EFFICIENT DISCOVERY OF RARE ALLELES AND *DE NOVO* MUTATIONS FROM PRE-EXISTING GENOMIC DATA

by

ROBERT ADAM ARTHUR

(Under the Direction of Jeffrey Bennetzen)

ABSTRACT

A few methods exist to identify the full spectrum of recent mutations in specific lineages, but all are costly, laborious and slow. We propose a novel strategy that requires only resequencing data and a reference genome sequence that are available at no cost from public databases. The comparison of differences between resequencing shotgun data and overlapping 50mers created *in silico* to represent the complete reference genome allows the discovery of reference-genome-specific *de novo* mutations, rare alleles, and sequencing errors unique to the reference genome. We investigated Nipponbare rice, and discovered thousands of candidate *de novo* sequence changes, of which ~51% are calculated to be events that occurred during the recent descent of this lineage. The remaining 49% were Nipponbare reference genome sequencing errors. Of the 148 validated mutations specific to Nipponbare, we found 143 single nucleotide substitutions, 4 tiny insertions, and 1 tiny deletion. Additionally, we applied our method to the reference genome for foxtail millet, Yugu1. However, the resequencing data for this species was not sufficient to mask ancient standing variation in the progenitors of Yugu1, so the analysis primarily yielded rare alleles and sequencing errors rather than *de novo* mutations. Of 119 confirmed sequence variations unique to Yugu1, we found 66 transitions, 40 transversions, and 13 indels (9 insertions, 4 deletions), all of which were only 1 bp. Surprisingly,

despite very high sensitivity to this type of genome change, we did not detect any recent transposable element activity in the origins of Nipponbare or Yugu1.

INDEX WORDS: whole genome, *de novo* mutation, rare alleles, rice, foxtail millet, mutation, genomics, genetics

EFFICIENT DISCOVERY OF RARE ALLELES AND *DE NOVO* MUTATIONS FROM PRE-EXISTING GENOMIC DATA

by

ROBERT ADAM ARTHUR

B.S., Georgia Institute of Technology, 2009

M.S., Georgia Institute of Technology, 2011

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2017

EFFICIENT DISCOVERY OF RARE ALLELES AND *DE NOVO* MUTATIONS FROM PRE-EXISTING GENOMIC DATA

by

ROBERT ADAM ARTHUR

Major Professor:	Jeffrey Bennetzen
Committee:	David Hall James Leebens-Mack Juan Gutierrez

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

DEDICATION

I dedicate this work to Bertha, my wife and best friend, who has been cheering me on for the whole duration of my graduate school career and has been encouraging, faithfully optimistic, and a great person to talk to at all times. Your encouragement despite hardships, your never-ending support, and your constant pushing me to become the best man and student I can be are why I am able to write this dissertation. I love you, and I am thankful for you every day.

I would also like to dedicate this work to my parents, Robert L. and Rosemary Arthur, who have been instrumental in encouraging (and partially funding) my tenure through post-secondary education. Without you two, I would not have ever had the confidence to attempt my doctorate, and so I thank you for all your devotion, love, and support that you have given me over the years.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

In biology, among the most intriguing subjects are the nature of mutation, how mutations arise and how they affect genomes. A mutation can be defined as a permanent change of the nucleotide sequence within the genome of an organism (1,2). Mutations are primarily responsible for diversity between organisms and within populations of the same species and are the raw materials on which natural selection acts. A mutation can range from a small scale change, to gene duplications or deletions, to chromosomal duplications or rearrangements, to genome duplication. Even the smallest of mutations can have effects on the entire genome. Mutations can either be heritable or not, with heritable mutations being passed onto the next generation and having the potential to become a permanent, fixed mutation in a population. Heritable mutations must be derived from germline cells and thus must arise in tissues that lead to the cell lineages that produce an egg or sperm cell. Mutations in exclusively somatic cells only affect the organism in question and cannot be passed down (2,3), except in organisms (like many plants) that can produce progeny asexually. Evolutionary studies tend to focus on germline mutations as they can be passed down to future generations, and can be traced to their ancestral form, while having the potential to elucidate changes driven by selection. Positive mutations increase fitness in future generations (4-6). Somatic mutations are not without consequences despite their inability to be transmitted to future generations, as diseases often arise from somatic mutations, with the most obvious being cancer (1,7).

Mutations are random, but not all mutations have equal probability to spontaneously arise. The ramifications of a mutation do not have a bearing on the probability for the mutation to

occur (3-6), with positive mutations not being more likely to occur due to their beneficial effects than negative mutations (7,8). Deleterious mutations are much more common than beneficial mutations because there are many more ways to debilitate a gene than there are to improve it. If a positive mutation happened to develop in somatic cells, it will not be passed down to the offspring despite its increase in the organism's fitness and chance at survival due to it not being within the germline. Mutation types may be non-random because some mutations are more likely to occur in a given organism due to its physiology and/or lifestyle (1,2). For instance, green plants generate vast numbers of oxidative molecules that can damage DNA from the process of photosynthesis (1,2), thus leading to an unusually high level of oxidatively-induced mutation types compared to less metabolically active organisms. Mutations are generated by many biological, chemical and physical processes, so they are unavoidable, but mutation rates in general are very low when compared to the size of an entire genome (3,5-6). Organisms minimize their mutation rates by the multiple DNA repair and proofreading mechanisms that exist, but the mutation rate never reaches zero.

