Development of genetic management methods for rice varieties and prediction of spontaneous mutation frequencies using next-generation sequencing technology

Hokkaido University, Graduate School of Agriculture
Frontiers in Production Sciences Doctor Course

A summary of thesis submitted to the Graduate School of Agriculture,
Hokkaido University, Sapporo, Japan in partial fulfillment of the requirements for
the award of the degree of Doctor of Philosophy (PhD) in Agrobiology
(Laboratory of Plant Breeding)

BALIMPONYA Elias George

June, 2023

Background

Plant breeding programs aim to improve crop traits for the benefit of farmers and consumers, particularly in light of the challenges posed by climate change and a growing global population. Neglecting varietal quality control can lead to the deterioration of crop quality due to spontaneous mutations. Understanding these mutations is crucial for identifying desirable traits and enhancing plant adaptation and genetic diversity. Spontaneous mutations arise from various factors, such as DNA repair mechanisms, mutagens, and genetic background, contributing to genetic diversity and adaptation. However, some mutations can be detrimental, impacting viability and uniformity in crop populations.

Spontaneous mutations in eukaryotes often occur on one of a pair of sister chromatids, resulting in a heterozygous state of mutation (HSM). Spontaneous mutation rates are relatively consistent across different plant species and populations, although various factors can influence them. Some mutations are non-heritable and have minimal effects on plant populations, while others are inheritable, either recessive or dominant, with varying impacts. Spontaneous genome mutations can lead to both beneficial or deleterious trait changes. Accurate data on spontaneous mutations is essential for efficient genetic improvement, conservation of genetic resources, and targeted breeding programs. For example, mutations in maize resulted in a brown midrib phenotype, affecting livestock digestibility but reducing forage and grain yields. Accurately detecting spontaneous mutations has been challenging due to factors like inaccurate tools and sequencing errors. Advances in next-generation sequencing (NGS) technology and bioinformatics tools have improved mutation detection. However, challenges like repetitive genome regions and sequencing errors still exist, affecting variant calling accuracy. Spontaneous mutations can impact crop quality by affecting viability, germination, genetic integrity, and nutritional composition. These mutations can lead to reduced seed capability, abnormal growth patterns, changes in grain composition, increased susceptibility to pests and diseases, and altered adaptation to environmental conditions.

Chapter 2 of the thesis discusses a study on removing spontaneous mutations from the population to maintain seed quality, using albinism in rice as a case study. I propose an NGS-WGS approach to remove deleterious alleles within a population in a single season. Chapter 3 explores the rate of spontaneous heterozygous mutations occurring per nucleotide per individual plant of rice and how they are inherited. The study includes various approaches to identify false positive callings, determining the rate of HSMs in the rice genome, and their genomic effects on protein synthesis, as well as the discovery of novel HSMs in different rice cultivars.

IDENTIFICATION AND REMOVAL OF SPONTANEOUS ALBINISM MUTATION IN RICE BREEDER SEEDS
(CHAPTER 2)

Introduction

Breeder seeds play a crucial role in ensuring the genetic purity and quality of subsequent generations of crops. Maintaining the Distinctness, Uniformity, and Stability (DUS) criteria is essential in plant breeding programs. However, breeder seeds can be susceptible to genetic mutations, such as albinism in rice, which can significantly impact the quality of seeds. This chapter aims to identify and remove deleterious mutations, particularly albinism, from breeder seeds to ensure seed quality. Molecular fingerprinting methods have long been used to identify genetic variations, but none of these methods can rapidly remove deleterious mutations in a single growing season. The SWL1 gene responsible for albinism in the Hinohikari rice variety was identified (Balimponya *et al.* 2022). The study recommends techniques for genotyping albinos, specifically those caused by the SWL1 gene interference, and proposes cost-effective procedures for removing hidden deleterious mutations from breeder seeds.

Methodology

The study used the Hinohikari rice variety, a progeny of Koganebare/Koshihikari, and its parents belong to the Japonica sub-species. Albino and green seedlings from this population were used for analysis. Seeds from the Hinohikari variety were tested at different locations and years to evaluate the frequency of albinism. The gDNAs were sequenced using the Illumina platform with paired-end reads. Raw sequencing data were processed, filtered, mapped to the rice genome, and analyzed for variants (SNPs and InDels) associated with albinism. The PCR products from the suspected albinism site were sequenced using the Sanger method to validate NGS results.

