulha natural, desta forma, se estabelece com a cultura, compete com esta, mas tem a colheita de seus grãos impossibilitada devido à debulha precoce de seus grãos. O número de afilhos, a estatura e a massa seca da parte aérea das plantas de cultivares de arroz são afetados negativamente devido à competição com arroz vermelho (Fleck *et al.,* 2008), demonstrando a habilidade competitiva superior desta planta daninha. Biótipos de arroz vermelho procedentes de lavouras de arroz irrigado do RS e SC apresentaram alta variabilidade quanto às características de sementes e à intensidade dá duração da dormência (Schwanke *et al.,* 2008) podendo alguns biótipos apresentaram sementes com período de dormência de até 150 dias após a colheita.

O arroz vermelho é considerado a planta daninha com maior dificuldade de controle no cultivo do arroz irrigado, sendo que em áreas com altas infestações, caso não seja feito seu controle, as perdas podem chegar a 90% do rendimento de grãos (Avila *et al.,* 2000). Isto se deve ao fato de arroz vermelho possuir alta capacidade de competição, o que influência de forma negativa o desenvolvimento do arroz cultivado em diversas etapas da cultura (Fleck et al 2008). Durante a última década a porcentagem de áreas orizícolas infestadas com arroz vermelho aumentou cerca de 30% (IRGA, 2010) e atualmente 56% de todas as áreas das regiões de cultivo de arroz do RS possuem altas infestações, com população variando entre 5-30 planta sementes por m² (IRGA, 2010).

O arroz vermelho é uma planta daninha de sucesso porque possui uma série de características que contribuem na eficiência da infestação de lavouras e na dificuldade de controle. Dentre elas destaca-se a adaptação a práticas agronômicas, ciclo de desenvolvimento sincronizado com o da cultura e a emergência rápida e vigorosa (Delouche *et al.,* 2007). Além disso essa planta daninha, pode ser facilmente dispersada através de contaminação das sementes de arroz cultivado, apresenta alto nível de degrane o que inviabiliza a retirada destas da lavoura e apresenta intensa e prolongada dormência das sementes, que mantém a viabilidade das sementes por longos períodos (Delouche *et al.,* 2007). Desta forma, o estudo destas características podem auxiliar no estudo e desenvolvimento de ferramentas para o controle do arroz vermelho.

2.4 O caractere degrene

O degrane ou debulha natural é um caráter evolutivo e adaptativo para a dispersão e distribuição de sementes em varias espécies, entre elas especies silvestres do arroz (Li *et al.,* 2006b; Lin *et al.,* 2007). No entanto, estes caracteres podem causar perdas consideráveis no rendimento de grãos no arroz domesticado. Através do processo de

domesticação têm sido selecionados biótipos com baixos níveis de degrane (Gu *et al.*,2005). Sendo que atualmente, o arroz cultivado apresenta grau de debulha considerado desejável, podendo variar conforme o cultivar e a forma de colheita.

O caráter degrane contribui para a dispersão e distribuição das sementes de arroz vermelho através de diversas formas. Primeiramente, o degrane permite que uma parte das sementes produzidas seja distribuída sobre a superfície do solo antes e durante a colheita, evitando que seja colhida com a cultura e removida do sistema de produção. Em segundo lugar, o principal fluxo de queda das sementes, na maioria dos tipos de arroz vermelho, ocorre alguns dias antes ou no momento da maturação fisiológica das sementes (Delouche *et al.*, 2007).

Normalmente a zona de abscisão entre o grão do arroz e o pedicelo é formada por uma camada de pequenas células com a parede celular fina. Nas plantas que apresentam degrane como as plantas silvestres, essa camada de células é contínua em toda a zona de abscisão. Já nas plantas que possuem pouca debulha natural essa camada é descontínua e completamente ausente na região dos feixes vasculares (Li *et al.*, 2006). A debulha natural do grão de arroz é causada pela diferenciação da camada de abscisão que delimita o grão do pedicelo. O processo de abscisão é gerado pela produção de etileno, que inibe a produção de auxina. Respondendo a certos sinais, enzimas hidrolíticas, como polygalacturonase e β-endo-glucanase, são ativadas nas células da camada de abscisão, causando a degradação da lamela média e da parede celular resultando na queda do grão (Roberts *et al.*, 2000; Patterson, 2001; Roberts *et al.*, 2002).

Os ecótipos de arroz vermelho tendem a apresentar pouca variabilidade quanto à intensidade do degrane, uma vez que as sementes de ecótipos que apresentam baixo degrane acabam sendo eliminadas da lavoura juntamente com grãos da cultura (Delouche *et al.*, 2007). Do mesmo modo, ecótipos com elevado degrane fazem com que sua erradicação seja dificultada, pois este caráter consiste em um dos principais meios de disseminação das sementes, causando reinfestação a partir do banco de sementes.

Estudos genômicos sobre a debulha natural em cruzamentos de *O. sativa spp. indica*com*O. rufipogon* (espécie silvestre e perene) têm demonstrado que o caráter é controlado por quatro (Cai e Morishima, 2000) ou cinco (Xiong *et al.*, 1999; Konishi *et al.*, 2006) QTLs (*quantitative trait loci*) principais. Por outro lado, análises genéticas de uma população F_2 proveniente do cruzamento entre *O. indica*e*O. nivara* (espécie silvestre e anual) indicou a existência de três QTLs (*sh3*, *sh4* e *sh8*) responsáveis pela redução da debulha natural (Li *et al.*, 2006a). Neste estudo, foi verificado que o QTL *sh4* localizado

no cromossomo 4 foi dominante e explicou 69% da variância fenotípica. Enquanto que os QTLs *sh3* e *sh8* explicavam 6,0 e 3,1%, respectivamente.

Da mesma forma, estudos que cruzaram*O. japonica* com três espécies silvestres (*O. rufipogon*, *O. glumaepetula* e *O. meridionalis*) verificaram um gene/QTLdominante e de grande efeito que também está presente no cromossomo 4 das três espécies era responsável pelo degrane (Sobrizal *et al.,* 1999; Nagai *et al.,* 2002), que provavelmente também pode ser o QTL *sh4*. Entretanto, analisando geneticamente uma população F_2 oriunda do cruzamento entre *O. indica* e *O. japonica* foram identificados 5 QTLs, sendo que o alelo *qSH1* presente no cromossomo 1 explicou 69% da variância fenotípica (Konishi *et al.,* 2006).

Analisando o gene *sh4*, Li*et al*. (2006) verificaram que uma única mutação não sinônima (substituição do nucleotídeo G por T, resultando na substituição do aminoácido asparagina por lisina) em uma região de 1,7 kb, presente na posição 237 do exon 1 do gene. Além disso, Segundo Li *et al.* (2006) e Thurber (2012) essa alteração é responsável pelo desenvolvimento incompleto da camada de abscisão e origem da ausência de debulha no arroz cultivado. Entretanto, Zhu *et al.* (2012), avaliaram acessos e cultivares oriundos de diferentes regiões orizícolas, sendo 166 acessos de arroz silvestre, 222 acessos de arroz daninho e 192 cultivares de arroz e verificaram que 73,5% dos acessos de arroz silvestre continham o nucleotídeo G na posição 237 do éxon 1, enquanto que 26,5% dos acessos silvestres apresentavam a mutação $G_{237}T$ no gene *sh4*. Além disso, verificou-se que todos os acessos de arroz daninho possuíam o nucleotídeo T, uma vez que todos esses acessos apresentaram degrane.

Ainda com relação ao gene *sh4*, Thurber *et al.* (2010) verificaram que a presença desta única mutação não é suficiente para conferir redução nadebulha natural, pois o único genótipo que não apresentou degrane dentre os acessos analisados, a espécie *O. rufipogon*, não possuia amutação $G_{237}T$ no gene *sh4*. Também, diferentemente do esperado, os biótipos de arroz vermelho avaliados que apresentam naturalmente alta propensão ao degrane, apresentaram a mutação T no gene *sh4* (Thurber et al 2010). Isto evidencia que a mutação T em *sh4* foi corrigida primeiramente em um conjunto de cultivares, e espalhou-se rapidamente para grupos de arroz domesticados através de fluxo gênico e seleção (Zhang *et al.,* 2009).

Nunes (2012) desenvolveu estudos no RS com duas cultivares de arroz cultivado (Batatais e Lacassine) e dois genótipos de arroz vermelho (AV31 e AV60), enão verificou relação direta da expressão do gene *sh4* com a ocorrência do degrane. Da mesma forma,

Markus (2013) também analisou a possível existência da mutação $G_{237}T$ no gene *sh4*, nos mesmos cultivares, no genótipo de arroz vermelho AV53 e na espécie silvestre *O. glaberrima* e os resultados mostraram ausência da mutação $G_{237}T$ no referido gene. Isso sugere que a mutação $G_{237}T$ está de alguma forma associada com o caractere degrane porém não trata-se de um variação genômica funcional que ativa a expressão do gene e que a presença de tal mutação também depende da população estudada.

Já o gene *OsCPL1* por sua vez, age como repressor da diferenciação da camada de abscisão (Ji et al 2010). O locus recessivo localiza-se entre marcadores RM7161 e RM8262 no cromossomo 7 (Ji *et al*., 2006). Alguns estudos identificaram que um SNP (G para T) localizado no éxon 8desse gene muda o aminoácido conservado serina para isoleucina, fazendo com que o fenótipo apresente degrane (Ji *et al*., 2010). Também foi observado que quanto maior a expressão do gene *OsCPL1*, menor é o nível de degrane e que as linhagens transgênicas com o gene *OsCPL1* inativado por RNA de interferência apresentaram níveis elevados de degrane (Ji *et al*., 2010). Contudo, outro estudo que avaliou a expressão do mesmo gene em dois ecótipos de arroz vermelho e duas cultivares revelou resultados contraditórios (Nunes, 2014). Neste estudo também se verificou que nos ecótipos que possuem alto degrane, a expressão do gene foi superior em relação às cultivares que possuem menor nível de degrane. Dessa forma, a expressão do gene *OsCPL1* estaria relacionada com a ativação do processo de abscisão, pois a expressão do gene estaria relacionada com a presença do degrane e não com a repressão do degrane como o observado nas análises histoquímicas do estudo de Ji *et al*. (2010).

2.5 Sequenciamento de nova geração e montagem de genomas

Na década de 1970, duas pesquisas estabeleceram um marco na ciência que ditaria o rumo a ser seguido nos próximos anos na área da biologia. Tais estudos possibilitaram o desenvolvimento de métodos de obtenção da sequência de fragmentos de DNA em laboratório, com duas técnicas diferentes, a enzimática dideoxi de Sanger (Sanger et al 1977) e por degradação química de Maxam e Gilbert (Maxan e Gilbert, 1977).

Estes primeiros esforços foram trazendo resultados de forma lenta, uma vez que o sequenciamento era manual e menos de duas centenas de bases eram produzidas num conjunto de quatro canaletas. O surgimento do sequenciador automático utilizando a técnica de Sanger na década de 1980 (Connell *et al.*, 1987), permitiu que avanços mais significativos pudessem ser feitos em grande escala.

A técnica Sanger foi o método de sequenciamento de ácidos nucléicos mais

utilizado nas décadas seguintes (Hutchison, 2007), uma vez que o concomitante desenvolvimento de algorítimos de montagem de genomas a partir de fragmentos sequenciados ao acaso tornaria o método Sanger a principal ferramenta do projeto de sequenciamento do genoma humano (Lander *et al.*, 2001; Venter *et al.*, 2001) e também foi utilizada no sequenciamento dos genomas das plantas Arabidopsis (Arabidopsis Genome Initiative, 2000) e arroz (IRGSP, 2005).

Apesar da capacidade inicial de produzir poucos milhares pares de bases por ano, a automatização do processo e aumento no número de capilares nas últimas décadas permitiu um incremento significativo na agilidade, possibilitando a elucidação de centenas de genomas de diversos organismos. Na última década, plataformas de sequenciamento que utilizam a tecnologia de nova geração ("Next-Generation Sequencing") foram disponibilizadas no mercado.

Essas técnicas inovadoras são baseadas em nanotecnologia e na construção de bibliotecas de fragmentos ou pareadas de DNA que não dependem da clonagem em vetores, abrindo grandes horizontes para estudos genéticos (Shendure; Ji, 2008). Os principais atributos desses sequenciadores são o custo, expressivamente inferior, e a rapidez com que os dados são gerados. Esses possuem a capacidade de processar milhões de sequências em uma única corrida, enquanto os sequenciadores convencionais de capilares processam apenas 96 ou 384 sequências simultaneamente. Além disso, uma quantidade pequena de DNA é requerida para a construção de bibliotecas (Mardis, 2008; Shendure; Ji, 2008).

As plataformas de sequenciamento de nova geração ou segunda geração comercialmente disponíveis incluem o sistema Roche 454 (Roche Applied Science), GenomeAnalyser IIx (Illumina, Inc.), HiSeq (Illumina, Inc.) e SOLiD (Applied Biosystems) (Tabela 2).

TABELA 2. Tamanho de leituras e rendimento das plataformas de sequenciamento Sanger e de nova geração de sequenciamento (NGS). Porto Alegre, 2015.

Ano de lançamento	Plataformas	Tamanho máximo das leituras	Rendimento *per run*
1977	Sanger	1000 bp	100 Kb
2005	454 (Life Science/Roche Diagnostics)	500 bp	500 Mb
2005	ABI Solid (Life Technologies)	50 bp	100 Gb
2007	Illumina Genoma Analyzer (Solexa)	150 bp	300 Gb
2010	Helicos (Helicos Biosciences)	55 bp	35 Gb
2010	Ion Torrent (Life Technologies)	200 bp	1 Gb
2010	SMRT (Pacific Biosystms)	2000 bp	100 Mb
2011	PacBio	5000 bp	1 Gb
2012	Nanopore MinION	1000 bp	1 Gb

O modelo 454 FLX da Roche foi o primeiro sequenciador de nova geração, introduzido no mercado no ano de 2004. Essa plataforma utiliza o princípio do pirosequenciamento, ou seja, a liberação de uma molécula de pirofosfato enquanto a DNA polimerase incorpora os nucleotídeos (sequenciamento por síntese de DNA). Essa reação produz a quebra da oxiluciferina pela luciferase, produzindo radiação luminosa em um comprimento de onda específico (Margulies*et al.*, 2005). A imagem emitida pela luciferase é então gravada enquanto um determinado nucleotídeo é adicionado. Atualmente esta plataforma possui como principal vantagem o tamanho da leitura gerada (tamanho médio de 400 pb) o que facilita o processo de montagem genômica, porém possui como principal limitação elevada taxa de erro (0,38%) em regiões de homopolímeros (Loman*et al.*, 2012).

Outro modelo de sequenciador de nova geração é a plataforma Illumina. Essa plataforma foi lançada no final de 2006, com a capacidade de gerar cerca de 100 milhões de segmentos de leitura por corrida. Essa tecnologia é baseada no princípio químico do sequenciamento por síntese, empregando quatro tipos de nucleotídeos proprietários, dotados da capacidade de terminação reversível e marcados com diferentes fluoróforos e, também, uma enzima DNA polimerase especialmente habilitada para incorporá-los (Figura 2) (Illumina, Inc., 2007; Ansorge, 2009). Inicialmente, o procedimento estava limitado a produzir sequências com um comprimento de apenas 36 bases. Contudo, foram surgindo novos equipamentos como o Genome Analyzer (Illumina GA) da geração de

2011, uma das tecnologias mais recentemente, que emprega tecnologia SBS - Sequencing By Synthesis, capaz de gerar até 600 Gb com segmentos de 76 pares de base em média. Equipamentos ainda mais recentes da tecnologia permitem leitura de segmentos de DNA com 100 bp, em média (www.illumina.com). A principal vantagem desta tecnologia é a elevada qualidade e cobertura dos dados gerados (Metzker, 2010) aliado ao baixo custo, fazendo com que essa plataforma torna-se uma das mais usadas atualmente.

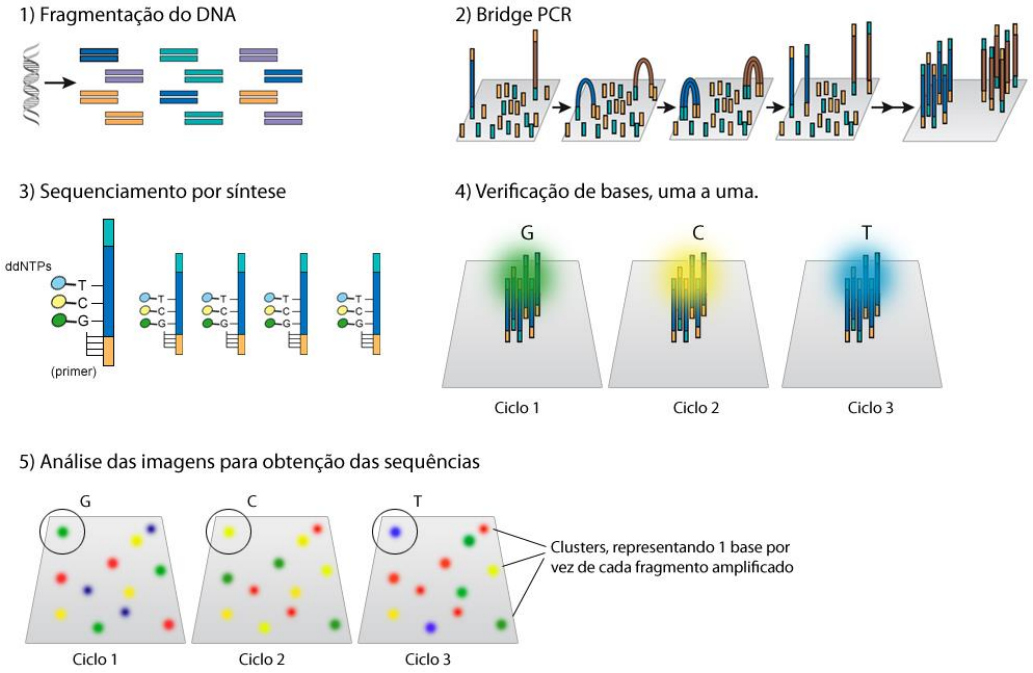


FIGURA 2. Ilustração do método de sequenciamento Illumina. Fonte:Tieppo (2014). Porto Alegre, 2015.