Mutations are not inherently negative or positive, per se, because they allow an organism to evolve over time and improve its fitness with the influence of selection in a particular biological context. Changes in environment lead to different selection forces at different times and locations (7). For instance, an allelic variation that is selected against in one environment may be selected for in another environment, such that calling a mutation good or bad is a purely conditional use of terminology.

Even the smallest of mutations, a single base pair change, can have physiological significance. The haploid human genome is estimated to have approximately 3 billion base pairs (9,10), but numerous common disorders arise from the change of just one nucleotide. For

instance, sickle-cell anemia, a disorder common among humans of African descent, can arise from a single mutation in a ß-hemoglobin gene. The gene coding for the beta chain of hemoglobin protein consists of 147 amino acids, but the most common type of sickle-cell anemia in the US originates from the change of a glutamic acid to valine due to a mutation from GAG to GTG. Individuals with sickle-cell anemia have blood cells that tend to clump together and form rods in the sickle shape, hence the name, and do not function up to par in terms of supplying oxygen to the body. Sickle red blood cells are often removed by the spleen, leading to anemia, and can cause clots in blood vessels. Despite this disorder being a type of genetically negative recessive disease, the sickle-cell allele is still maintained at high levels in some human populations because heterozygosity for sickle-cell passes on an increase in fitness due to decreased susceptibility to malaria (9). This example is one of many that illustrates the impact that mutations can make when they occur in coding regions within a genome, and that one seemingly innocuous mutation can have significant impacts on the organism as a whole. However, not all mutations are deleterious, as many can be neutral or "silent" and remain undetected in an organism's phenotype. Any base pair variances in a population are referred to as polymorphisms, but a "common" polymorphism typically describes a situation in which one of two or more sequences are found in at least 1% of the population (11). These common polymorphisms have been useful to human geneticists as a sort of road map for genetic change, with humans having a frequency of common single nucleotide polymorphisms (SNPs) occurring every 1000-2000bp in the genome (10).

Many different types of mutations exist, ranging from small mutations to mutations that occur on the whole-genome level. Small mutations are often referred to as point mutations, where only one nucleotide is changed to a different base. Point mutations may be described as

transitions, where a purine is substituted for a purine or a pyrimidine is substituted for a pyrimidine, or as transversions, where a purine is substituted with a pyrimidine or vice versa. Synonymous mutations are mutations in which a nucleotide change occurs in a coding region that does not alter the amino acid specified. In contrast, non-synonymous mutations alter the amino acid specified to either an early stop codon in a coding region or a different amino acid, with the sickle-cell anemia example above being a simple depiction of what non-synonymous mutations look like and the effects they can have on the physiology of an organism. Purifying selection, or the removal of deleterious alleles, is a type of selection that can purge harmful changes and maintain the "status quo". It is responsible for maintaining working sequences and genes within organisms and may remove any deleterious mutations detected that could alter a protein or gene or otherwise make it lose its functionality. An example of purifying selection at work is the conservation of specific coding and non-coding DNA for over 100 million years in vertebrates (12). Selection acts on mutations to ensure they are not spread to the population in general if deleterious or increases the chances that they spread throughout a population if beneficial.

Another type of mutation is an indel, the abbreviation for an insertion or deletion of DNA in a genome. As with point mutations, even a 1 base pair event can have dramatic effects. A 1 base pair insertion or deletion can cause a frameshift mutation in an organism. Frameshift mutations occur when a number not divisible by 3 is inserted or deleted into a coding sequence. If a single base pair is inserted or deleted in a given sequence, the reading frame, or grouping of codons, will shift and change the way the RNA is translated into a protein. If early enough in a coding sequence, these frameshifts usually lead to a premature stop codon or a completely different protein being encoded. Diseases caused by small deletions within genes include some

types of cystic fibrosis, while beta-thalassemia has arisen from small insertions. Intermediate sized indels can affect genes or gene groups, and the largest indels can cause chromosome-level changes. Chromosomes may break apart or fuse (1,2). Another class of mutations is those caused by transposable elements (TEs), or mobile DNA, that can "jump" around the genome of a given organism. Transposable elements are known to fall into two categories depending on the mechanism that they use to insert themselves into a genome, either by copy and paste methods (Class I) or by cut and paste methods (Class II). TEs can be a common type of large indel. Transposable elements can dramatically change the genomic landscape by potentially deactivating a current gene by inserting itself into the middle of the gene, reactivation of dormant genes by excising older insertions that deactivated the now-dormant gene, or even creating a new gene by pasting old genes or gene fragments together (1,2,13). Moreover, TE insertions are responsible for hundreds of documented cases of derived novel gene expression patterns by inserting into regulatory regions. They also cause numerous types of other rearrangements by serving as sites of ectopic unequal recombination or by their ability to initiate double-stranded DNA breakage (2).