Results

Albinism frequency in the Hinohikari rice variety remained consistent at 1.36% across different years and locations (**Table 1**). The gDNA quality and sequencing results were sufficient for NGS analysis. NGS analysis identified 756 genomic differences between albino and green plants, with 217 variants associated with albinism. A C-insertion in the *SWL1* gene was identified as the cause of albinism. This mutation resulted in a frameshift and premature stop codon (**Figure 1**). The albinism trait followed a Mendelian pattern of inheritance, confirming the recessive nature of the *swl1-R332P* allele. Albino plants exhibited disrupted chloroplasts, a lack of thylakoid membranes, and abnormal organelles compared to normal green plants.

Table 1 Occurrence of albino of the Hinohikari breeder seeds in MCAES

Materials tested	Albino occurrences (counts)	Total numbers of seeds tested	Incidence (%)	Year and place of testing
Original population	74	5438	1.361	2015 at MCAES*
Testing 1	31	2281	1.359	2019 at HU**
Testing 2	20	1476	1.355	2020 at HU
Total	125	9195	1.359	

* MCAES: Miyazaki Comprehensive Agricultural Experimental Station, Miyazaki Japan.

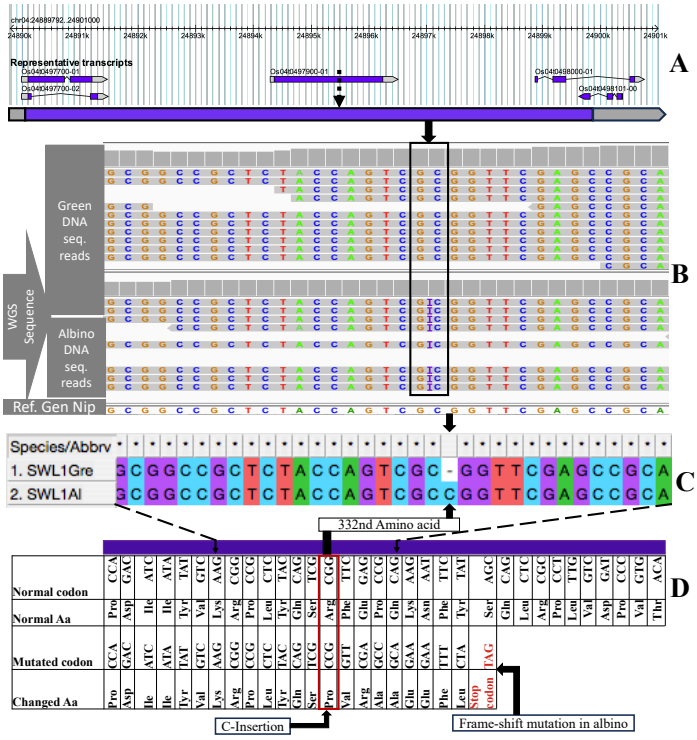

Figure 1 Identification of C-Insertion in the *SWL1* locus of albino mutants.
(A) The *SWL1* locus on Chr. 4. **(B)** WGS-NGS analysis of the genomic differences between pooled green plants and pooled albino plants, where the pooled albino sample contains the C-insertion in all mapped reads. **(C)** Sanger sequencing results confirming C-insertion in the pooled albino sample. **(D)** The effects of C-insertion, whereby the 332nd amino acid arginine is changed to proline, resulting in a premature TAG stop codon.

Discussion

The swl1-R332P allele responsible for albinism is recessive, making it challenging to eliminate from the population due to its presence in heterozygous individuals. The NGS analysis allowed for the rapid identification of the causative gene within a year, enabling the swift removal of the mutant allele from the population. The study highlights the importance of identifying and removing deleterious mutations to maintain seed purity and ensure seed quality. The use of WGS in identifying mutant genes offers a more efficient method for managing breeder seeds, allowing for the rapid removal of undesirable mutations.

Conclusion

The recessive nature of the mutation presents challenges in seed purity maintenance, but the rapid identification of the mutant allele allows for its removal from the population. This research underscores the importance of advanced genomic techniques in ensuring seed quality and purity in plant breeding programs. Surprisingly, 217 genomic variants were detected, including the SWL1 mutation. This suggests the presence of other unrelated mutations in the population and directed us to the chapter 3.