A plataforma SOLiD ("*Sequencing by Oligo Ligation and Detection*") da Applied Biosystems foi lançada no mercado em outubro de 2007. Essa plataforma utiliza a incorporação de dinucleotídeos marcados por meio da DNA ligase, seguida pela excitação do fluoróforo (o sinal emitido é captado por sensores) e a incorporação dos dinucleotídeos seguintes. Essa leitura gera um código de cores que é analisado por ferramentas de bioinformática e convertida à sequência de letras. Cada corrida no sequenciamento da plataforma SOLiD leva aproximadamente 5 dias, e produz de 3 a 4 Gb de sequências com o comprimento variando de 50 a 75 pb. O sequenciamento baseado na ligação de oligonucleotídeos ocorre por meio do anelamento de um primer universal do SOLiD, seguido pela ligação de uma sonda marcada que é detectada pela máquina em cada ciclo. O mecanismo de detecção de dinucleotídeo garante uma correção de erro, melhorando a

qualidade dos dados (Mardis, 2008; Metzker, 2010).

A plataforma de sequeciamento mais rescente ainda está em fase de testes mas promete ser ainda mais eficaz e barata, Nanopore MinION é o nome dessa plataforma produzida e comercializada pela Oxford Nanopore. Embora tenha sido lançado recentemente, esse método de sequenciamento foi inventado e patenteado em 1995 (Kasianowicz *et al*., 1996, Church *et al*., 1998). A teoria por trás desse mecanismo é que, quando um nanoporo é imerso num fluído condutor e uma corrente eléctrica é aplicada através dele possibilta identifcar bases nucleotídicas passando através desse poro. Além disso, o sequenciamento é feito diretamente sem a necessidade de amplificação previa do DNA via PCR. Essa plataforma é considerada como a terceira ou quarta geração de sequenciamento por alguns autores e têm o potencial para sequenciar de forma rápida e confiável todo o genoma humano por menos de $1000, e possivelmente até mesmo para menos de $100 nos próximos anos (Feng *et al.,* 2015). Não apenas a geração de dados terá seu custo drasticamente reduzido, mas também o instrumento em si será relativamente barato, e o tempo necessário para a cobertura de 6 vezes de um genoma humano pode ser inferior a um dia (Branton *et al.,* 2008). Contudo essa tecnologia continua em fase de testes, principalmente devido a elevada taxa de erro nos dados gerados.

Todas essas plataformas de sequenciamento estão promovendo grandes projetos de sequenciamento de plantas, como o IOMAP - International Oryza Map Genome Initiative, buscando sequenciar todas as espécies do gênero Oryza, o "1000 Plant Genomes Project" (www.onekp.com/), o "1001 Arabidopsis Genome Project" (www.1001genomes.org/) e o "1000 Plant and Animal Genome Project" (www.1d1.genomics.cn/). Igualmente, o "Genome 10K Project" foi criado para sequenciar e montar 10.000 genomas de vertebrados, incluindo pelo menos um de cada gênero (www.genome10k.org/).

Com o avanço nas tecnologias de pesquisa genômica, houve um crescimento nas informações biológicas disponíveis em bancos de dados, como o tamanho dos genomas de várias espécies de importância e o número de genes presentes (Tabela 3).O GeneBank e o RapDB são exemplos disto.

TABELA 3. Tamanho dos genomas e número aproximado de genes das principais de espécies cultivadas. Porto Alegre, 2015.

Espécies	Tamanho do genoma (Mb)	Número de genes
Cucumis sativus L.	367	26682
Prunnus persica	230	27852
Fragaria ananassa	720	34809
Brassica rapa	550	41174
Citrus sinensis	380	25,066
Theobroma cacao	346	28798
Beta vulgaris	758	27421
Carica papaya	372	13311
Vitis vinifera	487	30484
Musa acuminata	523	36542
Castor bean	320	31237
Manihot esculenta	760	30666
Eucalypytus grandis	900	34724
Solanum tuberosum	856	39031
Malus domestica	743	57386
Solanum lycopersicum	950	34727
Sorghum bicolor	730	34496
Oryza rufipogon	406	37071
Oryza sativa	430	41,620
Glycine max	1115	46430
Phaseolus vulgaris	520	31648
Zea mays	2300	39656
Triticum aestivum	17000	124201 (estimado)

Além de possibilitar a visualização de toda a coleção de genes de uma espécie, esse desenvolvimento forneceu ferramentas eficientes de genética reversa e de transcriptômica para o estudo de funções gênicas. Hoje em dia diversas outras espécies vegetais foram sequenciadas, entre elas o arroz (*Oryza sativa*, IRGSP, 2005), a uva (*Vitis vinifera*, TFIPCGGC, 2007), a soja (*Glycine max*, Schmutz *et al*., 2010) e o milho (*Zea mays*, Schnable *et al*., 2010). Não há dúvida de que a disponibilidade de sequências genômicas das principais espécies agrícolas vai possibilitar um conhecimento muito maior de sua biologia, impactando significativamente o desenvolvimento de novas variedades adaptadas a diferentes condições de ambientes.

Entretanto, essas novas plataformas de sequenciamento geram leituras pequenas (de 35 a 800 bases), que são menores que as geradas pelas tradicionais sequências da tecnologia Sanger, e com isso, à etapa de montagem dessas leituras em contigs, torna-se um dos maiores desafios na aplicação dessas novas tecnologias de sequenciamento (Figura 3). Isso se deve ao tamanho da leitura, pois nas plataformas NGS, é sacrificado em prol de uma vazão maior, permitindo que elas geram uma grande cobertura de sequenciamento (de 30 vezes ou mais) a um baixo custo relativo, se comparada com a

cobertura típica obtida com as plataformas de eletroforese capilar (da ordem de 10 vezes). Assim, com relação a esta última, conseguem promover um sequenciamento muito mais rápido, produtivo e bem menos oneroso (Chaisson *et al.*, 2004).

Assim, a bioinformática tem um papel muito importante nessas pesquisas, pois com o aumento na quantidade de informação de sequências através das modernas técnicas de sequenciamento em larga escala, se faz necessário o desenvolvimento de ferramentas computacionais e algoritmos mais eficientes para a análise dessa imensa quantidade de dados, maximizando-se o potencial de avanço nas áreas de biologia, genética e no melhoramento genético, como por exemplo, da construção da sequência completa de todos os cromossomos de um organismo, ou seja, desempenhando um papel fundamental, auxiliando na transformação da informação genética em conhecimento biológico aplicável (Martins, 2013).

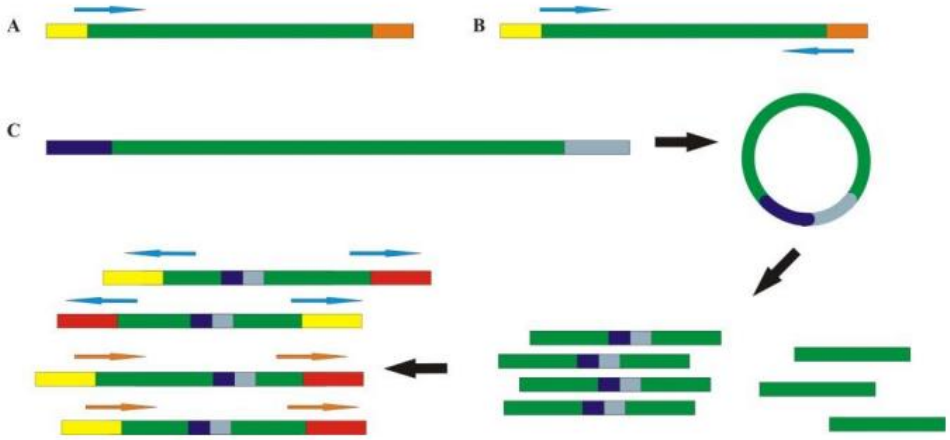


FIGURA 3. Ilustração das leituras geradas pelos sequenciamentos de nova geração. (A) Leituras simples, chamada de Single-end, a leitura é feita em apenas um lado (indicado pela seta em azul). (B) Nas leituras pareadas, paired-end, o sequenciamento é realizado em ambos os lados do fragmento, em sentidos contrários indicado pelas setas em azul). (C) Nas leituras mate-pair, são ligados nucleotídeos com biotina em ambos os lados dos fragmentos, e em seguida o fragmento é circularizado e cortado em sequências menores que são selecionadas com base na biotina para seguir com a leitura das sequências (Simples pela plataforma 454 e pareada pelas Illumina e SOLID. (Hamilton& Buell, 2012). Porto Alegre, 2015.

Atualmente, para a montagem de genomas existem várias opções de programas montadores que utilizam diferentes tamanhos de fragmentos, inúmeros formatos de arquivos, e são aplicados a genomas de diferentes complexidades. Novos avanços no

processo de montagem são esperados com a integração e bancos de dados e com a exploração de múltiplas estratégias de sequenciamento, sempre como propósito de enfrentar o desafio de montagem de genomas complexos.

Os programas podem ser classificados em três categorias que diferem conforme o algoritmo que é utilizado, todos são dinâmicos e baseados em grafos: Greedy; Overlap-Layout-Consensus; e Grafo de Bruijn. Greedy ou Guloso foi o primeiro algoritmo utilizado em bioinformática para realizar a montagem de genomas com leituras geradas pelas plataformas de sequenciamento de nova geração.

A construção dos *contigs* é feita da seguinte forma: primeiramente, esse algoritmo realiza uma busca por sobreposição par a par, entre todas as leituras e então, a sobreposição de maior pontuação encontrada é mantida e utilizada para uma nova busca por sobreposição, essa operação é repetida até que não seja possível a realização de mais nenhuma sobreposição. O chamado *Overlap-Layout-Consensus* (OLC) é amplamente utilizado para dados da tecnologia Sanger, e essa estratégia é realizada através da construção de um grafo de sobreposições, onde as leituras são consideradas vértices e as sobreposições como arestas, esse algorítmo é baseado em três etapas, primeiramente é computada a sobreposição entre todas as leituras, através do alinhamento par a par, e em seguida, é feita a ordenação e orientação das leituras de acordo com as sobreposições encontradas na fase anterior, e por último, através do alinhamento múltiplo das leituras (com base nos dados anteriores) a construção dos *contigs* é realizada (Farias, 2013).

Os algorítmos baseados em grafos de Bruijn, não realizam a comparação par a par entre todas as leituras, que são decompostas em sequências menores, com tamanho pré-definido (chamadas de K-mers), em seguida, são encontradas as sobreposições entre as K-mers, e desta forma, um grafo de Bruijn é formado, onde os vértices são compostos por uma série de sobreposições entre os K-mers, e se o sulfixo de um vértice pode ser conectado (através de sobreposição) ao prefixo de outro, então uma aresta é formada, e assim as sequências contíguas, e um caminho Euleriano desse grafo é formado (Miller *et al.*, 2010; Farias, 2013).

Portanto, de forma resumida, a montagem de genomas pode ser definida como a reconstrução da sequência de um genoma, a partir das várias leituras obtidas na etapa de sequenciamento. A maior dificuldade de realizar a montagem das leituras produzidas pelas plataformas de sequenciamento, em sequências contíguas, é devido ao tamanho dessas leituras e pela presença de regiões repetitivas nos genomas (Schatz *et al.*, 2010).

Grandes quantidades dessas regiões estão presentes em genomas vegetais,

chegando a alguns casos até 97% do total do conteúdo de DNA nuclear (Flavell*et al*.,1974; Murray*et al*., 1981). Em plantas, a maioria dessas regiões genômicas é composta por elementos transponíveis (ETs), principalmente por retrotransposons LTR (*Long Terminal Repet*) (Bennetzen, 2002; Ergman; Quesneville, 2007) que constitui uma grande porcentagem do total do tamanho do genoma, como em milho (58%, Messing*et al*., 2004), mamão (52%, Nagarajan *et al*., 2008), arroz (35%, IRGSP,2005) ou *Arabidopsis thaliana* (14%, Arabidopsis Genome Initiative, 2000).

Ainda com relação à montagem de genomas, do ponto de vista relacionado à análise dos dados, vale também ressaltar que ela pode ser de dois tipos: o primeiro, às vezes designado como ressequenciamento, utiliza um genoma de referência, contra o qual as leituras são alinhadas por similaridade. Normalmente, esse genoma de referência é escolhido levando-se em consideração a sua proximidade filogenética em relação ao genoma sequenciado e, em projetos deste tipo, a cobertura de sequenciamento necessária é menor (da ordem de 8 a 12 vezes) (Schuster, 2008).

Por causa do processo de alinhamento das leituras, tal abordagem de trabalho é, frequentemente, designada como mapeamento (Shendure & Ji, 2008; Horner *et al*., 2009; Bao *et al*., 2011) ou alinhamento (Paszkiewicz& Studholme, 2012) simplesmente. O segundo tipo é conhecido como montagem de novo ou ab initio ou sequenciamento de genomas desconhecidos. Nele, a montagem é executada usando-se as próprias leituras, ou seja, não há um genoma de referência para auxiliar o processo. Nesse caso, a cobertura de sequenciamento ideal é maior (da ordem de 25 a 70 vezes) (Schuster, 2008). Há situações, também, em que ambas as abordagens podem ser utilizadas em um mesmo projeto (Pop, 2009; Paszkiewicz; Studholme, 2012).

Dessa forma, a união de novas técnicas de sequenciamento com o estudo de caracteres adaptativos em arroz vermelho poderá se traduzir em um melhor entendimento da genética destes caracteres, facilitando o desenvolvimento de estratégias mais eficientes de controle destas plantas em condições de campo. Neste sentido, o caráter degrane é emblemático, uma vez que está diretamente relacionado com a manutenção de plantas de arroz vermelho a campo, sendo um fator adaptativo de grande importância. Além disso, futuramente, o sequenciamento de variedades de arroz cultivadas no sul do Brasil possibilitará o estabelecimento de um banco de dados nacional. Este banco de dados poderá ser pesquisado para genes de interesse dos melhoristas, permitindo o isolamento e clonagem de genes diretamente a partir da análise *in silico*.

3 MATERIAL E MÉTODOS

Este trabalho sequenciou e montou o genoma de dois genótipos de arroz vermelho com ocorrência em lavouras de arroz cultivado do estado do Rio Grande do Sul. Esses genótipos foram escolhidos a partir de trabalhos anteriores, que avaliaram o degrane 36 genótipos, sendo 18 cultivares de arroz, 16 ecótipos de arroz vermelho e duas espécies silvestres, em duas safras (2008/2009 e 2010/2011) (Nunes, 2012). Para melhor representação selecionou-se para o presente trabalho um genótipo com elevado grau de degrane (AV60) e outro com nível intermediário (AV53).

A condução dos trabalhos se deu no Laboratório de Biologia Molecular do Departamento de Plantas de Lavoura da Universidade Federal do Rio Grande do Sul (UFRGS) em Porto Alegre/RS e no Laboratoire Génome et Developpement de Plantes-Université de Perpignan via Domitia (UPVD), situado na cidade de Perpignan, França.

3.1 Material vegetal e extração de DNA

Os diferentes genótipos analisados foram semeados em casa de vegetação no ano de 2013. O DNA genômico foi extraído a partir de amostras de folhas jovens de plantas individuais através do protocolo CTAB (brometo de cetiltrimetilamônio) modificado (Doyle & Doyle, 1987).

O material vegetal foi macerado na presença de nitrogênio líquido (LN2) e acondicionado em microtubos de 1,5 mL previamente resfriados. Quinhentos µL de tampão de extração (0,1 M Tris-HCl [pH 8,0], 0,02 M EDTA [pH 8,0], 1,4 M NaCl, 2% CTAB) foram adicionados a cada tubo, seguido por agitação. Após, os tubos foram incubados a 65°C por 30 min em banho-maria, com agitação dos tubos a cada 10 min. Em seguida, 500 µL da mistura clorofórmio: álcool isoamílico (24:1) foi adicionada a cada tubo, misturado por 5 minutos e centrifugado a 10.000 rpm por 5 min em temperatura ambiente. O sobrenadante foi transferido para um novo microtubo 75 µL de RNAse (100 mg mL-1) foram adicionados e incubados a 37ºC durante 60 min. Após, o DNA foi precipitado com 300 µL de álcool isopropanol gelado, agitado gentilmente e

acondicionado por 12 h a 4ºC.

No dia seguinte as amostras foram submetidas a 10 min em temperatura ambiente e posteriormente foram centrifugadas por 30 min a 14.000 rpm e, o sobrenadante descartado. O precipitado foi lavado com etanol 70%, centrifugado por 5 min a 10.000 rpm e novamente lavado com etanol 70% e centrifugado por 5 min a 10.000 rpm. Após o descarte do etanol o precipitado foi seco por 20 a 30 min e ressuspendido em 30 µLde TE 0,1X, o DNA foi quantificado utilizando-se o espectofotômetro Genesys 2® (Thermo Spectronic) e gel de agarose 1%, após o DNA foi diluído em 50 ng μL^{-1} e enviado para o Laboratório Montpellier Genomix para o sequenciamento.

3.2 Sequenciamento

Para o sequenciamento dos genótipos utilizou-se a tecnologia Illumina por meio da plataforma Genome Analyzer Ilx, obtendo-se uma biblioteca pair-end para cada genótipo.O processo de preparação das amostras é padrão e realizado por meio de kits do equipamento elaborados pela empresa detentora da tecnologia (Illumina).

Após o recebimento da amostra de DNA, foram preparadas bibliotecas de tamanhos específicos (o sistema de kits da Illumina para preparação de bibliotecas de DNA pode gerar fragmentos paired-end (< 1 kb) com reads de até 2 x 100 pb), as quais foram submetidas à fragmentação mecânica ligando-se adaptadores aos terminais destes fragmentos, os quais desempenham o "papel" de discernir as amostras, auxiliando a análise bioinformática após a corrida.