On a larger scale, chromosomal mutations include inversions, deletions, translocations, and duplications. Inversions take place when a section of a chromosome is excised and reinserted in the opposite orientation. An example is seen in Opitz-Kaveggia syndrome in humans that leads to many phenotypic anomalies, including mental disability. Chromosomal deletions are, as the name suggests, when a region is deleted and the sequences contained therein are lost. An example in humans is Cri-du-chat syndrome that yields problems with the larynx and nervous system. Chromosomal segment duplications and translocations are associated with numerous forms of cancer and leukemia in humans (1,9). When entire chromosomes are either lost or

duplicated, copy number variation (CNV) is said to have occurred. Segmental or individual gene duplications or deletions also cause CNV. One type of CNV is for simple sequence repeats (SSRs), which are often associated with disease in humans if the SSR is inside the coding region or regulatory region of a gene. CNV can lead to many severe gene-dosage related issues. Gene amplification of a given locus has been associated with breast cancer, while trinucleotide repeat expansion is associated with Huntington's disease (9).

Mutational hot spots are common in many genomes, and are defined as areas that are prone to have higher numbers of mutations in comparison to the rest of the genome. A mutational hot spot is likely to contain areas with repetitive DNA. For example, areas with many SSRs are prone to polymerase slippage during DNA replication or repair, where DNA polymerase may "lose track" of the number of times it has used a given repeat as template and incorrectly replicate a region, thereby changing the number of times the repeat is found in the genome (14,15). In addition, the frequency of point mutations around indels increases significantly in many organisms including primates, rodents, and rice, although this higher frequency for point mutations dropped to background levels after moving only about 200 bp away from the indel (16). Further, hotspots for recombination involving double strand breaks (DSBs) can thereby create hotspots for additional sequence change, with the highest observed increase in mutation rates at approximately 2 kb from a DSB in *E. coli* (17).

There are many ways for mutation to arise both naturally and artificially. In brief, ultraviolet or ionizing radiation, chemical exposure, spontaneous mutations, tautomerization, and replication or DNA repair errors are some of the many pathways from which mutation can arise. Ultraviolet (UV) radiation has been a widely studied cause of mutation, with both naturally occurring UV radiation and applied UV radiation having been widely studied (18-19). UV-

exposure to a genome can generate mutations that lead to severe diseases like skin cancer in humans. One type of skin cancer susceptibility is caused by UV-induced damage to the *p53* gene, a widely studied oncogene (20-21). Another disease, xeroderma pigmentosum, is caused by a defect in the UV-damage repair pathway and leads to intolerance to sunlight exposure in a patient. Only UV-A and UV-B radiation are able to enter through the Earth's atmosphere due to the ozone layer. A specific type of UV-induced mutation is when a cytosine undergoes hydrolysis and changes to a hydrate form, which permits the base to incorrectly pair with adenine during replication and will eventually lead to it being replaced by a thymine. Cells associated with basal cell carcinoma, for example, have shown high amounts of this type of UV-produced C to T mutation (17). Ultraviolet radiation can also lead to the creation of covalent bonds between adjacent pyrimidine bases on a strand of DNA, which creates a cyclobutane pyrimidine dimer (CPD). Further, 6-4 photoproducts can be created by ultraviolet radiation exposure. UV-induced lesions in DNA cause warped or changed structure, typically described as bends or kinks, which can ultimately hinder both replication and transcription (22). Repair processes exist for these types of mutations, but are not foolproof, and will be discussed below.

A differing type of radiation, ionizing radiation, may also act as an engine for the creation of mutations, specifically the generation of DNA double-strand breaks (DSBs). Ionizing radiation includes gamma rays, cosmic rays, or X-rays, and these types of radiation are able to breach cells and tissue all over the body of a given organism. DSB repair can be a bit more complex than other forms of repair, and is also prone to errors, leading to great potential change in a genome when encountered (see below).

Tautomers, or a differing chemical form of a compound, exist in DNA and can contribute to improper base pairing due to the rarer *imino* and *enol* forms of some bases being unable to pair

with their normal counterpart (23). This phenomenon was first described by Watson and Crick and had to be resolved before they were able to finalize their model for the double-helix structure of DNA (2,23). These tautomers can lead to a base being replaced after it mispairs due to its usual *keto* form not being present.

Chemical exposure, particularly to oxidizing agents or free radicals, can also alter nucleotides and their ability to properly pair with other bases. Dioxin may intercalate between nucleotides and cause instability and increased probability of indels at a given base pair. Benzo[a]pyrene, found in cigarette smoke, is a known carcinogen that has been demonstrated to cause lesions for guanine bases in the *p53* tumor suppressor gene, leading to increased risk of lung cancer (24,25). Chemical agents that cause mutations can lead to oxidization, alkylation, or hydrolyzation. Base excision repair (BER) is the primary pathway to repair such changes with relatively high accuracy, such that less than 1 in 1000 of these types of mutations will become permanent (9,25). Drugs targeting DNA for cancer or disease treatment, such as temozolomide, can also become alkylating agents and may cause double-strand breaks in a patient's DNA (25). Additionally, metabolism-related chemical processes may also lead to mutations. For example, heterocyclic amines created during the cooking of meats can covalently bond to different sites on DNA to create bulky DNA adducts (23,26).