ACCURATE DETECTIONS OF DE-NOVO SPONTANEOUS MUTATIONS IN THE RICE GENOME (CHAPTER 3)

Introduction

The research focuses on detecting de-novo spontaneous mutations in the rice genome, particularly heterozygous state of mutations (HSMs). Spontaneous genome mutations play a crucial role in trait alteration and phenotypic expression in plants. While homozygous mutations are easier to identify, tracking and managing heterozygous mutations, particularly those controlled by recessive alleles, presents challenges. The study addresses the scarcity of research on heterozygous inheritance modes and their frequencies in plants. There have been several examples of phenotypic changes attributed to heterozygous mutations including albinism in Hinohikari due to a C-insertion in the SWL1 gene (Balimponya *et al.* 2022), gold color mutation in onions (Kim *et al*. 2004), and altered tomato shape due to mutations in the OVATE QTL (Liu *et al*. 2002). The mutation rates vary among species, with mammals accumulating mutations at 2.22×10^{-9} substitutions per site per year, while *Heliconius melpomene* exhibits a rate of 2.9×10^{-9}(Keightley *et al*. 2015).

The study aims to determine the rate of HSMs in the rice genome, map their locations, and assess their genomic effects on protein synthesis. Overall, the research contributes valuable insights into spontaneous mutations in the rice genome, particularly heterozygous mutations, and their potential implications for crop development and trait alteration.

Methodology

The methodology of the study involves the analysis of spontaneous de-novo mutations in the rice genome, focusing on heterozygous mutations (HSMs). Two approaches are used, and various plant materials are employed, including Nipponbare siblings (Nip1, Nip2, Nip3), Kitaake siblings (Kita1, Kita2), and progenies from Hinohikari (Parent, Progeny 1, Progeny 2, Progeny 3).

Initial Approach: Involves Nipponbare and Kitaake siblings. Two identical copies of Nip1 and three identical copies of Kita1 are used for assessing mapping and variant calling accuracy. A secondary reference genome is constructed to detect de-novo mutations (**Figure 2**). Sanger sequencing is used for validation. In an alternate Approach, I utilize a parent plant and its three progenies obtained from self-crossing. WGS is performed on these individuals using the Nipponbare reference genome, and the same analysis procedures as the initial approach are applied. In the Identification of the best pair of mapping tools and variant calling pipeline, Two mapping tools (bowtie2 and BWA-MEM 2) and variant calling pipelines (GATK4 HaplotypeCaller and BCFtools mpileup) validated by CLC Genomics Workbench are compared to identify authentic heterozygous variants. Specific filtration criteria are used to select the best combination of tools and pipelines. In estimating the sequencing, mapping, and variant calling errors, Errors are classified into three categories: sequencing, mapping, and variant calling errors based on the findings from the previous step.

Cases	1	2	3	4	5	6	7	8
Reference	**A**	**T**	**C**	**G**	**A**	**C**	**G**	**G**
Copy1	**A**	**T**	**C**	**G/C**	**AT**	**-**	**T**	**C**
Copy2	**A**	**A**	**TC**	**G/C**	**A**	**-**	**GA**	**T**
Conse Sec. Ref	**A**	**T**	**C**	**G/C**	**A**	**-**	*	*

A

Cases	1	2	3	4	5	6	7	8
Copy1	G	A	C	AG	T/C	A/G	G	T
Copy2	G	A	CA	AG	T/C	A/G	A	N
Copy3	G	G	C	A	T/C	G	T	-
Conse. Sec. Ref	**G**	**A**	**C**	**AG**	**T/C**	**A/G**	*	*

B

Figure 2 The hypothetical edited rice genome that gives the consensus reference genome. **(A)** involves the same two copies and references. **(B)** involves three copies of same gDNA. The consensus secondary reference genome was created by CLC genomic workbench by considering the common nucleotide among the three sets and was used as a reference for mapping the rest of the siblings. The sign * represents positions in the consensus genome where no nucleotide was assigned and the sign – represents deletion(s).

Two cases are considered, involving Nipponbare and Kitaake. Consensus secondary reference genomes are constructed from multiple identical gDNA samples, incorporating positions corrected for errors. The secondary reference genome is used for mapping. In the detection of HSMs in

Nipponbare and Kitaake, in the case of Nipponbare siblings, five repetitions of mapping and variant calling are performed, and only consistent heterozygous variants are retained. CLC Genomics Workbench is used for validation, and variants are visually inspected. For Kitaake siblings, the same approach is applied, and common heterozygous variants and those exhibiting a segregation pattern are considered.To validate heterozygous variants, genomic segments containing these variants are amplified and subjected to Sanger sequencing.

In the mapping and variant calling process for the parent and its progenies, the parent and three progenies are mapped to the Nipponbare reference genome, and variants are called. Five repetitions are conducted, and only common heterozygous variants are selected, validated using IGV and CLC Genomic Workbench. Unique heterozygous variants in progenies are considered as novel genetic variations.