Após, as bibliotecas foram depositadas em uma lâmina (flowcell) contendo 12 canaletas (lanes) através de um instrumento robótico chamado cBot (Illumina). Na flowcell há uma superfície de oligos que se complementam aos adaptadores das bibliotecas, onde ocorreu a etapa de amplificação dos fragmentos. A flowcell foi então colocada no equipamento, onde ocorreu a incorporação de nucleotídeos marcados por fluorescência contendo os terminadores dideoxi, nos fragmentos ligados aos primers de sequenciamento. Ao ocorrer a incorporação de um nucleotídeo, a fluorescência é excitada com uma série de lasers e captada por câmeras. Após a captura da imagem, os terminadores foram clivados e o próximo ciclo de incorporação ocorreu (até 100 ciclos por direção do read). Ao final da corrida, prossegue-se com a análise bioinformática.

3.3 Montagem e alinhamento dos genomas

Após a obtenção das bibliotecas com leituras pareadas (pair-end), os dados foram disponibilizados em formato *fastaq*, padrão da tecnologia Illumina.Para verificar a qualidade das leituras obtidas e remoção de possíveis adaptadores remanescentes foram utilizados os programas: *FASTX-toolkit*, utilizado para identificar possíveis adaptadores presentes nas leituras; *FastQC* que foi utilizado para a visualização da qualidade das bases. Após essa verificação, utilizou-se o script Trim.sliding.Window.pl para a realizar a limpeza das leituras, através da retirada de bases com baixa qualidade.

Com isso, foi obtido a cobertura total do genoma pelo sequenciamento, que foi calculada de acordo com a seguinte fórmula:

$$Cobertura\ teórica = \frac{Tamanho\ dos\ reads \times Número\ de\ clusters\ após\ limpeza}{Tamanho\ do\ Genoma}$$

Onde: tamanho dos reads = tamanho das leituras realizadas pelo sequenciador, número de clusters após a limpeza = número de clusters encontrados após a análise de qualidade das bases geradas e limpeza pelo programa Trimm e o tamanho do genoma = tamanho do genoma de referência de Oryza sativa, 430 megabases (IRGSP, 2005).

Para o processo de montagem foi utilizado uma série de programas montadores de genomas (*assemblers*), com o objetivo de buscar o melhor montador para o genoma do arroz vermelho. Assim, utilizou-se os programas ABySS (Simpson*et al*., 2009); RAY (Boisvert et al. 2010); SOAPdenovo (Li et al. 2010); Velvet (Zerbino; Birney, 2008); e SSPACE (Boetzer*et al*., 2011), que são empregados na montagem de genomas sequenciados por meio da tecnologia Illumina.

Depois de montar o genoma, buscou-se alinhar as leituras obtidas dos genótipos de arroz vermelho (AV53 e AV60) com as sequências do genoma de referência de *Oryza sativa japônica* (IRGSP 1.0), utilizando-se o programa Bowtie. Da mesma maneira, o programa BLAST (Altschul *et al*., 1990) foi usado para o alinhamento dos *scaffolds* dos dois genótipos de arroz vermelho com as sequências do genoma referência de *Oryza sativa indica*.

3.4 Análise estrutural de genes relacionados ao degrane

Sequências individuais em formato *fasta* do *locus* e da *cds* dos principais genes relacionados ao degrane, previamente analisados por Nunes (2012), foram obtidas do

banco de dados público RAPDB. Em seguida, realizou-se análises *blastn* e *tblastn* contra os genomas de arroz vermelho sequenciados nesse estudo e também contra os genomas de referência *O. indica* e *O. japonica*. Após recuperar as sequências similares, com auxílio do programa Gepard, construiu-se um dotplot para comparar a estrutura dessas sequências.

Em seguida foi realizado um alinhamento das sequências nucleotídicas usando o algoritimo ClustalW (Higgins *et al*., 1994) no progrmana MEGA7 (Kumar *et al*., 2015). Posteriomente as sequências foram traduzidas para proteína e avaliadas em relação a sua distância evolutiva por meio do método do vizinho mais próximo.

3.5 Detecção de SNPs e INDELs

Após o alinhamento das leituras obtidas dos genótipos de arroz vermelho (AV53 e AV60) com as sequências do genoma de referência de *Oryza sativa japônica* e *O.sativa indica* (IRGSP 1.0 variedade 9311), por meio do programa Bowtie, realizou-se uma busca pelas variações nucleotídicas (SNPs e INDELs) utilizando-se a estratégia Samtools, com a opção *mpileup* e filtragem *bctools*. A frequência dos variantes genômicos foi determinada através da divisão do número total de pares de bases após limpeza pelo número de SNPs e INDELs detectados na análise descrita anteriormente.

4 RESULTADOS E DISCUSSÃO

4.1 Montagem e comparação dos genomas

Inicialmente foi analisado sequências brutas (fastq) e os seus correspondentes valores de qualidade atribuídos pelo Illumina. O software calcula um índice de qualidade para cada base nucleotídica refletindo a probabilidade dessa base ter sido sequenciada errada. Esse cálculo leva em conta a ambiguidade do sinal para a respectiva base, bem como a qualidade da bases vizinhas e a qualidade de toda a leitura. Sendo assim, o escore de qualidade Q é definida por $Q = -10\ log_{10}\ (P)$ (Minoche *et al*., 2011). Por exemplo, Q=30 corresponde à probabilidade *P*=0,001 que uma base foi chamada incorretamente. O maior valor possível para Q atribuído pela software é de 40 (Figuras 4 e 5), correspondendo a *P* = 0,0001. A qualidade das bases foi adequada nos ciclos de 1-90 das duas leituras, tendo uma ligeira diminuição no final da leitura para os genótipos AV53 e AV60 e suas duas respectivas amostras (Figura 4). Entretanto, pode-se afirmar que a qualidade das amostras foi ideal.

Quanto maior a cobertura do sequenciamento, mais informação se obtém para cada posição, facilitando o processo de montagem do genoma e tornando-o mais robusto. Neste trabalho esperava-se uma cobertura teórica de 25 vezes, mas utilizando-se a fórmula para cobertura do sequenciamento, citada acima, onde: o tamanho dos reads foi igual a 2*100 para um sequenciamento paired-end e com o genoma do arroz com cerca de 430 megabases, a cobertura final foi de 36,6 vezes para o genótipo AV53 e 32,1 vezes para o genótipo AV60, indicando que a cobertura de sequenciamento foi satisfátoria. Viera *et al*. (2014), encontraram resultados semelhantes para algumas espécies de coníferas, onde obteve uma cobertura média do genoma de: 24,63 para *A. angustifolia*, 135,97 para *A. bidwilli*,1196,10 para *P. lambertii* e 64,68 para *P. patula*, ressaltando também que esse é um importante fator facilitador no momento da montagem do genoma devido a elevada cobertura do genoma.

A qualidade das bases sequenciadas tende a diminuir ao final das leituras, conforme ilustrado na Figura 4 para as amostras 1 e 2 dos genótipos AV53 e AV60. Sendo

que as áreas em verde, amarelo e vermelho nos gráficos indicam respectivamente, elevada (≥28), intermediária (20 a 28) e baixa qualidade (<20) das bases sequenciadas.

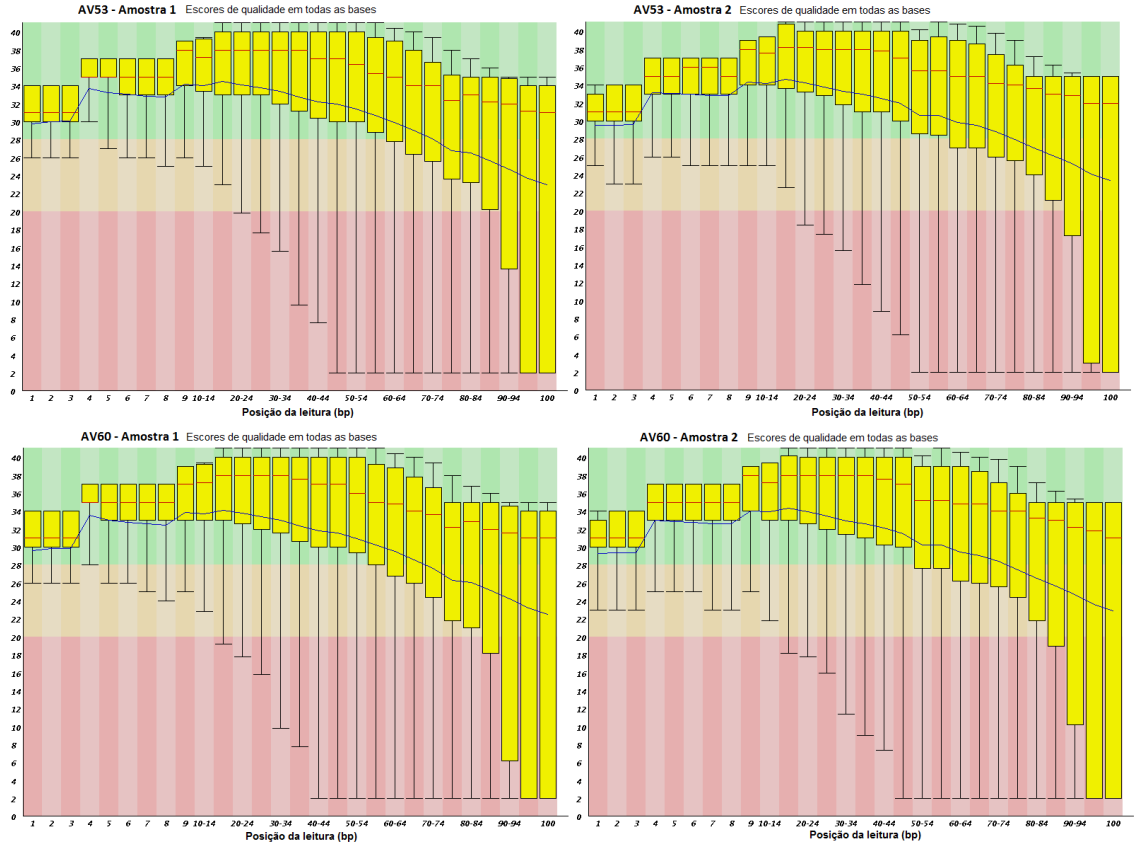


FIGURA 4. Boxplot mostrando a distribuição da qualidade das bases nas leituras para as amostras provenientes do sequenciamento dos genótipos AV53 e AV60 antes do refinamento. As áreas em verde, amarelo e vermelho nos gráficos indicam respectivamente elevada (≥ 28), intermediária (20 a 28)e baixa (<20) qualidade das leituras. Porto Alegre, 2015.

A limpeza das leituras visa obter um refinamento dos dados, mantendo unicamente com bases de boa qualidade. Para a obtenção de bases com alta qualidade, foi utilizado o script Trim.sliding.Window.pl, que trabalha com a média de qualidade dentro de "janelas". Desta forma, foram selecionadas apenas bases com qualidade superior a Q=15 (Figura 5), dessa forma, as bases que não alcançaram esses valores foram removidas.

Dois valores foram utilizados para a limpeza das leituras, sendo o primeiro no intervalo de 4-20 (Q20W4), onde as bases com qualidade superior a Q=20 seriam mantidas, e o segundo intervalo de 5-15 (Q15W5), onde todas as bases com qualidade acima de Q=15 foram mantidas. Contudo, quanto maior o valor da qualidade escolhido,

maior a quantidade de dados eliminados. Por exemplo, usando o intervalo Q20W4 cerca de 30% dos dados seriam eliminados para o AV53 e quase 40% para o AV60. Dessa maneira, optou-se por um refinamento menos crítico escolhendo-se um valor intermediário (Q15W5), porém com satisfatória qualidade das bases (Figura 5).

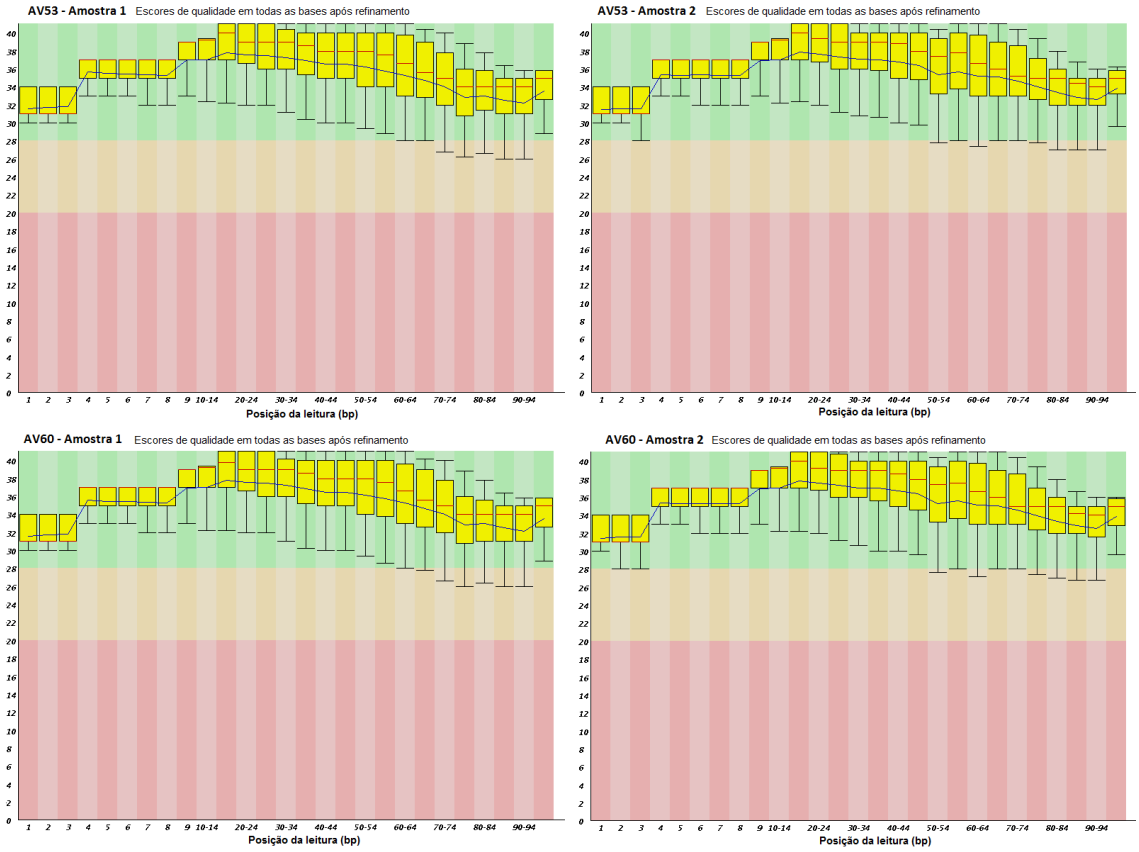


FIGURA 5. Boxplot mostrando a distribuição da qualidade das bases nas leituras para as amostras provenientes do sequenciamento dos genótipos AV53 e AV60 após o refinamento. As áreas em verde, amarelo e vermelho nos gráficos indicam respectivamente elevada (≥ 28), intermediária (20 a 28)e baixa (<20) qualidade das leituras. Porto Alegre, 2015.

Dessa forma, do total das bases sequenciadas, cerca de 16% (AV53) e 17% (AV60) foram removidas durante o processo de limpeza e refinamento, resultando em uma cobertura final do sequenciamento de 84,2 e 83% respectivamente para os genótipos AV53 e AV60 (Tabela 4). Qiu *et al.*(2014) encontraram resultados semelhantes no resequenciamento de três genótipos de arroz daninho, onde após a limpeza dos dados, cerca de 89% foram mantidos para análises daquele estudo.

TABELA 4. Qualidade das leituras do sequenciamento dos genótipos de arroz vermelho AV53 e AV60 após limpeza pelo Trimm. Porto Alegre, 2015.

Acessos		Nº leituras	Remanescente (%)
AV53	Raw	183317826	100
	Q15 W5	154313786	84,2
	Q20 W4	123771964	67,5
AV60	Raw	162784310	100
	Q15 W5	135043253	83,0
	Q20 W4	102147814	62,8

Diferentes programas montadores de genomas, como o Abyss, Soapdenovo, SSpace-Velvet e Ray foram testados para este trabalho. Dentre estes, o Abyss foi o que produziu maior número de *scaffolds*, sendo 1,140,984 para o genótipo AV53 (Tabela 5), enquanto que o programa Sspace-Velvet gerou 165,756 scaffolds mas também produziu o maior *scaffold* 118,873. Já o Ray gerou o menor número de *scaffolds* de 102,796. O programa Ray obteve o maior tamanho médio de *scaffolds* sendo de 2,816 pb contra 712 pb produzidos pelo programa Soapdenovo e 343 pb pelo programa montador Abyss. O tamanho médio dos contigs foi elevado para os programas Ray e SSpace-velvet sendo de 2,370 pb e 1,407 respectivamente. Esses dois programas montadores foram os que produziram menor número de *contigs*, onde o programa Ray produziu 120,157 contigs e o programa SSpace-Velvet produziu 224,643 contigs, enquanto que o programa Soapdenovo gerou 712,008 *contigs* e o programa Abyss mais de um milhão de *contigs* (1,175,841).

O software Abyss gerou um maior tamanho total dos *scaffolds* de aproximadamente 392 Mb, representando cerca de90% do tamanho total do genoma de referência do arroz cultivado (430 Mb). Porém a porcentagem de contigs não montados em *scaffolds* também foi elevada (57,1%), o que pode ser explicado pelo elevado número de contigs encontrados no sequenciamento (1,175,841). Já o programa Ray possui uma elevada porcentagem de contigs não montados em scaffolds (53,9%), mas produziu um baixo número de *contigs* (120,157).

TABELA 5. Resumo da montagem do genoma do acesso de arroz vermelho AV53, com detalhes das sequências, *contigs* e *scaffolds*. Porto Alegre, 2015.

AV53	Abyss	Soapdenovo	Ray	Sspace-Velvet
Número de Scaffolds	1,140,984	489,144	102,796	165,756
Tamanho Total de Scaffolds	391,707,156	348,249,259	289,520,149	321,716,683
Maior Scaffold	92,751	99,293	74,394	118,873
Menor Scaffold	55	100	101	101
Tamanho Médio Scaffold	343	712	2,816	1,941
N50	6,955	10,833	9,030	13,424
L50	13,941	8,577	9,588	6,814
% de contigs montados em Scaffold	42,9	81,5	46,1	76,6
% de contigs não montados em	57,1	18,5	53,9	23,4
Número de Contigs	1,175,841	712,008	120,157	224,643
Maior Contig	60,894	34,381	51,512	53,049
Menor Contig	54	3	101	51
Tamanho Médio de Contigs	331	469	2,370	1,407

TABELA 6. Resumo da montagem do genoma do acesso de arroz vermelho AV60, com detalhes das sequências, *contigs* e *scaffolds*. Porto Alegre, 2015.