Mutations may also arise spontaneously within a genome. In the case of depurination, a purine base may be lost from a nucleotide as a result of hydrolysis that can occur without any environmental cues. If the depurination is not rectified, the base will change; with the most common permanent change being that an incorrectly paired adenine is introduced where the depurination event occurred. This is because replication will stop when the replication complex does not recognize a proper base at the apurinic site, and an incorrect base may be inserted to

close the gap, often by one of many lesion-bypassing DNA polymerases. Another example of spontaneous mutation is the deamination process, where the amine group in a base is removed. If a cytosine is deaminated, it becomes uracil, which will improperly pair with adenine as opposed to the guanine the cytosine would have paired with, resulting in a spontaneous substitution. This spontaneous deamination is carried out through hydrolysis of the cytosine to uracil that releases ammonia as a byproduct. Bisulfite can also remove the amine from cytosine but not from a methylated cytosine (21). Uracil-DNA glycosylase removes the uracil and creates an AP (apurinic/apyrimidinic) site. AP Endonucleases will remove the AP site, polymerase repairs the lesion, and the repaired strand is then ligated to complete the repair (17, 25). The described repair process is an example of base excision repair, which will be discussed in further detail below.

Lastly, DNA mutations may derive from replication or repair errors. As mentioned above, trinucleotide repeat expansion leads to a common form of replication error in which a hairpin loop of the repetitive region is created during replication that ultimately leads to polymerase slippage and the formation of an indel. The strand slippage that occurs can happen between the template and newly synthesized strand of DNA when one of the strands creates the hairpin loop, and an insertion would be the result of the template strand slipping where a deletion would be the result of the newly synthesized strand slipping and polymerase incorrectly resolving the hairpin. DNA polymerase is typically very accurate, but still will make a mistake approximately 1 in every 100,000,000 nucleotides (25). When genome size is taken into account, the polymerase error rate would translate to approximately <48 errors per S phase in humans, <8 errors per S phase in rice, or <10 errors per S phase in foxtail millet.

DNA repair pathways are not 100% efficient, leading to the inevitability of mutations being transmitted. There are many types of repair pathways associated with DNA repair, each with its own subspecialty and timing in the cell cycle. The repair mechanisms are base excision repair (BER), mismatch repair (MMR), nucleotide excision repair (NER), translesion synthesis, or double-strand break repair (DSBR). DSBR is typically divided into nonhomologous end joining (NHEJ), homologous recombination repair (HRR), or microhomology-mediated end joining (MMEJ) mechanisms (1,19,25).

Base excision repair is characterized by its handling of DNA damage caused by chemicals. This can be when a base has been alkylated, oxidized, hydrolyzed, or deaminated and can also be triggered in response to reactive compounds created during metabolism (19). Examples of base lesions repaired by base excision repair are bases like hypoxanthine formed from the deamination of adenine, 3-methyladenine and 7-methlyguanosine which occur from alkylation, and 8-oxoguanine and 2,6-diamino-4-hydroxy-5-formamidopyrimidine that arise from oxidization. Unlike nucleotide excision repair, BER repairs bases that do not alter the structure of the DNA double helix. Base excision repair is initialized by glycosylases that are able to identify and eliminate damaged bases by cleaving the covalent bonds between the damaged bases and the sugar-phosphate backbone, which leads to the formation of AP sites. The AP sites are cleaved by AP endonucleases, and the resulting single-strand break is repaired. Depending on the length of the break, either short-patch BER where a single nucleotide is replaced or long-patch BER where 2-10 nucleotides are replaced is used (19). In either scenario, the gap is filled by polymerase and then sealed by a ligase. The downstream steps of BER can also sometimes be used to fix spontaneous single strand breaks as the mechanism is the same by excising nearby bases, filling in the gap with polymerase, and ligation. In humans, a noted

reduction in base excision repair occurs during aging, and increased risk for cancers can be the end result. Additionally, defective mutations in genes encoding BER enzymes in humans like polymerase ß are found in roughly 30% of cancer patients (19, 21).

Mismatch repair is a type of DNA damage repair that occurs post-replication and specifically targets incorrect nucleotides that have been added to the growing strand and thus have stalled replication due to the exposed 3'-OH group being improperly positioned (26). Defects that can be repaired by MMR are commonly caused by tautomerization during DNA replication. Proofreading during replication performed by DNA polymerase enzymes can detect this type of error and can fix the improper base with approximately 99% accuracy. MMR enzymes are able to detect the structural disfigurements in DNA when bases are improperly paired and assist in replacing the improper bases. If a mutation were to survive both proofreading and mismatch repair, it may become a permanent mutation. Mismatch repair is carried out in 3 general steps: 1) recognition of the mismatched base(s) on a specific strand, 2) excision of the mismatched bases and creation of a gap, and 3) repair synthesis and ligation. It is strand specific and can correct indels that arose as a result of strand slippage or base substitutions caused by improper base placement. Mismatch repair is highly conserved in organisms ranging from prokaryotes such as *E. coli* all the way to eukaryotes like humans (19,26). The MutS and MutL protein domains are associated with MMR, and defects in either have been shown to cause microsatellite instability, increased mutation rates, and cancer in humans (26,27).