Results

Optimizing Mapping Tools and Assessing Errors:

Four different sets of mapping tools and variant calling pipelines were tested, with BWA-MEM2 combined with GATK4 HaplotypeCaller proving to be the most accurate, achieving similarity rates of around 99% across multiple experiments. In contrast, BCFtools mpileup demonstrated poorer performance, with similarity rates as low as 66.5%. Furthermore, BCFtools mpileup exhibited a significant limitation in detecting known heterozygous variants in some samples, highlighting the need for caution when using this variant calling pipeline.

Detection of Heterozygous State of Mutations (HSMs)

By combining the results of multiple independent analyses, the the number of false positives was reduced. The calculated rate of HSMs per nucleotide in the rice genome was found to be relatively high, consistent across different samples and methodologies. HSMs were observed to be randomly distributed across all 12 chromosomes of the rice genome, with no specific pattern or genomic regions associated with their presence. The majority of HSMs were located outside of genic regions, but some were found within exons and had the potential to impact protein synthesis. The study extended its analysis to include a parent plant and its progenies. The HSM rates per nucleotide in the progenies were found to resemble those in Nipponbare and Kitaake, indicating similar efficiency in detecting HSMs across different approaches.

Mapping and Variant Calling Errors and Sanger Sequencing Validation.

Sequencing errors, variant calling errors, and mapping errors contributed to false positive mutations and accounted for a significant proportion of identified heterozygous variants. Sanger

sequencing was employed to validate identified heterozygous variants. Out of the sequenced positions, a high percentage were confirmed as true heterozygous variants, aligning with the results obtained through computational methods. A heterozygous was confirmed after showing a clear two peaks of climatographic (**Figure 3**).

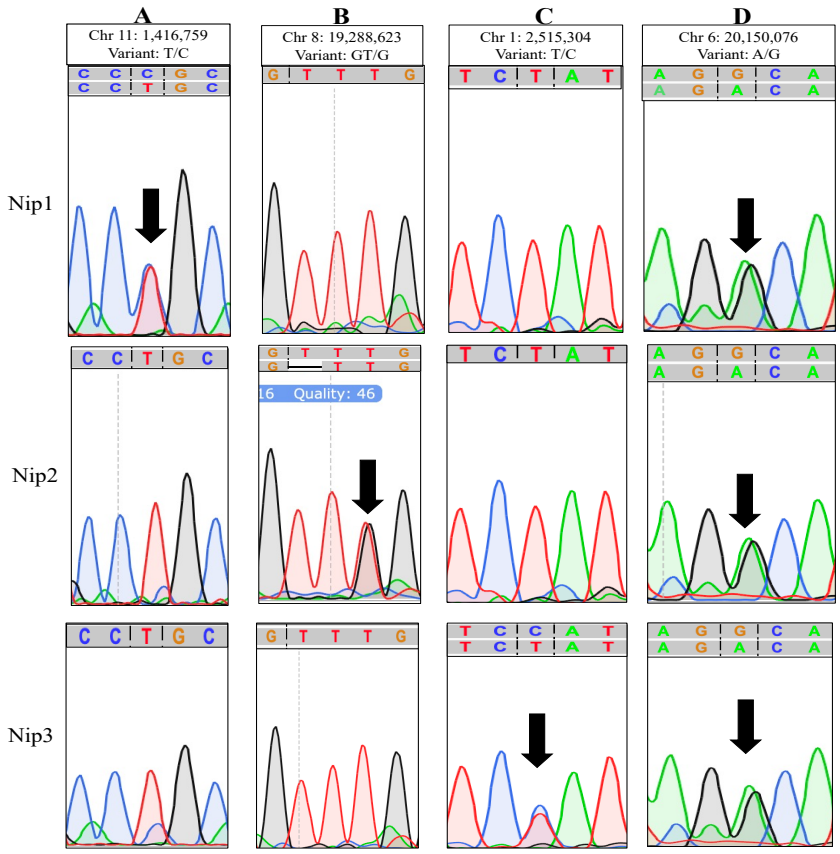

Figure 3 The Sanger sequencing results for four cases of Nipponbare.

(A) the case where Nip1 had unique HSM as compared to Nip2 and Nip3. **(B)** the case where Nip2 had unique HSM as compared to Nip1 and Nip3. **(C)** the case where Nip3 had unique HSM as compared to Nip1 and Nip2. **(D)** the case of having the common HSMs among all three samples.