AV60	Abyss	Soapdenovo	Ray	Sspace-Velvet
Número de Scaffolds	1,154,517	463,775	110,626	169,829
Tamanho Total de Scaffolds	390,265,005	343,986,957	284,688,908	320,944,801
Maior Scaffold	66,365	124,422	80,554	125,039
Menor Scaffold	55	100	100	101
Tamanho Médio Scaffold	338	742	476	1,890
N50	6,535	10,628	8,195	12,878
L50	14,892	8,731	10,399	7,118
% de contigs montados em Scaffold	42,2	81,8	47,1	76,8
% de contigs não montados em Scaffold	57,8	18,2	52,9	23,2
Número de Contigs	1,189,197	678,442	129,613	230,635
Maior Contig	57,411	30,441	49,880	49,791
Menor Contig	54	3	100	51
Tamanho Médio de Contigs	326	487	889	1,368

Para o genótipo AV60 (Tabela 6), o programa Abyss também foi o que gerou maior número de *scaffolds*(1,154,517). Já o programa Ray gerou o menor número de scaffolds de 110,626 e os programas SSpace-Velvet e Soapdenovo geraram valores

intermediários aos programas anteriores de 169,829 e 463,775 *scaffolds* gerados respectivamente.

Para este genótipo, o software Abyss gerou maior tamanho total dos *scaffolds*, cerca de 392 Mb,representando cerca de 90% do tamanho total do genoma de referência do arroz cultivado (430 Mb). Os programas Ray e SSpace-Velvet geraram menos *scaffolds* de 284,688,90 e 320,944,801, o que pode ser explicado pelo número de *contigs* gerados (129,613 e 230,635 respectivamente) e pelo tamanho médio dos *contigs* montados de 889 pb no programa Ray e 1,368 para o programa Sspace-velvet.

Para o genótipo AV60, também foi elevada a porcentagem de *contigs* não montados em *scaffolds* do programa Abyss (57,8%), podendo ser atribuído ao elevado número de *contigs* produzidos, por volta de 1,189,197. O programa Ray obteve uma elevada porcentagem de *contigs* não montados em *scaffolds* (52,9%), contudo o número de *contigs* produzidos foi de 129,613.

O valor de N50 representa o comprimento do menor *contig* que, quando adicionado a um conjunto de *contigs* maiores produz pelo menos 50% do genoma. Essa abordagem em relação ao número de *contigs* e tamanho médio de *contig*, é amplamente utilizado para comparação dos programas de montadores, pois quanto maior o valor N50, melhoré o conjunto de dados, conferindo alta cobertura do genoma (Farias, 2013). Além disso, o número de *contigs* e a média do tamanho de contigs fornecem uma estimativa do tamanho das peças que fazem a montagem. Dessa forma, um número pequeno de *contigs* e uma média alta do tamanho deste *contigs* são indicadores de uma boa montagem do genoma.

Nesse sentido, novamente o programa Abyss gerou em média os menores *contigs* sendo 331pb e 326pb para os genótipos AV53 e AV60 (Tabelas 5 e 6), com um valor de N50 de 6,955 para o AV53 e 6,535 para o AV60. Sendo que este valor é menor do que os apresentados pelos outros programas montadores, como o Soapdenovo e o Sspace-Velvet, que apresentaram um valor de N50 igual a 10,833 e 13,424 respectivamente para o AV53 e de 10,628 e 12,878 respectivamente para o AV60.

Esse fato, juntamente com o maior número de *scaffolds* gerados e o maior tamanho de *scaffolds* montados, demonstra a maior eficácia do software Abyss na montagem e qualidade da montagem de genomas, em relação aos outros programas montadores. Dessa forma, esse programa montador foi o escolhido para ser utilizado na montagem que foi empregada nas análises realizadas ao longo deste trabalho.

Resultados semelhantes foram encontrados em outros estudos, como Farias

(2013), que trabalhou com o sequenciamento e montagem do genoma de *Oryza glumaepatula*, onde o programa Abyss também apresentou a maior quantidade de bases montadas em scaffolds, por volta de 94%. Similarmente, Desai *et al*.(2013) encontraram até 95% de cobertura do genoma utilizando este programa para montagem do genoma de *C. elegans*. Também Qiu *et al*. (2014) encontraram alta cobertura e qualidade na montagem do genoma de três genótipos de arroz daninho, utilizando o mesmo programa montador.

4.2 Análise estrutural de genes relacionados ao degrane

A estrutura gênica de sequências nucleotídicas podem ser comparadas por meio de gráficos do tipo *dot plot*. Nesse tipo de gráfico uma sequência é colocada no eixo y e a(s) outra(s) no eixo x, sendo que a intersecção de cada nucleotídio idêntico entre as sequencias, é representado por um ponto. Dessa forma, a formação de uma linha diagonal contínua representa perfeita homologia entre as sequências comparadas e a formação de linhas descontínuas pode apresentar inserções ou deleções.

O degrane é um dos mais importantes caracteres selecionados durante a domesticação dos cereais (Fuller *et al*., 2009; Harlan, 1992). Recentemente, vários genes relacionados com a domesticação começaram a ser identificados, principalmente no arroz cultivado, e entre eles os genes relacionados ao degrane. Entre eles, já foram clonados 2 QTLs de grande efeito, *qsh1* e *sh4*, que explicam cerca de 70% da variação fenotípica do degrane (Li*et al*., 2006; Konishi *et al*., 2006).

Além dos genes anteriores, o*OsCPL1*, *Os01g0849100*, *OsCel9D* e *OsXTH8*também possuem efeito significativo na expressão do caracter degrane. Dessa forma, a sequências desses 6 genes foram buscadas no banco de dados RAPDB para serem comparadas com as suas sequências homólogas nos acessos de arroz vermelho sequenciadas no presente estudo.

O *sh4* está localilizado no cromossomo 4 e possui 2035 pares de base e conta com dois éxons separados por um intron (Figura 6), estruturalmente nos genótipos de arroz vermelho estudados (AV53 e AV60) são basicamente idênticos à estrutura dos genomas utilizados como referência para esta análise, *japonica* e *indica*, exceto por uma leve descontinuidade nas sequências, sendo a do AV60 um pouco maior. Estas descontinuidades podem ser consideradas deleções e podem ter sido originárias de diversos fatores. Em primeiro lugar o sequenciamento utilizou apenas uma biblioteca de 330 kb e além disso, regiões de sequências repetitivas, como ocorre no final do gene *sh4*,

são difíceis de montar (Markus, 2013). Outro fator a ser considerado é a possível inserção de elementos tranponíveis (Panaud *et al*., 2009).

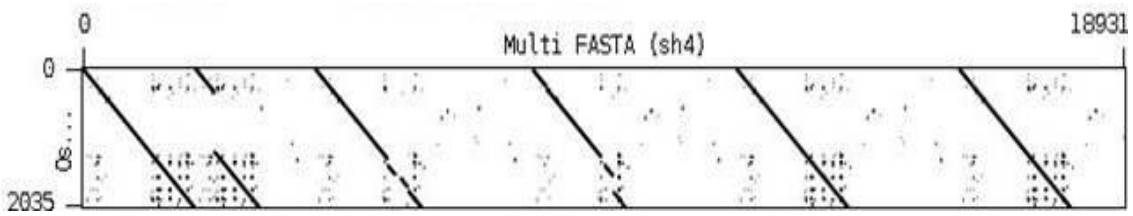


FIGURA 6. Dotplot do locus do gene Sh4, região codificante (cds), sequênciasprovenientes dos acessos AV53 e AV60 (2,1,3), e sequênciasoriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015.

Além disso, análise de distância evolutiva mostra que as sequências protéicas traduzidas a partir do gene *sh4* dos acessos AV53 e AV60 apresentam elevada similaridade entre si e com a espécie *O. indica* (Figura 7).

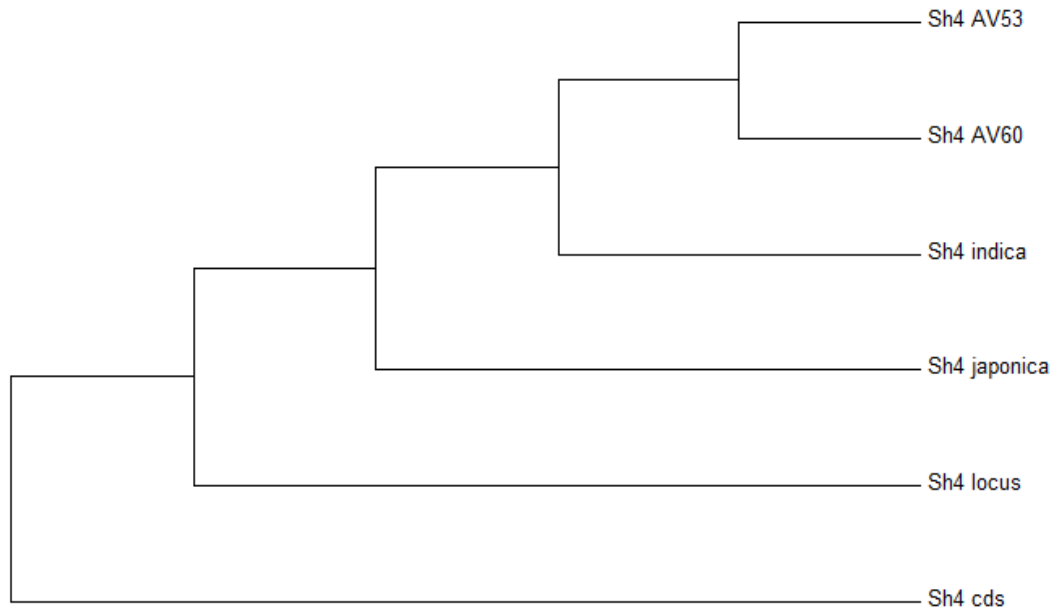


FIGURA 7. Distância evolutiva inferida pelo método de agrupamento de vizinho mais próximo entre as sequências de proteína do gene de degrane sh4 provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1026 aminoácidos. Porto Alegre, 2015.

Já o gene *OsCPL1* por sua vez, age como repressor da diferenciação da camada de abscisão (Ji *et al*., 2010). Alguns estudos identificaram que um SNP (G/T) localizado no éxon 8 do gene *OsCPL1* muda o aminoácido conservado serina para isoleucina,

fazendo com que o fenótipo apresente degrane (Ji *et al*., 2010).

O locus do gene *OsCPL1* possui 5709 pb, com 6 éxons e 6 introns. A estrutura do gene aparece de forma completa em quase todos os genótipos estudados (Figura 8). Porém há uma aparente deleção de uma parte da sequência da região anterior ao início do gene no genótipo AV60 que tambem pode ser vizualizada no Apêndice 2.

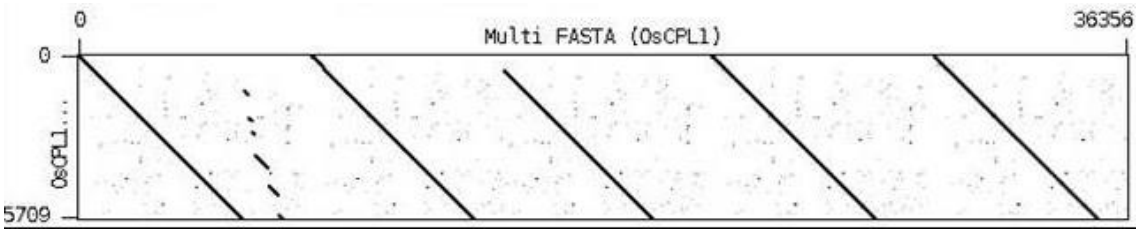


FIGURA 8. Dotplot do locus do gene OsCPL1, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015..

Além disso, a análise de distância evolutiva mostrou uma maior identidade entre as sequências do AV53 e *O. japonica*, enquanto que a sequência do *OsCPL1* no AV60 foi mais similar ao locus gênico. Nunes (2012) verificou que ecótipos de arroz vermelho que possuem alto degrane, a expressão desse gene foi superior em relação as cultivares que possuem menor nível de degrane. Dessa forma, a expressão do gene *OsCPL1* estaria relacionada com a ativação do processo de abscisão, pois a expressão do gene estaria relacionada com a presença do degrane e não com a repressão do degrane como o observado nas análises histoquímicas do estudo de Ji *et al*. (2010).

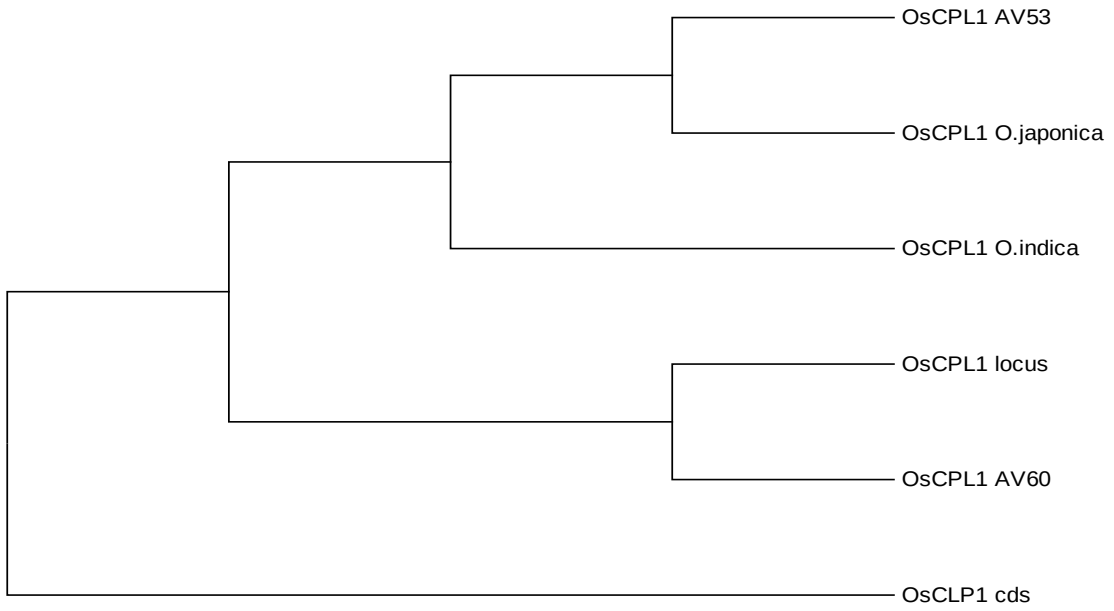


FIGURA 9. Distância evolutiva inferida pelo método de agrupamento de vizinho mais próximo entre as sequências de proteína do gene de degrane OsCPL1 provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1246 aminoácidos. Porto Alegre, 2015.

Os01g084910 é outro gene relacionado ao degrane que age codificando uma proteína com um domínio que estimula a troca de nucleosídeos difosfatados por nucleosídeos trifosfatados (Berkenet *etal*., 2005). Este gene está posicionado a 34kb *upstream* ao gene *qsh1* e apresenta expressão na região entre o pedicelo e a flor.

Este gene apresenta 4507pb, possui 7 éxons e uma região intergênica proeminente (Figura 10). Pode ser observada a possível presença de um elemento transponível para os genótipos analisados, embora estruturalmente são idênticos. A análise de distância evolutiva (Figura11) novamente mostrou mair similaridade entre as sequências do *Os01g0849100* nos acessos AV53 e AV60 com a a referência de O. indica, sugerindo que pode ser a espécie de origem do arroz vermelho.

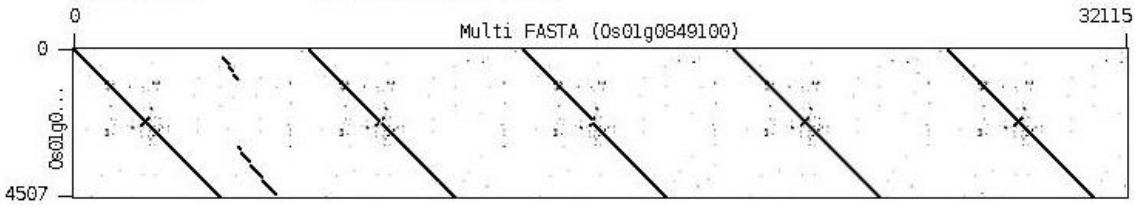


FIGURA 10. Dotplot do locus do gene Os01g0849100, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015.

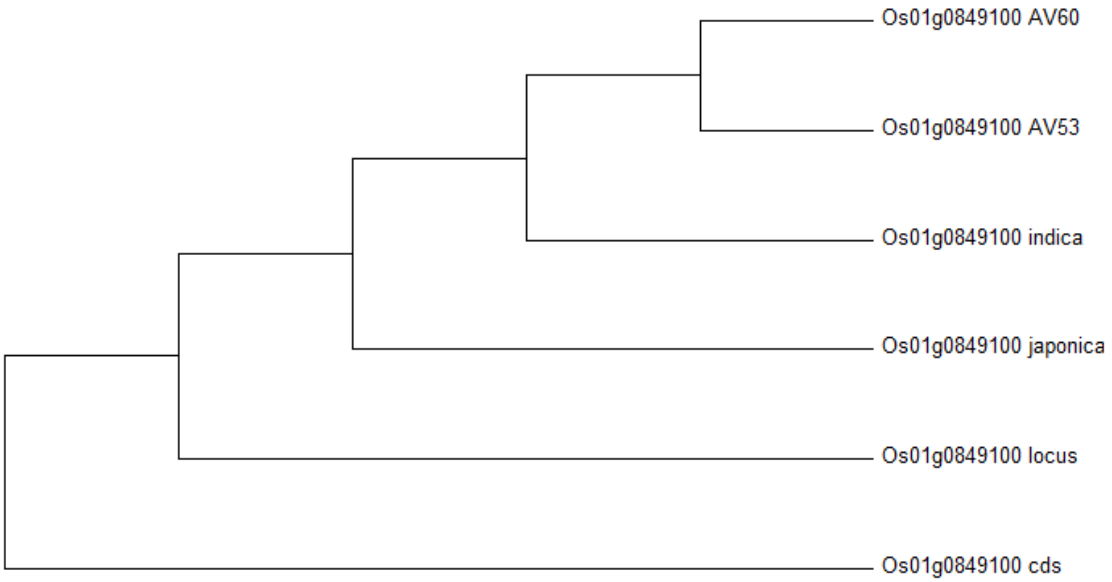


FIGURA 11. Distância evolutiva inferida pelo método de agrupamento vizinho mais próximo entre as sequências de proteína do gene de degrane Os01g0849100 provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1637 aminoácidos. Porto Alegre, 2015.