Nucleotide excision repair is one of the potential DNA damage responses (DDR) to exposure to ultraviolet radiation. UVA or UVB can cause two classes of lesions that result in bulky adducts, cyclobutane pyrimidine dimers, or 6-4 photoproducts. Ultraviolet radiation-related lesions cause structural changes (bands or kinks) in DNA and lead to the inhibition of

both transcription and replication. NER is used to repair such damage, and involves multiple genes in different organisms, with at least 30 implicated in humans (19,26). Nucleotide excision repair generally comprises 4 steps, with the first being the UV-induced damage to the structure of the DNA double helix. In the second step, multiple damage detection proteins are utilized that scan the genome for helix distortions, including the DNA-damage binding (DDB) and XPC-Rad23B complexes. In prokaryotes, the main enzyme complexes are the UvrABC endonuclease enzymes and DNA helicase II. Next, the strands are separated and proteins support the intermediate single strands, and the damaged DNA is cleaved. Lastly, the gap is filled in by polymerase and is finally sealed by DNA ligase. Defects in the NER pathway cause an inability to handle lesions created due to exposure to ultraviolet radiation and are associated with diseases such as xeroderma pigmentosum, mentioned above, or Cockayne's syndrome, which is a neurodegenerative disorder that can lead to growth suppression, sensitivity to light, issues with vision, and early aging in humans. It is mechanistically similar to base excision repair, but as NER can involve 9 proteins and 30 genes, there is a bit more room for error or defects in the NER pathway. There are two subclasses of nucleotide excision repair, global genomic or GG-NER, or transcription coupled or TC-NER. GG-NER uses the DDB and XPC-Rad23B complexes to constantly scan the genome for helix deformities as mentioned above, with XPC-Rad23B being responsible for detecting the helix distortions and DDB being responsible for the detection of UV damage. When damage is found by the XPC-Rad23 complex, repair is initialized and damaged regions of the double helix are corrected. Xeroderma pigmentosum can be a result of defective GG-NER pathways. Transcription coupled NER is a bit different in that it does not require XPC or DDB proteins in mammalian cells, something that GG-NER's damage detection and repair are based upon. In contrast to GG-NER, TC-NER begins when RNA

polymerase undergoes a stall at a lesion in DNA, with the polymerase itself being substituted as the signal for detection of double helix structural damage. After, CS protein complexes, namely CSA and CSB, are able to bind and repair the damaged area in lieu of the XPC-Rad23B-driven repair in GG-NER. Known diseases associated with TC-NER in humans are Cockayne syndrome, described above, and trichothiodystrophy, a disorder associated with ichthyosis and retardation. As also observed in the case of base excision repair, aging cells have been shown to exhibit a decreased ability to successfully carry out nucleotide excision repair in many organisms (29).

Translesion synthesis (TLS) is another damage tolerance and repair process in DNA that is a bit unusual in its approach when compared to other simpler excision-based mechanisms. During replication, if a lesion is encountered that may stall the replication machinery such as an AP site or thymine dimers, the TLS process will swap the normal replication DNA polymerases for a more specialized translesion polymerase. Translesion polymerases are typically members of the DNA polymerase IV or V families, and are able to handle the "usually-proper" insertion of bases across from damaged nucleotides (30). The swapping of the polymerases to translesion polymerases is mainly handled by the replication processivity factor PCNA (proliferating cell nuclear antigen). Though the translesion polymerases are known to have relatively low accuracy when inserting bases with templates with no damage, their specialty lies in the recognition and bypassing of specific types of damage. A particular example would be damages involving lesions caused by UV radiation, in which polymerase η is able to add an adenine across from a T^T photodimer using standard Watson-Crick base pairing and then adding a second adenine with Hoogsteen base pairing (30). Hoogsteen base pairing in this case between the A and T is a deviation from standard base pairing in which the nucleotides can be held together by hydrogen

bonds in the major groove. TLS also enables the bypassing of certain types of lesions, such as the guanine-thymine intra-strand crosslink G(8,5-Me)T and the subsequent relocation of the replication fork (31). The bypass mechanism is facilitated by the PCNA at the location of the lesion by RAD6 and RAD18 proteins that allow the PCNA to bypass the lesion and continue replication afterward. Next, TLS requires extension after the bypass or repair, which is typically carried out by another specialized polymerase, polymerase ζ. The final step involves PCNA switching back to the usual processive polymerase and for replication to go on as it had before encountering the lesion.

The final major form of DNA repair is double strand break repair and can be divided into three subtypes: non-homologous end joining (NHEJ), homologous recombination repair (HRR), and microhomology-mediated end joining (MMEJ). Double strand breaks are an extremely deleterious type of damage to DNA and can be the result of ionizing radiation, ultraviolet radiation, genotoxic chemicals such as oxidative free radicals, transposable element action, or mechanical stress. They can lead to cell death or severe chromosomal and/or genomic rearrangements and, if left unrepaired, can result in drastic genomic changes for an organism. In humans, DSBs are also known to cause cancer if tumor suppressors are inactivated (27, 32). No S phase can be completed if there is even one unrepaired DSB, and cells with such unrepaired DSBs will undergo apoptosis in mammals (2,19).