Discussion

Accurate detection of spontaneous mutations is crucial in plant breeding programs. While identifying homozygous dominant mutations with immediate phenotypic effects is relatively straightforward, detecting heterozygous mutations is challenging. Previous studies employed various procedures to estimate spontaneous mutations, such as creating F1 generations or BC2F7 populations. In contrast, this study introduced a novel approach using siblings (plants raised from seeds of the same plant) to detect spontaneous heterozygous mutations. The study successfully

identified unique heterozygous mutations in siblings and validated them using Sanger sequencing and parental data.

The study highlights the importance of using a combination of mapping tools and variant calling pipelines to estimate HSMs accurately. Different tools and pipelines may yield varying results due to differences in alignment and calling algorithms, computational optimization, and tunable parameters (Blackburne and Whelan 2013; Yu and Sun 2013; Medvedev *et al*. 2009; Reumers *et al*. 2012). Using multiple tools helps identify a more comprehensive set of HSMs. The calculated HSM rate in rice is consistent with rates observed in other plants with similar genome sizes. For example, *Arabidopsis thaliana* and maize exhibited mutation rates in a comparable range (Yang *et al*. 2015; Yang *et al*. 2017). This suggests that the HSM rate observed in rice is within the expected range for plants with genomes of similar sizes. Spontaneous mutations can affect various parts of the genome, including regulatory elements, genes, and coding sequences. Mutations in regulatory elements may impact gene expression, while those in genes can affect mRNA splicing, stability, and translation. Non-synonymous mutations can modify protein activity, while synonymous mutations are functionally neutral. Mutations causing premature stop codons can result in truncated or non-functional proteins (Robert and Pelletier 2018). The study identified HSMs in both genic and non-genic regions of the rice genome, with potential effects on gene expression and protein synthesis.

Conclusion

This study provides valuable insights into the detection of spontaneous mutations in rice genomes. The novel approach of using siblings for mutation detection, combined with a rigorous validation process, yielded an average HSM rate per nucleotide per generation. These findings have implications for plant breeding programs and contribute to our understanding of mutation rates in plant genomes.

References:

Balimponya, E.G., M.S. Dwiyanti, T. Ito, S. Sakaguchi, K. Yamamori, Y. Kanaoka, Y. Koide, Y. Nagayoshi and Y. Kishima (2022). Seed management using NGS technology to rapidly eliminate a deleterious allele from rice breeder seeds. Breeding Science 72: 362 – 371.

Blackburne, B.P. and S. Whelan (2013). Class of multiple sequence alignment algorithm affects genomic analysis. Mol Biol Evol 30:642-653.

Keightley, P.D, A. Pinharanda, R.W. Ness, F. Simpson, K.K. Dasmahapatra, J. Mallet, J.W. Davey and C.D. Jiggins (2015). Estimation of the spontaneous mutation rate in Heliconius melpomene. Mol. Biol. Evol 32:239 – 243.

Kim, S., R. Jones, K. S. Yoo and L. M. Pike (2004). Gold color in onions (Allium cepa): A natural mutation of the chalcone isomerase gene resulting in a premature stop codon. Mol. Genet. Genom 272: 411–419.

Liu, J., J. Van Eck, B. Cong and S. D. Tanksley (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. Proc. Natl. Acad. Sci. U.S.A 99:13302–13306.

Medvedev, P., M. Stanciu, and M. Brudno (2009). Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods, 6: pp.S13-S20.

Reumers, J., P. De Rijk, H. Zhao, A. Liekens, D. Smeets, J. Cleary, P. Van Loo, M. Van Den Bossche, K. Catthoor, B. Sabbe and E. Despierre (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. Nat. Biotechnol 30:61-68.

Robert, F and J. Pelletier (2018). Exploring the impact of single-nucleotide polymorphisms on translation. Front. Genet 9:1 – 11.

Yang, N., X.W. Xu, R.R. Wang, W.L. Peng, L. Cai, J.M. Song, W. Li, X. Luo, L. Niu, Y. Wang and M. Jin (2017). Contributions of Zea mays subspecies mexicana haplotypes to modern maize. Nat. Commun 8:pp1874.

Yang, S., L. Wang, J. Huang, X. Zhang, Y. Yuan, J. Q. Chen, L. D. Hurst and D. Tian (2015) Parent-progeny sequencing indicates higher mutation rates in heterozygotes. Nature 523: 463–467.

Yu, X. and S. Sun (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinform 14:1-15.