O gene *qsh1* localiza-se no cromossomo 1 e alguns estudos observaram a presença de um SNP que corresponde a substituição do nucleotídeo G por T na região regulatória 5' deste gene, a 11841 bases upstream (Konishi et al 2006). De acordo com a análise estrutural deste gene (Figura 12), verifica-se que o referido gene apresenta 4495pb, quatro éxons e três introns e está presente em ambos os genótipos de arroz vermelho estudados (AV53 e AV60), assim como nas espécies referência de arroz (*japônica* e *indica*) de maneira completa, sem a intersecção por elementos transponíveis ou demais variáveis estruturais. Evolutivamente, o gene *qsh1* presente nos genótipos sequenciados, estão relacionados com o genótipo de referência *O. japonica*(Figura 13). Também, Konishi et al (2006), avaliando a expressão deste gene, localizou um SNP que ocasiona a perda da expressão do gene *qsh1* somente na região entre o pedicelo e a flor, que é responsável pela origem da ausência de debulha no arroz cultivado, corroborando com os dados do estudo de Nunes (2012) onde verificou-se que o gene *qsh1* não foi expresso na região entre o pedicelo e a flor aos 10 dias após a polinização.

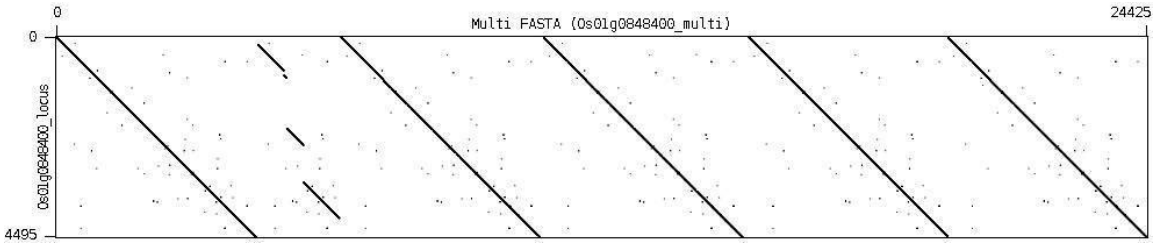


FIGURA 12. Dotplot do locus do gene qsh1, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015.

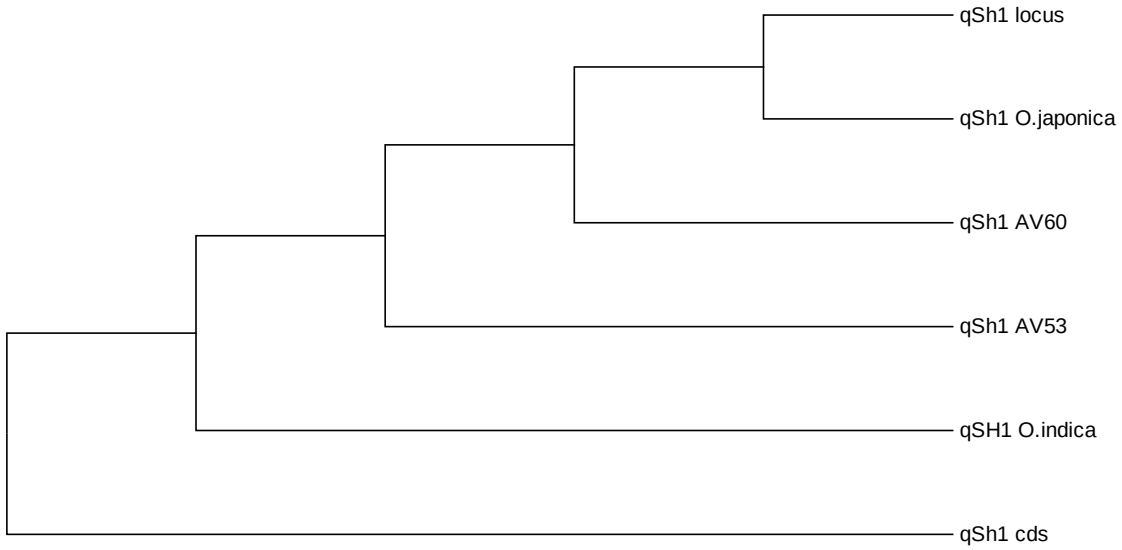


FIGURA 13. Distância evolutiva inferida pelo método de agrupamento de vizinho mais próximoentre as sequências de proteína do gene de degraneqsh1 provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1831 aminoácidos. Porto Alegre, 2015.

O gene *OsCel9D* codifica uma proteína do tipo endo-1,4-betaglucanase. Este gene possui 4216 pb, 6 éxons e 7 introns (Figura 14). Além disso, a análise estrutural este gene indica que pode haver a inserção de um elemento transponível para todos os genótipos estudados e uma possível deleção no genótipo AV53 e em escala menor, no genótipo AV60, o que não aparece nas espécies utilizadas como referência, que são identicas entre si.

Nunes (2012) avaliou a expressão relativa desse gene em dois genótipos com alto degrane (AV31 e AV60) e dois genótipos de baixo degrane (Batatais e Lacassine) e verificou maior expressão relativa em genótipos com baixo nível de degrane, sugerindo que há uma relação inversa entre o nível de expressão relativa e o nível de degrane das

sementes. Isto também indica que a superexpressão deste gene reduziria o nível de degrane em espécies de *O. sativa* cultivadas. Isso também pode ser observado na análise de distância evolutiva (Figura 15), onde o gene *OsCeI9D* presente no genótipo AV60, que fenotipicamente apresenta elevado degrane, está mais proximamente relacionado com a sequência de referência de*O. japonica*.

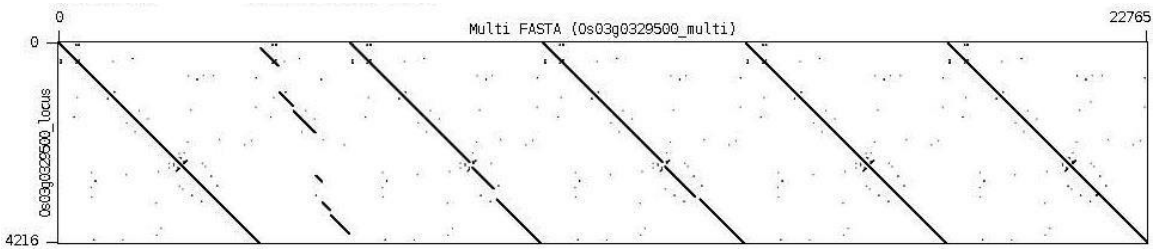


FIGURA 14. Dotplot do locus do gene OsCel9D, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015.

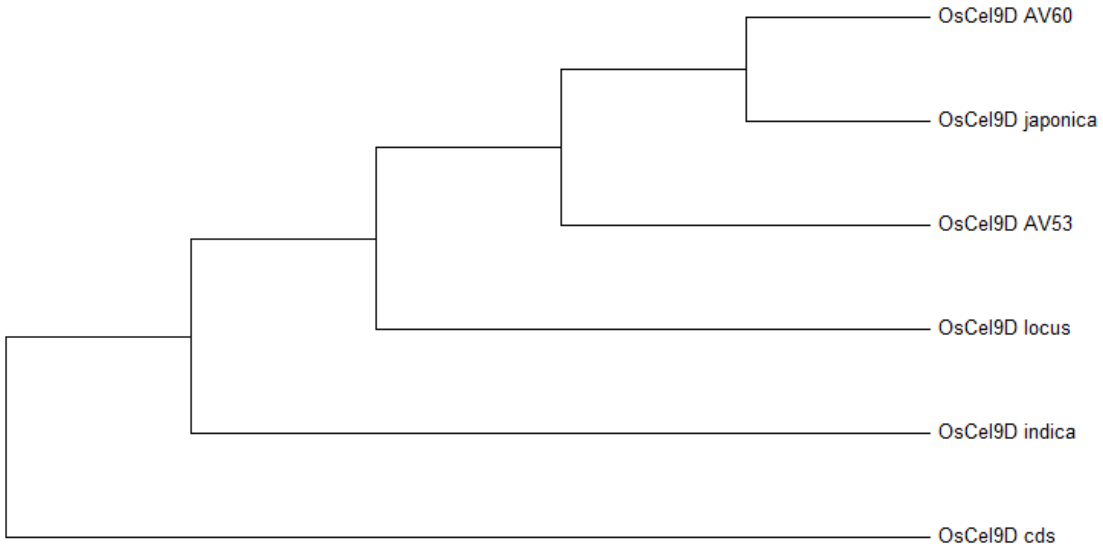


FIGURA 15. Distância evolutiva inferida pelo método de agrupamento vizinho mais próximo entre as sequências de proteína do gene de degrane OsCeI9D provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 1783 aminoácidos. Porto Alegre, 2015.

O gene OsXTH8 é expresso em elevados níveis em células que estão ativamente em processo de alongamento e diferenciação e, plantas contendo contruções com o gene silenciado apresentam crescimento limitado (Jan et al 2004).

Estruturalmente este gene possui 1224 pb e 2 éxons e 3 íntrons. A estrutura gênica

parece ser idêntica para os genótipos estudados (Figura16).Porem o genótipo AV60, parece ocorrer uma pequena deleção, que pode ser explicada pela ação de um elemento transponível. Assim como em estudo realizado por Nunes (2012), encontrou uma variação nucleotídica do gene OsXTH8 demonstrou apenas mutações no intron 1 do gene. No entanto estas mutações não parecem ter uma relação com o nível de degrane, pois elas estão presentes tanto em genótipos com degrane, quanto naqueles sem degrane (Nunes, 2012).

Pelo estudo da distância evolutiva (Figura 17), o gene OsXTH8 presente nos genótipos sequenciados (AV53 e AV60), está relacionado com o genótipo de referência *O. indica*.

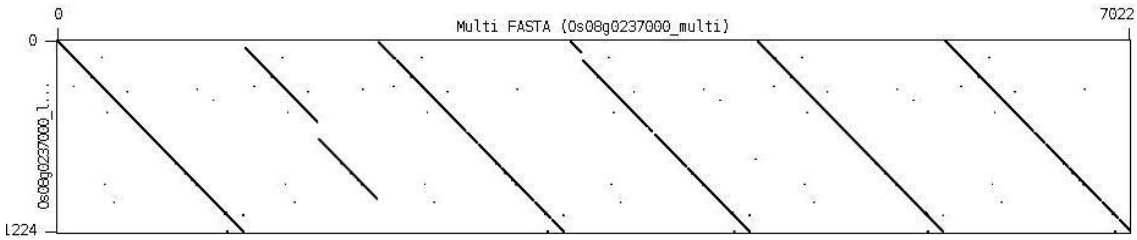


FIGURA 16. Dotplot do locus do gene OsXTH8, região codificante (cds), sequências provenientes dos acessos AV53 e AV60 (2,1,3), e sequências oriundas dos genomas de referência do *O. japonica* e *O. indica*, respectivamente. Porto Alegre, 2015.

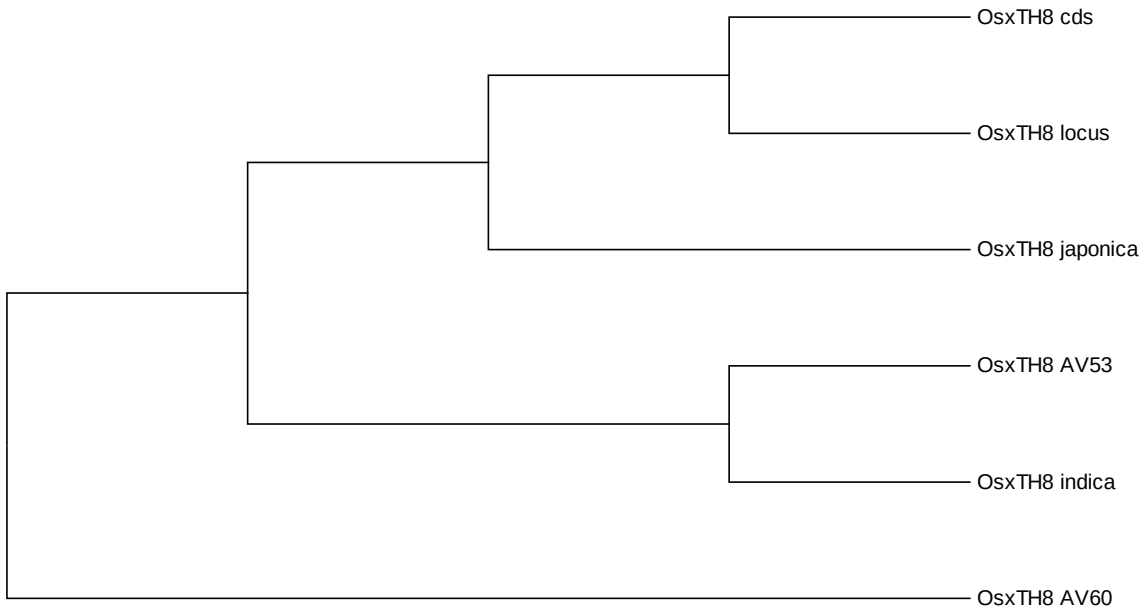


FIGURA 17. Distância evolutiva inferida pelo método de agrupamento vizinho mais próximo entre as sequências de proteína do gene de degrane OsxTH8 provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Análise foi baseada no alinhamento de 845 aminoácidos. Porto Alegre, 2015.

Em estudo similar, Wang *et al*. (2014) investigaram um conjunto de genes relacionados ao degrane em *O. sativa*que também foram artificialmente selecionados durante a domesticação do arroz africano. Ainda nesse estudo, verificou-se que foi selecionado inconscientemente para mutações que previnem ou reduziem o degrane localizadas nos genes ortólogos *OsSh1*, *Sh4* e *qSh1*. Os genótipos de arroz vermelhos sequenciados no presente estudo não foram domesticados e dessa forma espera-se que os mesmos não apresentem tais mutações, entretanto seriam necessárias análises adicionais para comprovar tal hipótese.

Nenhuma diferença estrutural de grande relevância foi encontrada por esta análise. Contudo, estes dados, contribuem para o melhor conhecimento dos genes relacionados ao degrane dos genótipos de arroz vermelho do Sul do Brasil. De maneira geral, os genomas utilizados como referencia são em estrutura idêntica entre si, já os genótipos de arroz vermelho diferem entre si e dos genomas de referência para os genes, com exceção do gene *qsh1*. Estas diferenças podem estar relacionadas a elementos transponíveis, mutações, inserções ou deleções e também pode ser pelo fato de, nesta análise, utilizarmos as sequencias oriundas do sequenciamento e montagem dos genomas utilizando apenas uma biblioteca pair-end, o que gerou alguns *scaffolds* menores.

4.3 Detecção de SNPs e INDELs

Variantes genômicos do tipo SNP (*Single Nucleotide Polymorphism*) são polimorfismos de DNA mais abundantes no genoma e pode ser definido como um sítio do DNA onde se observa a substituição de uma única base entre indivíduos de uma mesma população (Risch & Merikangas, 1996) Enquanto que INDELs referem-se ao número de inserções e deleções de bases nucleotídicas no genoma.

Teoricamente é possivel ocorrer quatro alelos diferentes para cada nucleotídeo em um sítio SNP, uma vez que o DNA é composto por quatro bases nitrogenadas (A, C, T, G). Entretanto, observa-se que a presença em maior frequência de apenas duas possíveis variações, fato que pode ser explicado pela ocorrência desigual de substituições de base do tipo transição (TS) (A ↔ G, T ↔ C) e transversão (TV) (A ↔ C, A ↔ T, G ↔ C, G ↔ T). Sendo assim, a razão TS/TV refere-se ao número de SNPs originados por transição dividida pelo número de SNPs causados por transversões (Figura 20 e 21). Apesar do número possível de transversões ser em média duas vezes maior do que o de transições, o que se observa na prática é que a ocorrência de transições é cerca de duas vezes maior

do que de transversões (Vignal, 2002), fato que pode ser explicado pela alta taxa de deaminação espontânea da 5-metil-citosina em timina em dinuclotídeos CpG (Coulondre, 1978), tendência essa também comprovada no presente estudo, onde os valores de TS/TV variam de 2.30 a 2.56 (Tabela 7) e corroborando com os resultados relatados por DePristo *et al*. (2011).

TABELA 7. Estatísticas da detecção de SNPS e INDELS entre os genomas de arroz vermelho (AV53 e AV60) alinhados com os genomas refêrencia *O. japonica* e *O. indica*. Porto Alegre, 2015.

Genomas alinhados	SNPs	TS/TV	INDELs	Singletons (AC=1)			Multialélicos	
				SNPs	TS/TV	INDELs	sites	SNPs
AV53 x *O.japonica*	2,518,726	2.56	419685	7.8%	2.30	11.4%	6275	552
AV60 x *O.japonica*	2,567,308	2.56	421990	3.80%	2.24	6.9	5375	386
AV53 x *O.indica*	1,268,507	2.36	259443	14.9%	2.36	14.30%	2846	216
AV60 x *O.indica*	1,774,756	2.45	338591	5.10%	2.23	6.5	3627	227

No alinhamento dos genomas do AV53 com o *O. japonica*, foram encontrados 2,518,726 SNPs e 419,685 INDELs, enquanto que para o AV60 foram encontrados 2,567,308 SNPs e 421,990 INDELs, respectivamente (Tabela 7). Por outro lado, o número de variantes genômicos foi relativamente menos frequente quando alinhados contra o genoma do *O. indica*, indicando que há maiores semelhanças entre os genomas dessas espécies e sugerindo que o arroz vermelho orginou-se a partir dessa espécie (Figura 18).