Homologous recombination when used as a repair mechanism for DSBs guarantees a higher level of accuracy as an undamaged sister chromatid or a homologous chromosome is used as a temple to repair DNA damage. It is mediated by the Rad52 protein family. When a DSB is detected within a cell, the ataxia-telangiectasia mutated (ATM) complex activates the damage recognition and response pathways and will phosphorylate downstream of the DSB (19,27). The

cell cycle is halted and ATM will either mediate repair or apoptosis depending on the severity of the damage to the DNA. The MRX complex will bind to DNA on both sides of the DSB, and the resection process begins. MRX recruits the Sae2 protein, and both then create 3' overhangs on either side of the break by exonucleolytic removal of 5' end sequences. Sgs1 helicase is used to unzip the double stranded DNA and Exo1 and Dna2 nucleases cut the single stranded DNA produced by Sgs1. The single stranded DNA 3' overhangs are then bound by the RPA protein. Rad51 forms a filament of protein and nucleic acids on the single stranded DNA coated with RPA, which then starts to look for DNA that has high homology to the 3' overhangs. The filament will invade the similar strand when found, forming a displacement loop (D-loop), and DNA polymerase will extend the end of the invading 3' strand. The extension of the invading strand converts the D-loop to a structure called a Holliday junction in the shape of a cross. The second 3' overhang that did not invade another strand also forms a Holliday junction with the homologous chromosome. Both Holliday junctions are converted using nicking endonucleases to recombination products, which sometimes results in chromosomal crossover. DNA synthesis occurs and restores the strand on the homologous chromosome that was displaced during invasion (33). Homologous recombination has a high level of accuracy but does require undamaged DNA to exist that can be used as a template for DSB repair.

Non-homologous end joining, in contrast to HRR, does not require homologous template DNA to repair double strand breaks but may be less accurate. NHEJ employs microhomologies typically found at the single stranded overhangs around the double strand break and repairs the damaged DNA using these small areas as nucleation sites for templates to restore the DNA. NHEJ is strongly conserved throughout the tree of life and is the most widely used DSBR pathway in mammals (34). In mammals, several proteins are involved in NHEJ. The Mre11-

Rad50-Nbs1, or MRN, complex and a Ku70/Ku80 heterodimer are recruited and bound to the double strand break. Ku then recruits other factors to the site of the DSB, including DNA-PKcs, X-ray cross complementing protein 4 (XRCC4), DNA ligase IV, XRCC4-like factor (XLF) and APCF. The DNA-PKcs tether the ends of the DSB and undergo phosphorylation. Next, the end processing step removes damaged or mismatched bases with nucleases (Artemis digestion) and subsequent gap filling with polymerases. The final step is ligation, which is carried out by DNA ligase IV and the XRCC4 complex. In summary, the DSB is detected, the ends are bound and tethered, and the ends are then processed to repair damaged or incorrectly paired nucleotides, the correct nucleotides are synthesized, and, lastly, the strands are ligated. If the NHEJ pathway has been deactivated or damaged, microhomology-mediated end joining (MMEJ) may be used instead. The downside to MMEJ is that it is very error-prone and is always associated with causing deletions flanking the original double strand break. In contrast to NHEJ, MMEJ is known for its use of 5-25 base pair microhomologous sequences to align and repair the double strand break (35-36). First, the mismatched strands of hanging DNA caused by the double strand break are ligated. Then, the hanging nucleotides are removed. Lastly, gaps are filled in by new nucleotides. After a DSB takes place, 5-25 complementary base pairs are identified and then used on both strands to align the broken strands. Any overhangs or mismatched base pairs are removed and missing nucleotides are inserted to fill in gaps. Due to the use of microhomologies, lost base pairs are not accurately detected since the homology had to have occurred upstream or downstream of the double strand break. So, there is no step in MMEJ concerning error checking and it always causes deletions when used as a mechanism to repair DSBs. As such, MMEJ is typically only employed if a cell cannot use the more accurate standard NHEJ mechanism to repair itself (35).

Although multiple repair pathways exist to attempt to handle and process the many ways that DNA can undergo mutation, no species has a mutation rate of zero. Mutations are an inevitable consequence of life and reproduction across all the kingdoms of life, and are a necessary component for selection to act. The understanding that even a single mutation can cause a phenotypic response or a potentially harmful disease such as sickle-cell anemia leads to an appreciation of the mechanisms by which mutations can arise despite the multitude of repair and damage response pathways that exist to mitigate the numbers of mutations. However, the majority of evolutionary changes derive from the cumulative effects of multiple mutations that may have only small effects if considered individually. Depending on its nature and location, a mutation may be neutral, beneficial, or deleterious; with the majority of non-neutral mutations being deleterious (1,8). A protein that has undergone millions of years of evolution may be thought of as a protein that has been polished and that is functioning with very little room for improvement. However, a frameshift mutation or nonsynonymous point mutation may "break" the protein and change what is being coded for at a given site, so even a well-conserved protein may be easily damaged or altered. For evolutionary studies, the focus is often on how mutations may affect an organism's level of fitness, or the ability to successfully pass its genetic material onto the next generation, and if certain allelic variations or mutations lead to selection and advantages in a given environment. Some mutations have a degree of interplay by which a given set of mutations may interact with one another and alter the expression or effects of other allelic variants, which is known as epistasis. Despite the complexity involved when considering the outcomes of mutations in a given species, the vast majority of mutations have relatively small effects (4), but the most interesting changes occur when considering the accumulation of numerous mutations over time. Because the current sequence variation in any individual

organism is a combined outcome of *de novo* mutation rates/types, selection and population history, then it would be informative to dissect out these components.

Recently, a few studies have focused on estimating *de novo* mutation rates and spectra across whole genomes in multiple different organisms, including *Escherichia coli*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Schizosaccromyces pombe* and humans (37-39). The general strategy for these studies is to use an organism with a precisely known genome sequence as a starting point to compare to derived sequences, with inbred reproduction used in the case of plants (38) to maintain progeny continuity. As mutation rates can vary between genomes and within genomes, using a well-studied pedigree with self-crossing done where possible eliminates potential noise in the study and facilitates the detection and tracking of mutations in a given lineage over a relatively short amount of time (several years). Many lines can be generated and have their genomes sequenced, and all resulting mutations will be identified to allow calculation of the mutation types, rates, locations, etc. (5,37).

Because mutations are rare, and might often be confused with sequencing errors, the best studies have employed "mutation accumulation" strategies. Of the many species that mutation accumulation studies have been conducted in, only 6 have had multiple groups conduct a study on the same species and generate their own data as of this writing. The number of generations and lineages studied varies between groups, yet the mutation rates and spectrum tend to be quite similar even down to the relative number of substitutions versus small indels detected in both studies (37). The advent of modern mutation accumulation studies can be attributed to widespread whole-genome sequencing projects and data, with lower sequencing costs and existing analytical methods allowing a deeper level of accuracy at a quicker pace. Mutation rates for humans have been estimated to be between $1x10^{-4}$ to $1x10^{-6}$ per gamete (17,39). Rates

estimated for other species were $7x10^{-9}$ per base/generation for point mutations in *A. thaliana*, $2.2x10^{-10}$ per base/generation in *E. coli*, $2.4x10^{-10}$ per base/generation in *S. pombe*, and $2.8x10^{-9}$ per base/generation in *D. melanogaster* (37,40-42). In general, base substitutions are more likely to occur than an insertion or deletion in a given genome, with the number of base substitutions sometimes being an order of magnitude higher than the number of insertions and deletions in organisms such as *A. thaliana* and *D. melanogaster* (38,41).

In *E. coli*, mutations were found to be evenly distributed throughout the genome (40). As is common in multiple organisms, *E. coli* mutations were clustered near areas with known base methylation as well as repetitive regions known to be hotspots for CNV, with the indels being smaller than 4 bp (40).

In *D. melanogaster*, Keightley et al. found an estimated $2.8x10^{-9}$ substitutions/site/generation where Schrider et al. found a rate of $\sim 5.5x10^{-9}$ substitutions/site/generation (41,42). Keightley et al. found randomly distributed mutations, with 6 substitutions and 3 small deletions arising after only one generation between 12 progeny lines (41). Schrider et al. reported 732 substitutions and 60 small indels in 8 lines after 147 generations, with roughly 2% of mutations clustering in regions that could be considered mutational hotspots (42). Both studies in *D. melanogaster* reported a bias towards G:C->A:T mutations.

It is a considerable challenge to attempt to find *de novo* mutations in a species. Few studies have focused on uncovering *de novo* mutations in plant systems. One mutation accumulation study was completed in *Arabidopsis thaliana* (38). Ossowski et al. generated 5 separate mutation accumulation (MA) lines of *Arabidopsis thaliana*, maintained 30 generations by self crossing with single-seed descent. The mutation accumulation lines were sequenced at

23-31x coverage using the Illumina platform, and substitution and indels were identified as any sequence changes from that of the reference genome (which was the initial self-cross parent) that were specific to one and only one of the 5 MA lines. Ossowski et al reported a total of 99 substitutions and 17 small indels that occurred *de novo* in the *A. thaliana* genome, thus predicting rates of $7x10^{-9}$ substitutions/site/generation, $0.3x10^{-9}$ insertions/site/generation and $0.6x10^{-9}$ deletions/site/generation. As *Arabidopsis thaliana* is known to have a heavily methylated genome of approximately 135 Mb, Ossowski et al. were not surprised to observe a significantly higher number of G/C -> A/T transitions than any other substitution type. Their explanation for this is that the deamination of methylated cytosine and heavy amounts of ultraviolet radiation seemed to have skewed their data toward this particular type of transition, with deamination of methylated cytosine being responsible for many C->T transitions (26,38). In addition, ultraviolet radiation is also known to cause G/C -> A/T transitions if the cytosine is either in a CC or CT group, with CC and CT groups being dipyrimidine sites. Ossowski et al. conjectured that, despite the high number of this transition, real-world estimates should be even higher because their plants were shielded from some ultraviolet radiation in a greenhouse environment during population generation. Their study was the most accurate work to date with stringently selected and rigorously verified results, but it took several years to complete. Many lines of *A. thaliana* had to be meticulously monitored and selectively bred in order to keep the 5 MA lineages pure, and 30 generations had to be grown, giving this type of study a long data generation time. Jiang et al. (38) looked at 9 MA lines and 10 generations for *A. thaliana,* as opposed to Ossowski and colleagues' 5 lines and 30 generations. Jiang et al. found similar results, with 44 substitutions and 7 indels confirmed and a strong skewing of substitutions toward G/C -> A/T transitions (38,43).