Esses valores sao relativamente maiores do que aqueles reportados por Feltus et al (2004) que depois de filtrar sequencias com múltiplas copias e baixa qualidade encontraram 384341 SNPs e 24557 INDELS de base única no alihamento das subespecies *O. indica* e *O.japonica*, resultando numa taxa de polimorfismo de 1,70 SNPs/kb e 0.11 INDEL/kb. Alem disso, a maior parte desses polimorfismos foram transições (61,8%) seguido de transversões 32,8% (TS/TV=~2.), e INDELs representando 6,0%. A menor taxa de polimorfismo entre essas subespecies pode em função de ambas terem sido domesticadas, enquanto que no presente estudo foram alinhados os genomas de acessos selvagens (arroz vermelho) com os genomas referencia do arroz cultivado (*O. indica* e *O. japonica*), resultado em maior numero de variantes genomicos.

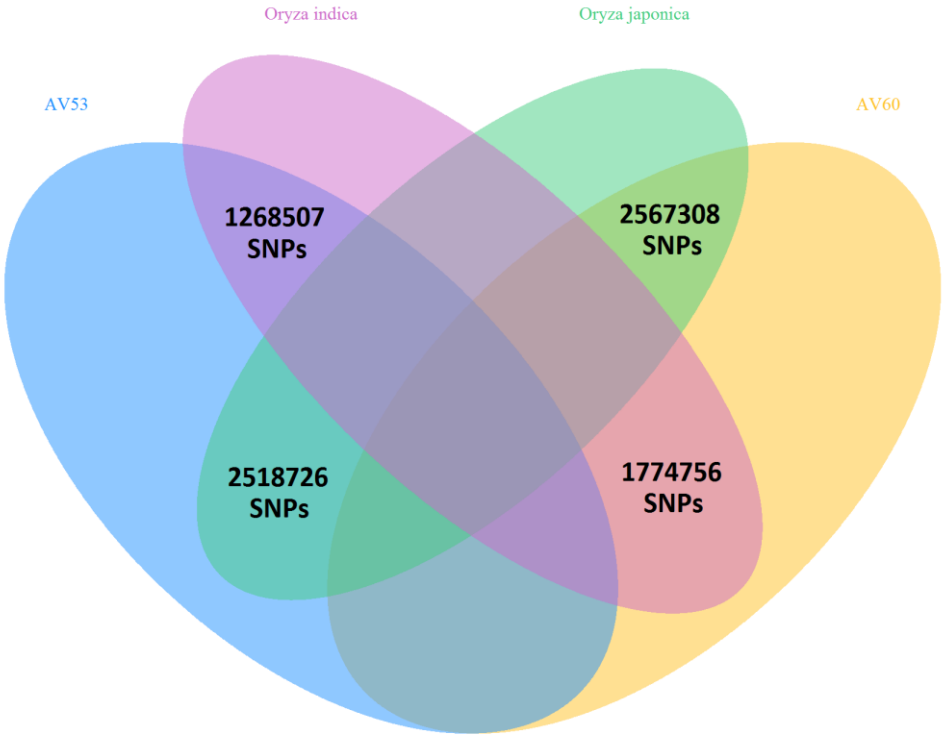


FIGURA 18. Diagrama de Venn com o número de SNPs compartilhados entre os genomas dos acessos AV53, AV60, *Oryza indica* e *Oryza japonica*. Porto Alegre, 2015.

A frequência de SNPs e INDELs foi determinada a partir da divisão do número de pares de bases dos *scaffolds* gerados pelo programa Abbys (Tabelas 5 e 6) pelo número respectivo de variantes encontrados em cada alinhamento (Tabela 7). Verificou-se em média a ocorrência de um SNP a cada 154 bp no alinhamento dos genótipos de arroz vermelho contra o genoma do *O. japonica*, enquanto que quando alinhado contra o *O. indica* a occorência foi em media de um SNP a cada 264 bp (Figura 19). Esses resultados estão de acordo com outros estudos similares; por exemplo, em *arabidopsis*observa-se em média 1 SNP a cada 3300 (Drenkard *et al*., 2000); em soja, 1 SNP a cada 200 pb (Graef & Diers, 2004); em milho são ainda mais frequentes, podendo chegar a 1 SNP a cada 70 pb (Bhattramakki, 2000), e em arroz, observa-se em média 1 SNP a cada 232 pb entre variedades escolhidas ao acaso (Nasu, 2002). Além disso, estima-se que o número de inserções e deleções curtas (1^{-10} pb) em gramíneas representam mais de 90% do total de INDELS detectadas e que as espécies mais estreitamente relacionados tendem a ter uma maior proporção de INDELS curtos (Xu*et al*., 2012).

Cabe ainda ressaltar que o número de SNPs observados em uma espécie depende, naturalmente, das relações de vínculo genético entre as amostras de acessos utilizadas na

análise. Se as amostras apresentam grande diversidade genética, a tendência é se observar maior número de SNPs a cada kpb analisado, como foi verificado no presente estudo entre os acessos de arroz vermelho e o genoma do *O. japonica*. Isso contribui para confirmação de que o arroz vermelho econtrado no sul do Brasil é proveniente da especie *O. indica.*

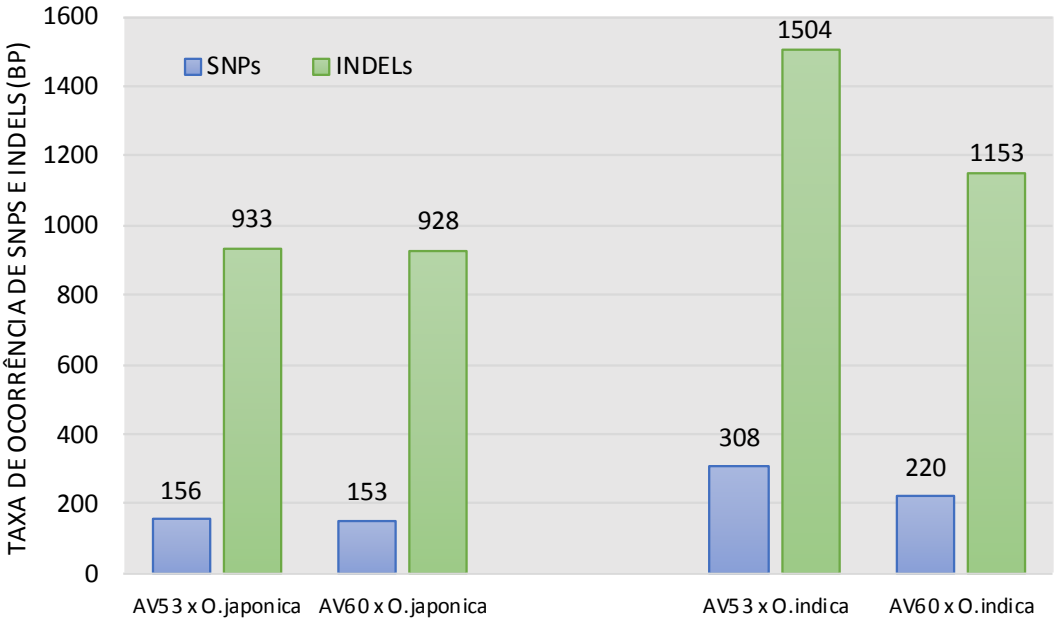


FIGURA 19. Ocorrência de SNPS e INDELs encontrados nos alinhamentos do AV53 e AV60 contra os genomas de referência *O. japonica* e *O. indica*. Porto Alegre, 2015.

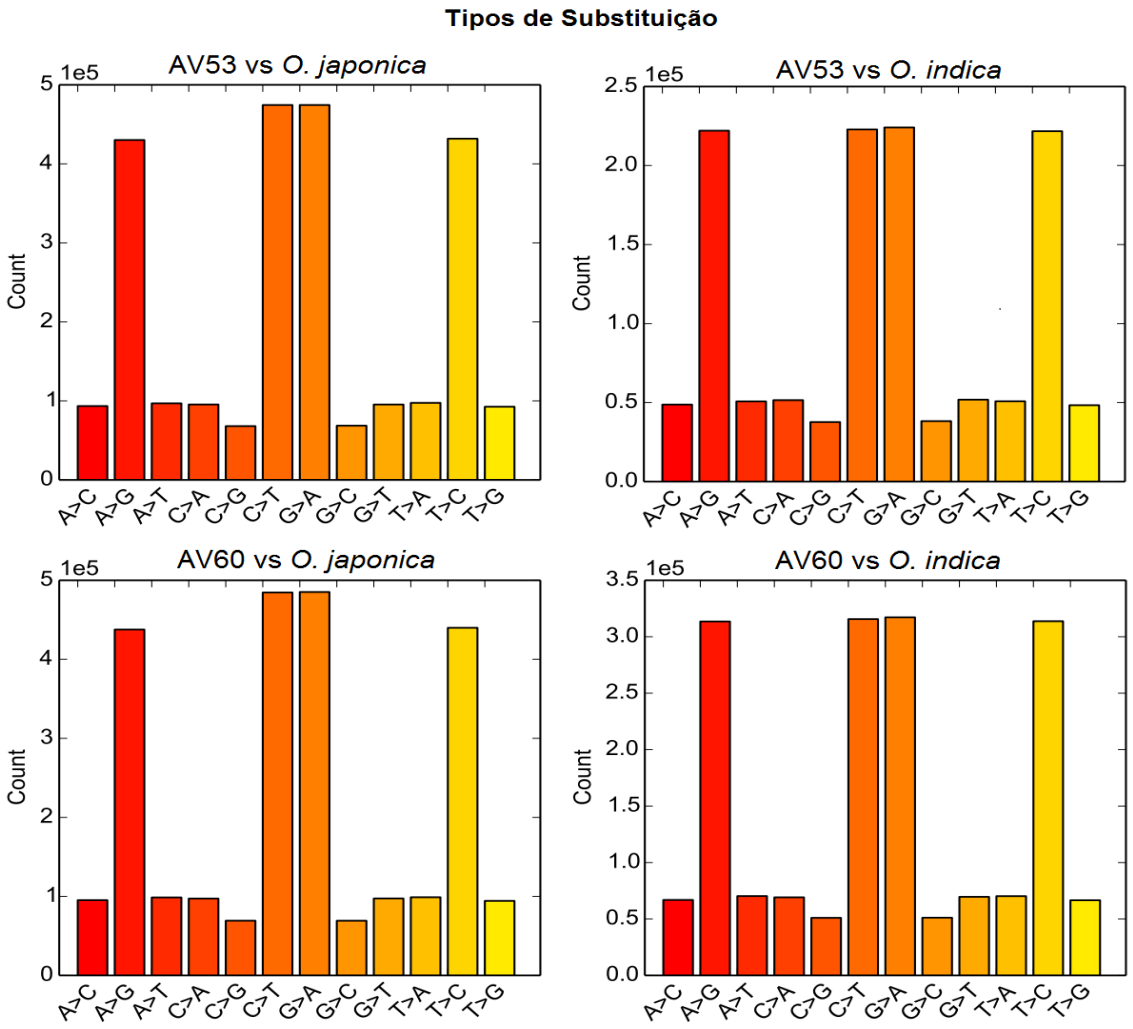


FIGURA 20. Gráfico mostrando frequências dos tipos de substituição, transversões (TV) e transições (TS). Porto Alegre, 2015.

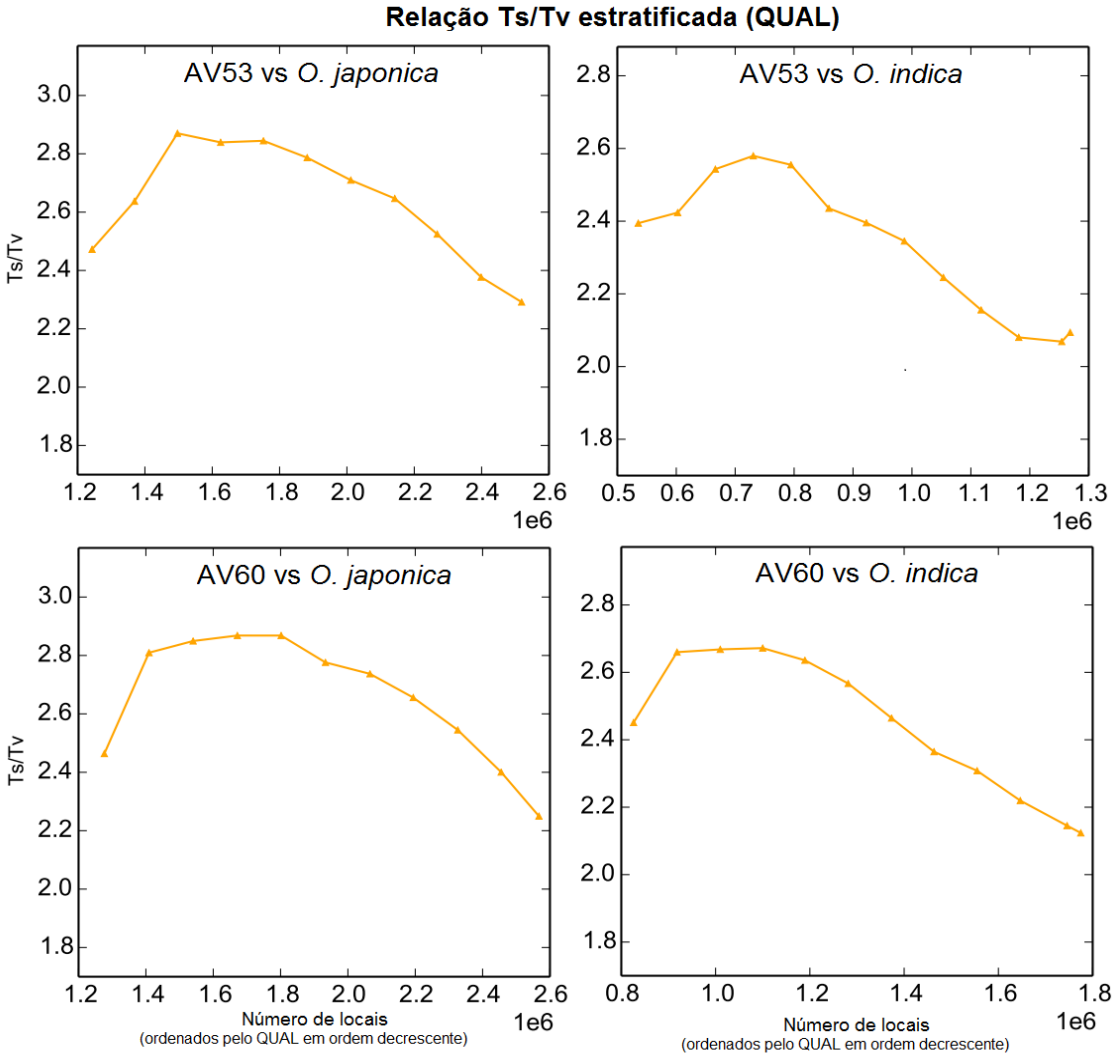


FIGURA 21. Distribuição dos valores da relação TS/TV (substituição transição/transversão), ordenadas pela qualidade do sequenciamento ao longo do genoma. Porto Alegre, 2015.

A disponibilidade de grandes quantidades de dados seqüência de DNA, de forma cada vez mais barata e rápida vem causando enorme impacto na nossa compreensão da biologia das plantas. Com isso, análises de bioinformática tornam-se o maior o desafio atual, em função da necessidade de computadores cada vez mais potentes para analisar grande volume de dados. Além disso, também é requerido que usuários bem treinados capazes de manusear diversos programas montadores de forma eficiente do ponto de vista computacional e financeiro (Hadfield & Eldridge, 2014). Por fim, o uso de várias ferramentas simultaneamente pode ser confuso, por isso na maioria dos casos usuários limitam-se a poucos métodos de analisar os dados.

5 CONCLUSÕES

A cobertura do sequenciamento foi de 32.1 a 36.6 vezes e a montagem pelo programa Abyss gerou um tamanho do genoma de 391,7 e 390,2 Mb, respectivamente para os genótipos AV53 e AV60.

Visando obter uma montagem mais precisa dos genomas recomenda-se que em estudos futuros sejam construídas múltiplas bibliotecas *paired-end* e *mate-paired*.

A análise estrutural de seis genes relacionados ao degrane relevou que a composição gênica é similar entre os genótipos analisados, corroborando com resultados de estudos anteriores.

A menor frequência de SNPs e INDELs no alinhamento dos genótipos de arroz vermelho com o genoma de *Oryza sativa spp indica*, evidencia maior similaridade entre os genomasalinhados, sugerindo que o arroz vermelho é originário dessa subespécie.

6 REFERÊNCIAS BIBLIOGRÁFICAS

AGOSTINETTO, D. et al. Arroz vermelho: ecofisiologia e estratégias de controle.**Ciência Rural**, Santa Maria, v. 31, n. 3, p. 341-349, 2001.

ALTSCHUL, S.F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, Cambridge, v.215, n.3, p.403-410, 1990.

ARABIDOPSIS GENOME INITIATIVE. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. **Nature**, London, v.408, n.6814, p.796–815, 2000.

AVILA, L. A. D. et al. Banco de sementes de arroz vermelho em sistemas de semeadura de arroz irrigado. **Ciência Rural**, Santa Maria, v. 30 p. 773-777, 2000.

BRANTON, D. et al. The potential and challenges of nanopore sequencing. **Nature biotechnology**, London, v. 26, p. 1146-1153, 2008.

BAO, S. et al. Evaluation of Next-Generation Sequencing Software In Mapping And Assembly. Retraction. In: EVALUATION of next-generation sequencing software in mapping and assembly. **Journal Human Genetics**, New York, v. 56, p.406-414, 2011.

BENNETZEN, J.L. Mechanisms and rates of genome expansion and contraction in flowering plants. **Genetica**, Dordrecht, v.115, p.29–36, 2002.

BERGMAN, C.M.; QUESNEVILLE, H. Discovering and detecting transposable elements in genome sequences. **Briefings in Bioinformatics**, London, v.8, p.382–392, 2007.

BOISVERT, S.; LAVIOLETTE, F.; CORBEIL, J. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. **Journal of Computational Biology**, New York, v.17, n.11, p.1519–1533, 2010.

BULLER, F. et al. High-throughput sequencing for the identification of binding molecules from DNA-encoded chemical libraries. **Bioorganic & Medicinal Chemistry Letters**, Amsterdam, v. 20 p. 4188-419, 2010.

CAI, H. W.; MORISHIMA, H. Genomic regions affecting seed shattering and seed dormancy in rice. **Theoretical and Applied Genetics**, New York, v. 100, n. 6, p. 840-846, 2000.

CHURCH, G.M. et al. **Characterization of individual polymer molecules based on monomer-interface interaction**. US 5,795,782. 1998.

COMPANHIA NACIONAL DE ABASTECIMENTO (CONAB). **Perspectivas para a agropecuária.**Brasília, 2014. v.2, 155p.

CONNELL, C. et al. Automated DNA-Sequence analysis. **Biotechniques**, Natick, v.5, p.342–348, 1987.

COUNCE, P.A.; KEISLING, T.C.; MITCHELL, A.J. A uniform, objective, and adaptive system for expressing rice development. **Crop Science**, Madison, n.40, p.436-443, 2000.

CRONN, R. et al. Multiplex sequencing of plant chloroplast genomes using Solexa-sequencing-by-synthesis technology. **Nucleic Acid Research**, London, v. 36, n.19, e122, 2008.

DELOUCHE, J.C. et al. Weedy rices-origin, biology, ecology, and control. Rome: FAO, 2007.144 p. (FAO Plant Production and Protection Paper, 188)

DEPRISTO MA., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. **Nature Genetics**, London, v. 43, n.5 p.491-498, 2011.

DESCHAMPS, S. et al. Rapid genome-wide single nucleotid polymorphism discovery in soybean and rice via deep resequencing of reduced representation libraries with the Illumina Genome Analyzer. **The Plant Genome**, Madison, v.3 p. 53-68, 2010.

DIARRA, A.; SMITH JR., R.J.; TALBERT, R.E. Growth and morphological characteristics of red rice (Oryza sativa) biotypes. **Weed Science**, Champaign, v.33, p.310-314, 1985.

EMBRAPA. **Cultivo do arroz irrigado no Brasil**. Pelotas: Embrapa Clima Temperado, 2005. v. 1, 85 p.

FARIAS, D. **Desenvolvimento de ferramenta in silico para estudos de genômica em arroz e montagem da espécie selvagem Oryza glumaepatula Steud**. 2013. 79 f. Tese (Doutorado) - Programa de Pos-Graduação em Fitomelhoramento, Universidade Federal de Pelotas, Pelotas, 2013.

FLAVELL, R.B.; et al. Genome size and proportion of repeated nucleotidesequence DNA in plants. **Biochemical Genetics**, New York, v.12, n.4, p.257–269, 1974.

FELTUS F.A et al. An SNP Resource for Rice Genetics and Breeding Based on Subspecies Indica and Japonica Genome Alignments. **Genome Research**, Cold Spring Harbor, v. 14, p. 1812–1819, 2004.

FENG Y, et al. Nanopore-based Fourth-generation DNA Sequencing Technology. **Genomics, Proteomics & Bioinformatics**, Beijing, v. 13, p. 4-16, 2015.

FENG, Q. et al. Sequence and analysis of rice chromosome 4. **Nature**, London, v. 420, n. 6913, p. 316-320, 2002.

FINKELSTEIN, R. et al. Molecular aspects of seed dormancy. **Annual Review of Plant Biology**, Palo Alto, v.59, p.387-415, 2008.

FLECK, N. G. et al. Relative competitivity among flooded rice cultivars and a red rice biotype. **Planta Daninha**, Londrina, v. 26, n. 1, p. 101-111, 2008.

GEALY, D. R.; MITTEN, D. H.; RUTGER, J. N. Gene flow between red rice (*Oryza sativa*) and herbicide-resistant rice (O. saliva): implications for weed management. **Weed Technology**, Champaign, v. 17, n. 3, p. 627-645, 2003.

GOMES, A. S.; MAGALHÃES JUNIOR, A. M. M. **Arroz Irrigado no Sul do Brasil**. Brasília: Embrapa Informação Tecnológica, 2004. 900 p.

GU, X. Y. et al. Genetic analysis of adaptive syndromes interrelated with seed dormancy in weedy rice (*Oryza sativa*). **Theoretical and Applied Genetics**, New York, v.110, p.1108-1118, 2005.

HABERER, G.; FISCHER, T. C.; TORRESRUIZ, R. A. Mapping of the nucleolus organizer region on chromosome 4 in Arabidopsis thaliana. **Molecular & GeneralGenetic**, Berlin, v.15, n.250, n.1, p.123-128, 1996.

HADFIELD, J.; ELDRIDGE, M. D. Multi-genome alignment for quality control and contamination screening of next-generation sequencing data. **Frontiers in genetics**, Lausanne,v.5, artigo 31**,** 2014.

HAMILTON, J.P.; BUELL, C.R. Advances in plant genome sequencing. **Plant Journal**, Malden, v.70, p.177–190, 2012.

HIGGIS, D.; THOMPSON, J.; GIBSON T. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequences weighting, position-specific gap penalties and weight matrix choice. **Nucleic Acids Research,** London, v.22, p.4673-4680, 1994.

HOMER, N.; MERRIMAN, B.; NELSON, S. F. BFAST: an alignment tool for large scale genome resequencing. **PLoS One**, New York, v.4, n.11, e7767, 2009.

HUANG, X. et al. High-throughput genotyping by whole-genome resequencing. **Genome Research**, Cold Spring Harbor, v.19, p.1068-1076, 2009.

HUANG, X. H. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature Genetics**, London, v. 42 p. 961-976, 2010.

INTERNATIONAL RICE GENOME SEQUENCING PROJECT. The map-based sequence of the rice genome. **Nature**, London, v.1, n.7052, p.793-800, 2005.

IRGA. **Relatório final de colheita do arroz irrigado no Rio Grande do Sul** - Safra 2009/10. Porto Alegre: Irga, 2010. v. 1, 6p.

ISHIKAWA, R. et al. Allelic interaction at seed-shattering loci in the genetic backgrounds of wild and cultivated rice species. **Genes and Genetic Systems**, Shizuoka, v.85, p.265-271, 2010.

JENA, K. K. The species of the genus Oryza and transfer of useful genes from wild species into cultivated rice, O. sativa. **Breeding Science**, Tokyo, v. 60, n.5, p. 518-523, 2010.

JI, H. S. et al. Characterization and mapping of a shattering mutant in rice that corresponds to a block of domestication genes. **Genetics**, Baltimore, v. 173, n. 2, p. 995-1005, 2006.

JI, H. et al. Inactivation of the CTD phosphatase-like gene OsCPL1 enhances the development of the abscission layer and seed shattering in rice. **Plant Journal**, Oxford, v. 61, n. 1, p. 96-106, 2010.

KASIANOWICZ, J.J. et al. Characterization of individual polynucleotide molecules using a membrane channel". **Proceedings of the National Academy of Sciences of the United States of America**, New York, v. 93 n. 24 p. 13770–13773, 1996.

KE, X.; TAYLOR, M.S.; CARDON, L.R. Singleton SNPs in the human genome and implications for genome-wide association studies. **European Journal of Human Genetic,** Basel, v.16, n.4, p.506-515, 2008.

KHUSH, G. S. Origin, dispersal, cultivation and variation of rice. **Plant Molecular Biology**, Dordrecht, v. 35, n. 1-2, p. 25-34, 1997.

KOBAYASHI, A. Varietal adaptability for mechanized rice cultivation: direct seeding adaptability, shattering habit, smoothness. **Journal of Agricultural Science**, Cambridge, v.45 p.186-18, 1990.

KONISHI, S. et al. An SNP caused loss of seed shattering during rice domestication. **Science**, Washington, v. 312, n. 5778, p. 1392-1396, 2006.

KUMAR, S.; STECHER G.; TAMURA K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. **Molecular Biology and Evolution**, [Chicago],2015. (submitted)

LANDER, E. S. et al. Initial sequencing and analysis of the human genome. **Nature**, London, v.409, p.860-921, 2001.

LANDER, E. S. et al. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. **Genomics**, Berlin, v.1, n.2, p.174-81, 1987.

LANGMEAD, B. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. **Genome Biology**, London, v.10, p.25, 2009.

LI, C. B.; ZHOU, A. L.; SANG, T. Genetic analysis of rice domestication syndrome with the wild annual species, Oryza nivara. **New Phytologist**, Cambridge, v. 170, n. 1, p. 185-193, 2006.

LI, C. B.; ZHOU, A. L.; SANG, T. Rice domestication by reducing shattering. **Science**, Washington, v. 311, n. 5769, p. 1936-1939, 2006.

LIN, Z. W. et al. Origin of seed shattering in rice (Oryza sativa L.). **Planta**, Berlin, v. 226, n. 1, p. 11-20, 2007.

LI, H. et al. The sequence alignment/map format and SAMtools. **Bioinformatics**, Oxford, v. 25, n. 16, p. 2078-2079, 2009.

LOMAN, N. J. et al. Performance comparison of benchtop high-throughput sequencing platforms. **Nature Biotechnology**, London, v. 30, n. 5, p. 434-439, 2012.

MARCHESAN, E. et al. Controle do arroz-vermelho. In: GOMES, A.S.; MAGALHÃES JÚNIOR, A.M. (Ed.). **Arroz irrigado no Sul do Brasil**. Brasília: Embrapa Informação, 2004. p. 547-577

MARDIS, E. R. The impact of next-generation sequencing technology on genetics. **Trends in Genetics**, Amsterdam, v. 24, n. 3, p. 133-141, 2008.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, London, v. 437, n. 7057, p. 376-380, 2005.

MARKUS, C. **Caracterizacao e padrao de expressao genica relacionados ao degrane e dormencia de sementes em arroz vermelho**. 2013. 131 f. Dissertação (Mestrado) - Programa de Pós-Graduação em Fitotecnia, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

MARTINS, E. R. et al. Distribution of pilus islands in Streptococcus agalactiae causing human infections: insights into evolution and implication for vaccine development. **CVI**: **Clinical and Vaccine Immunology**, Washington, v.20, n.2, p.313-316, 2012.

MAXAM, A.M.; GILBERT, W. A new method for sequencing DNA. **Proceedings of the National Academy of the United States of America**, Washington, v.74, n.2, p.560–564, 1977.

MESSING J. et al. Sequence composition and genome organization of maize**. Proceedings of the National Academy Sciences of the United States of theAmerica**, Washington, v.101, p.14349–14354, 2004.

METZKER, M. L. Sequencing technologies - the next generation. **Nature Reviews Genetics**, London, v. 11, n. 1, p. 31-46, 2010.

MILLER, J.R.; KOREN, S.; SUTTON, G. Assembly algorithms for next-generation sequencing data. **Genomics**, Berlin, v. 95, n. 6, p. 315, 2010.

MINOCHE A. E.; DOHM J. C.; HIMMELBAUER, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. **Genome Biology**, London, v.12, n.11, R112, 2011.

MOLINA, J. et al. Molecular evidence for a single evolutionary origin of domesticated rice. **Proceedings of the National Academy of Sciences of the United States of America**, New York, v. 108, n. 20, p. 8351-8356, 2011.

MURRAY, M.G.; PETERS, D.L.; THOMPSON, W.F. Ancient repeated sequences in the pea and mung bean genomes and implications for genome evolution. **Journal of Molecular Evolution**, New York, v.17, n.1, p.31–42, 1981.

NAGAI, Y. S. et al. Sh3, a gene for seed shattering, commonly found in African in wild rices. **Rice Genetics Newsletter**, Mishima, v. 19, n. 1, p. 74-75, 2002.

NAGARAJA, N., et al. Genome-wide analysis of repetitive elements in papaya. **Tropical Plant Biology**, New York, v.1, n. 3-4, p.191–201, 2008.

NOLDIN, J.A.; CHANDLER, J.M.; MCCAULEY, G.N. Red rice (Oryza sativa) biology. I. Characterization of red rice ecotypes. **Weed Technology**, Champaign, v.13, p.12-18, 1999.

NUNES, A.L. et al. Avaliação da expressão de genes relacionados ao degrane em arroz através de PCR em tempo real. In: CONGRESSO BRASILEIRO DA CIÊNCIA DAS PLANTAS DANINHAS, 27., 2010, Ribeirão Preto. **Anais do...** Londrina: SBCPD, 2010. p. 426-430

NUNES, A. L. **Variabilidade genética de características associadas ao degrane em arroz vermelho**. 2012. 128 f. Tese (Doutorado) - Programa de Pós Graduação em Fitotecnia, Faculdade de Agronomia, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

PANTONE, D.J.; BAKER, J.B. Reciprocal yield analysis of red rice (Oryza sativa) competition in cultivated rice. **Weed Science**, Champaign, v.39, p.42-47, 1991.

PATTERSON, S. E. Cutting loose. Abscission and dehiscence in Arabidopsis. **Plant Physiology**, Rockville, v. 126, n. 2, p. 494-500, 2001.

PASZKIEWICZ, K., STUDHOLME, D.J. De novo assembly of short sequence reads. **Briefings in Bioinformatics**, London, v. 2, n. 5, p. 457-472, 2010.

PASZKIEWICZ, K., STUDHOLME, D.J. High-Throughput Sequencing data analysis software: current state and future developments. In: RODRÍGUEZ-EZPELETA, N.; HACKENBERG, M.; ARANSAY, A.M. (Ed.). **Bioinformatics for High Throughput Sequencing**. New York: Springer, 2012. p. 231- 48

POP, M. Genome assembly reborn: recent computational challenges. **Briefings in Bioinformatics**, London, v.10, n. 4, p. 354-366, 2009.

ROBERTS, J. A.; ELLIOTT, K. A.; GONZALEZ-CARRANZA, Z. H. Abscission, dehiscence, and other cell separation processes. **Annual Review of Plant Biology**, Palo Alto, v. 53, n. 1, p. 131-158, 2002.

ROBERTS, J. A. et al. Cell separation processes in plants - models, mechanisms and manipulation. **Annals of Botany**, Oxford, v. 86, n. 2, p. 223- 235, 2000.

ROUNSLEY, S.D.; LAST, R.L. Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. **Plant Journal**, Oxford, v.61 p.922-927, 2010.

SANGER, F. et al. DNA sequencing with chain- terminating inhibitors. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v.74, p.5463–5467, 1997.

SANMIGUEL, P. et al. Nested retrotransposons in the intergenic regions of the maize genome. **Science**, London, v. 274, n. 5288, p. 765-768, 1996.

SCHMUTZ, J. et al. Genome sequence of the paleopolyploid soybean. **Nature**, London, v. 463 p. 178-183, 2010.

SCHNABLE, P.S. et al. The B73 Maize Genome: Complexity, diversity, and dynamics. **Science**, London, v.20, p.1112-1115, 2010.

SCHWANKE, A.M.L. et al. Caracterização morfológica de ecótipos de arroz daninho (Oryza sativa) provenientes de áreas de arroz irrigado. **Planta Daninha**, Viçosa, v.26, p. 249-260, 2008.

SCHATZ, M.C. et al. Assembly of large genomes using second-generation sequencing. **Genome Research**, Cold Spring Harbor, v.20, n.9, p.1165–1173, 2010.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature Biotechnology**, London, v.26, p.1135–1145, 2008.

SHUICHI, F.; KAWORU, E.; TOSHIO, Y.; ET AL. Integration of genomics into rice breeding. **Rice**, London, v. 3, p. 131-137, 2010.

SCHUSTER S.C. Next-generation sequencing transforms today's biology. **Nature Methods**. New York, v.5, n.1, p.16-18, 2008.

SIMPSON, J.T., et al. ABySS: a parallel assembler for short read sequence data. **Genome Research**, Cold Spring Harbor, v.19, p.1117–1123, 2009.

SMITH, C. W.; DILDAY, R. H. **Rice:** origin, history, technology, and production. Hoboken: John Wiley & Sons, 2003.

SOCIEDADE BRASILEIRA DE ARROZ IRRIGADO. **Arroz irrigado:** recomendações técnicas da pesquisa para o Sul do Brasil. Pelotas, 2010. 188 p.

SOCIEDADE BRASILEIRA DE ARROZ IRRIGADO. **Arroz irrigado**: recomendações técnicas da pesquisa para o Sul do Brasil. Itajaí, 2012. 177 p.

SOBRIZAL, K. et al. RFLP mapping of seed shattering gene on chromosome 4 in rices. **Rice Genetics Newsletter**, Mishima, v. 16, n. 1, p. 74-75, 1999.

SMITH, C. W.; DILDAY, R. H. **Rice:** origin, history, technology, and production. Hoboken: John Wiley & Sons, 2003. 627 p.

TAIZ, L.; ZEIGER, E. **Fisiologia vegetal**. 5. ed. Porto Alegre: Artmed, 2013. 954 p.

THE FRENCH-ITALIAN PUBLIC CONSORTIUM FOR GRAPEVINE GENOME CHARACTERIZATION. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. **Nature**, London, v.449, p. 463-467, 2007.

THURBER, C. S.; HEPLER, P. K.; CAICEDO, A. L. Timing is everything: early degradation of abscission layer is associated with increased seed shattering in U.S. weedy rice. **BMC Plant Biology**, London, v. 11, n. 14, p. 1-10, 2011.

THURBER, C.S. et al. Molecular evolution of shattering loci in U.S. weedy rice. **Molecular Ecology**, Oxford, v.19, p.3271-3284, 2010.

VENTER, J. C. et al. The sequence of the human genome. **Science**, New York, v. 291, n.5507, p.1304-1351, 2001.

WANG M et al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. **Nature Genetics**, London, v. 46, n.9, p. 982-991, 2014.

WEBER, A.P. et al. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. **Plant Physiology**, Minneapolis, v.144, p.32-42, 2007.

XIONG, L. X. et al. Identification of genetic factors controlling domestication-related traits of rice using an F-2 population of a cross between *Oryza sativa* and *O. rufipogon*. **Theoretical and Applied Genetics,** New York, v.98 p. 243-251, 1999.

YANG Q.; SUK-MAN, K.; XINHUA, Z. Identification for quantitative trait loci controlling grain shattering in rice. **Genes and Genetic Systems**, Shizuoka, v.32 p.173-180, 2010.

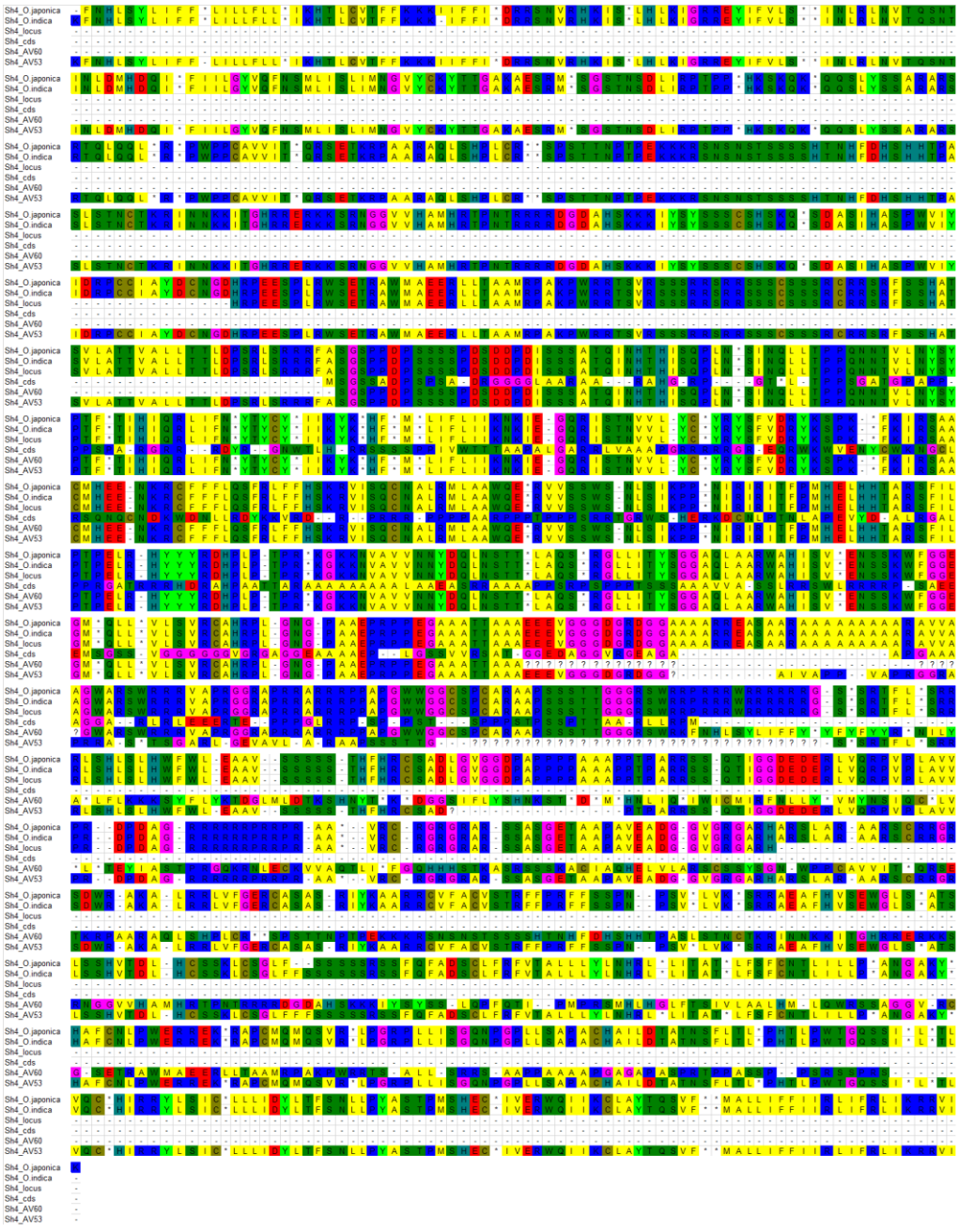
ZHU, Y. Q.; ELLSTRAND, N. C.; LU, B. R. Sequence polymorphisms in wild, weedy, and cultivated rice suggest seed-shattering locus Sh4 played a minor rolein Asian rice domestication. **Ecology and Evolution**, London, v. 2, n. 9, p. 2106 - 2113, 2012.

ZHANG, Z. H. et al. Mapping quantitative trait loci (QTLs) for seedling-vigour using recombinant inbred lines of rice (Oryza sativa L.). **Field Crops Research,** Oxford, v. 91, n.3, p. 161–170, 2004.

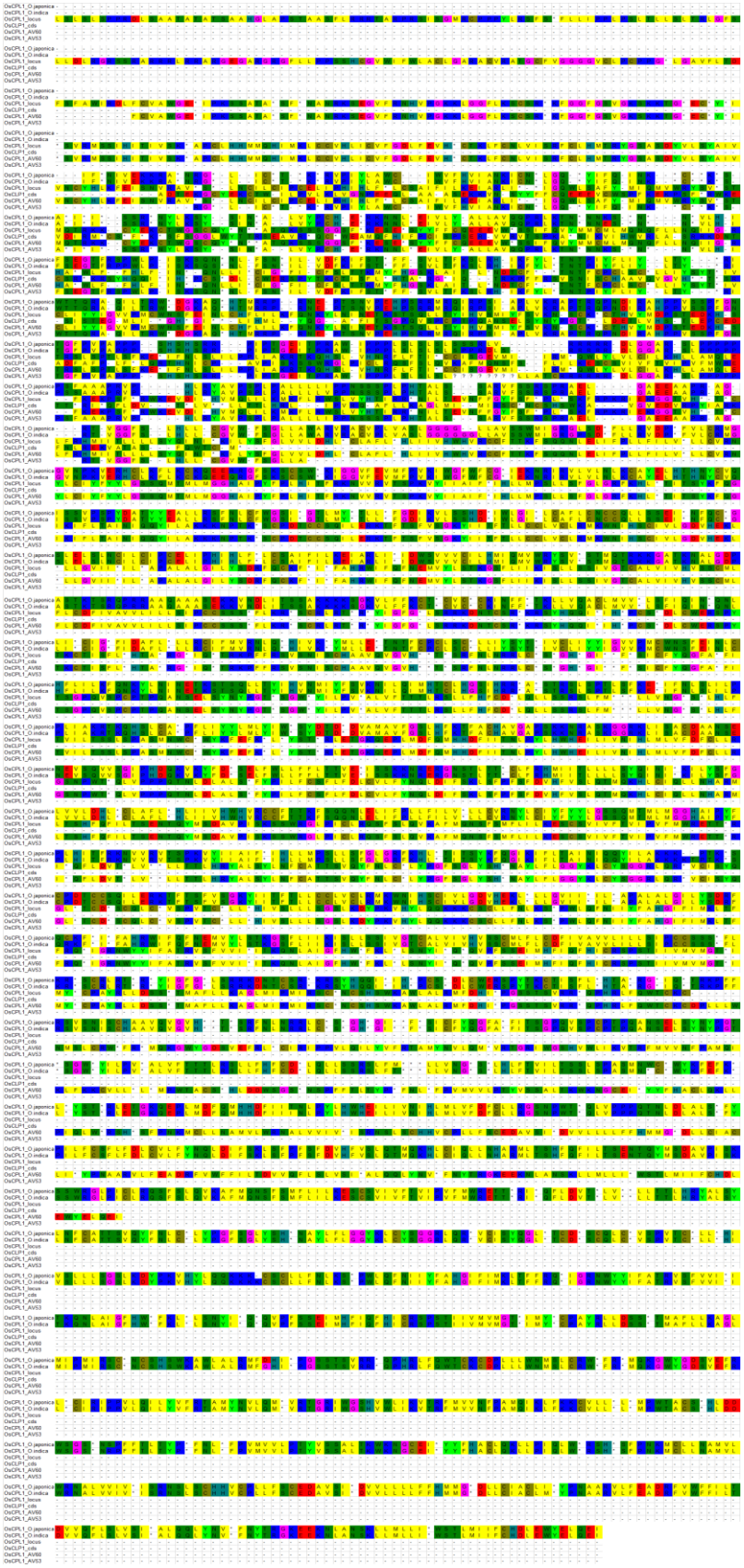
ZHANG, L. B. et al. Selection on grain shattering genes and rates of rice domestication. **New Phytologist**, Malden, v. 184, n. 3, p. 708-720, 2009.

7 APÊNDICES

APÊNDICE 1. Alinhamento de proteínas resultantes da tradução do gene de graneSh4provenientes dos acessos de arroz vermelho AV53 e AV60, Oryza indica, Oryza japonica, locus gênico e região codificante (cds). Indels são representados por "-" e dados faltando "?" e regiões conservadas por "*".Porto Alegre, 2015.

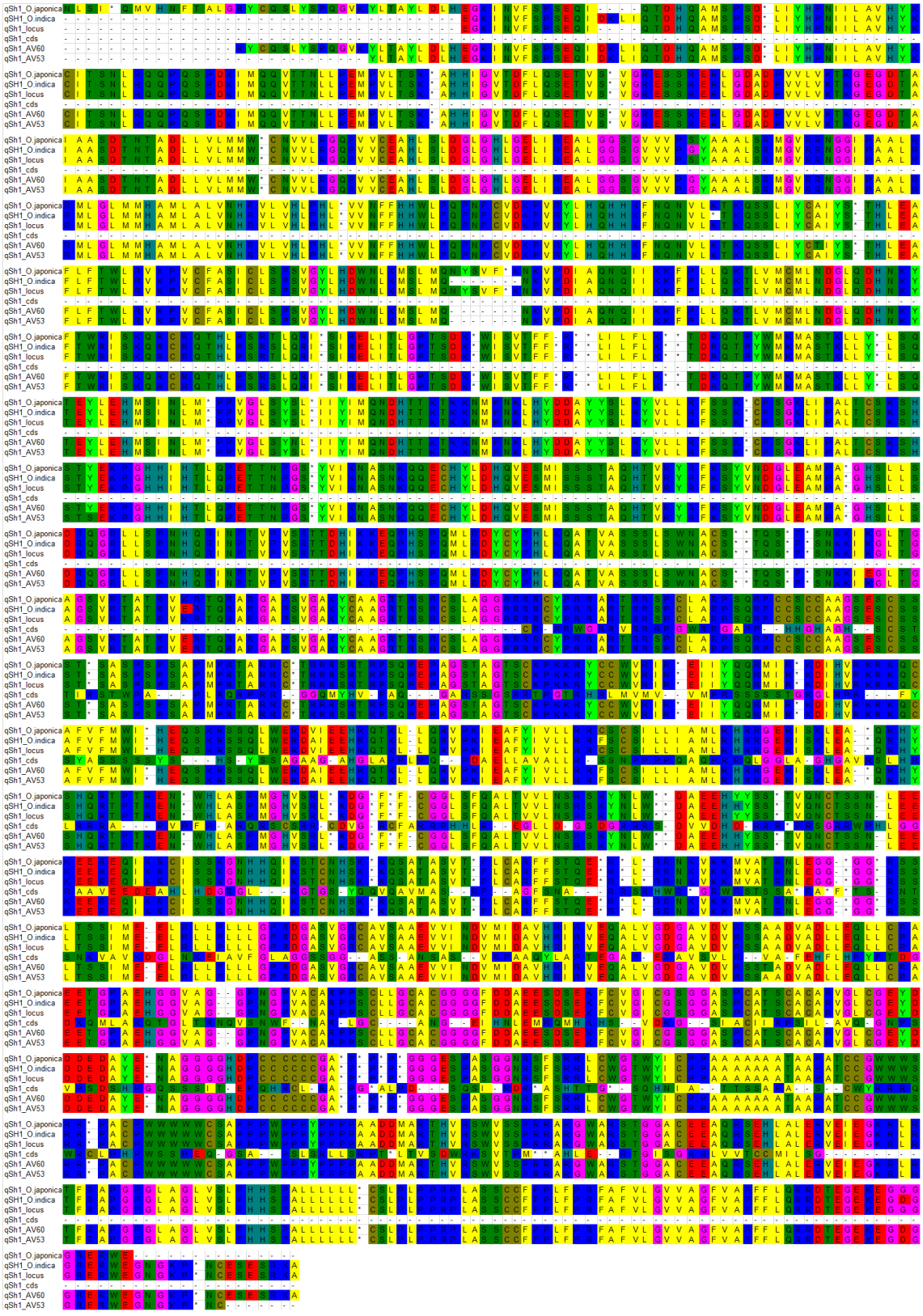


APÊNDICE 2. Alinhamento de proteínas resultantes da tradução do gene de degrane OsCPL1provenientes dos acessos de arroz vermelho AV53 e AV60, Oryza indica, Oryza japonica, locus gênico e região codificante (cds). Indels são representados por "-" e dados faltando "?" e regiões conservadas por "*". Porto Alegre, 2015.

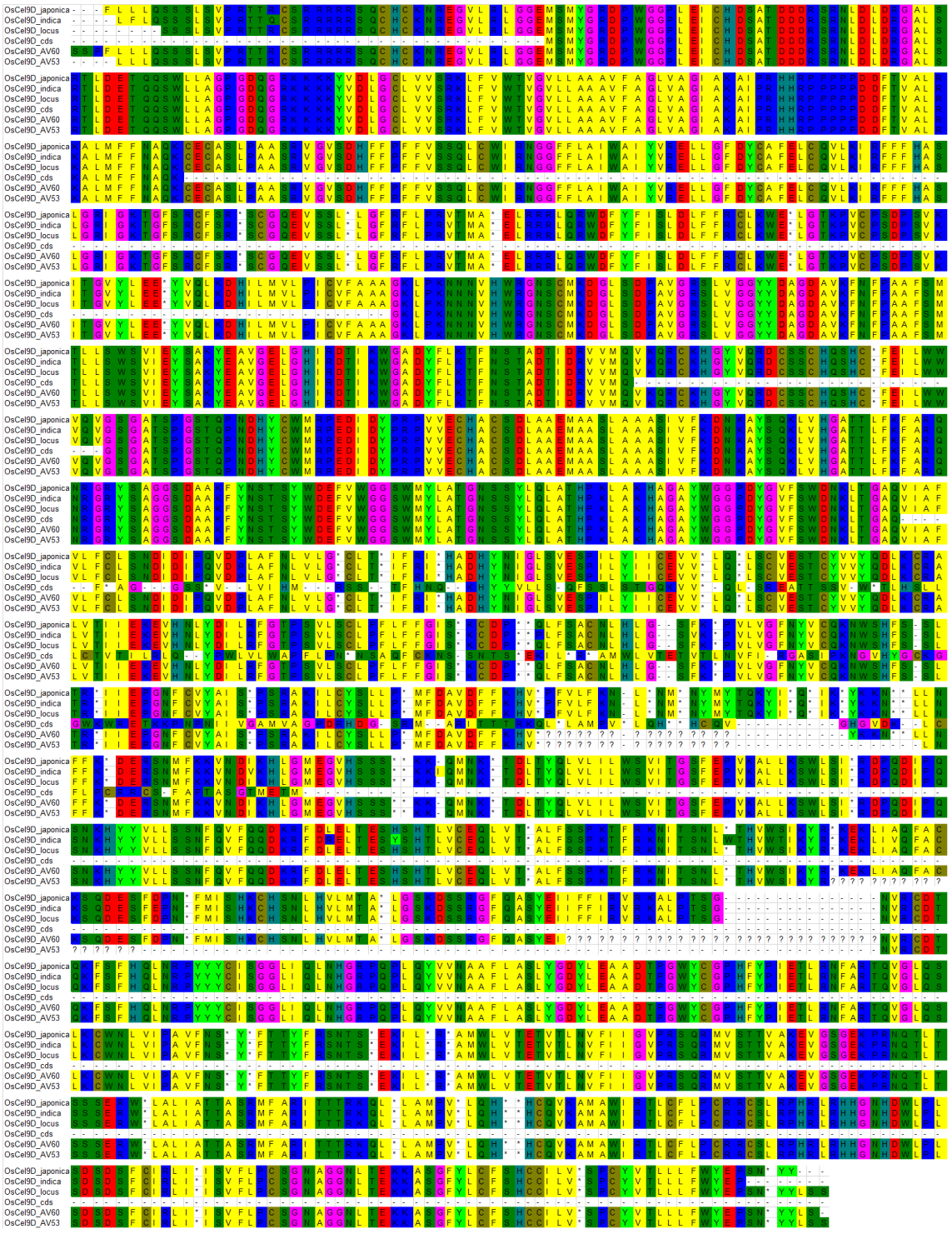


APÊNDICE 3. Alinhamento de proteínas resultantes da tradução do gene de degrane Os01g0849100provenientes dos acessos de arroz vermelho AV53 e AV60, Oryza indica, Oryza japonica, locus gênico e região codificante (cds). Indels são representados por "-" e dados faltando "?" e regiões conservadas por "*". Porto Alegre, 2015.

APÊNDICE 4. Alinhamento de proteínas resultantes da tradução do gene de degrane qSH1provenientes dos acessos de arroz vermelho AV53 e AV60, Oryza indica, Oryza japonica, locus gênico e região codificante (cds). Indels são representados por "-" e dados faltando "?" e regiões conservadas por "*". Porto Alegre, 2015.



APÊNDICE 5. Alinhamento de proteínas resultantes da tradução do gene de degrane OsCeI9D provenientes dos acessos de arroz vermelho AV53 e AV60, Oryza indica, Oryza japonica, locus gênico e região codificante (cds). Indels são representados por "-" e dados faltando "?" e regiões conservadas por "*". Porto Alegre, 2015.



APÊNDICE 6. Alinhamento de proteínas resultantes da tradução do gene de degrane *OsxTH8* provenientes dos acessos de arroz vermelho AV53 e AV60, *Oryza indica*, *Oryza japonica*, locus gênico e região codificante (cds). Indels são representados por "-" e dados faltando "?" e regiões conservadas por "*". Porto Alegre, 2015.