In our work, we propose an alternative approach to discover *de novo* mutations. This novel strategy can work in any system, often at zero data generation costs and zero time for population generation by the investigator. Myriad genome resequencing studies have generated thousands of data sets for genomes of many organisms, with all of the resequencing data publicly available to any researcher for free. Rather than the costly approach of generating and maintaining multiple MA lines, our method utilizes freely available resequencing data and a reference genome to search for recent mutations in any given organism. The comparison of multiple lines to a reference genome uses a diverse pool of resequencing data then enables the alignment of one well-studied reference genome to many other lines simultaneously to find recent mutations. Classically, reference genome lines are unique lines that are kept separate from their species' germplasm and will thus have the ability to accumulate mutations that are unique to that particular genome. In the case of rice (*Oryza sativa* ssp. *japonica*), the reference genome Nipponbare was derived from a unique cultivar of rice that has been genetically isolated for many years (44) and sequenced as part of an international collaborative effort. As Nipponbare is the best-studied and annotated reference genome for rice, we used comparative genomic techniques to study the differences between the Nipponbare genome and other *Oryza sativa* ssp. *japonica* genome sequences generated by the Bin Han group (45) in order to find recent Nipponbare-specific mutations. We also applied the same technique in *Setaria italica* (foxtail millet) using the Yugu1 reference genome from the Bennetzen lab (46) and resequencing lines of *Setaria italica* also from the Bin Han group (47).

We employed this strategy as a way to identify recent mutations that were specific to the Nipponbare line in Japanese rice and recent mutations in the Yugu1 line of foxtail millet. Although our method dramatically enriches for *de novo* mutations, we cannot determine rates of

mutation because we do not precisely know how far removed the reference lines are from the lines used to generate the resequencing data analyzed. Moreover, our approach will not detect *de novo* mutations if they generate a variant that was already present in one or more of the resequenced populations. However, the strategy is robust and generates a great deal of novel observations at low cost and with minimal demand on experimental resources. We propose this strategy as a cost-effective method to correct reference genome errors and to find novel mutations between a reference genome and resequencing data that yields similar results to the more detailed but expensive and time-consuming MA line protocols.

CHAPTER 2

EFFICIENT DISCOVERY OF *DE NOVO* GENOME CHANGE IN NIPPONBARE RICE

Introduction

One of the primary questions in biology is the nature and origin of genetic change. Evolutionary biologists routinely use comparative genomic analyses to identify the changes that differentiate individuals within a species or between species. However, these methods uncover variations that are the outcomes of multiple phenomena, including rates and natures of *de novo* mutation, natural selection acting on these changes, and transmission issues associated with mating strategies, population sizes, and geographical distributions. Mutation may arise due to spontaneous or environmentally-driven base modification, errors during DNA replication, inaccurate DNA repair, transposon insertion/deletion or chromosome breakage (48-49). Multiple DNA repair mechanisms work in concert to minimize change, such that the tens of thousands of DNA changes generated every cell generation still yield only mutation rates of $1x10^{-9}$ to $1x10^{-12}$ per base per organismal generation (8,38,50). For instance, in the model plant Arabidopsis, with just over 10^8 bp in its nuclear genome, ~1 *de novo* mutation is expected to be transmitted in a single plant generation (38).

The rate at which mutations occur can vary within or between species, and taxa also differ in the relative frequencies of types of mutation, although point mutations (both substitutions and tiny indels) are routinely far more common than larger indels (1,3,6,38). Within genomes, genic regions exhibit a lower number of accumulated mutations, at least partly due to

the fact that coding sequences are usually subject to purifying selection (3). Regions in genomes that are methylated, such as CG dinucleotides in many animals and all plants, also display a higher point mutation rate because 5-methyl cytosine deaminates at a higher rate than unmethylated cytosine, leading to frequent cytosine to thymidine transitions (26). Taxa with active transposable elements (TEs) can accumulate dozens of *de novo* insertion mutations per generation, while sister lineages with quiescent TEs can go hundreds, perhaps millions, of years without any new TE insertions (44,45,51).

While most mutation analysis studies have focused on changes that have accumulated over evolutionary time, few studies have investigated *de novo* change because of the cost and temporal demands of such investigations. Estimation for the spontaneous mutation rate in *E. coli* was $\sim 2.1x10^{-10}$ *de novo* changes per genome per generation, with point mutations outnumbering indels by >9:1 (40). In *S. pombe*, the rate of point mutations was $2.4x10^{-10}$ base/generation (37). In humans, sperm DNA sequencing was utilized to investigate *de novo* DNA change, and predicted a mutation rate of $\sim 2.4x10^{-8}$ mutations/base/generation (39). In the plant kingdom, Ossowski and colleagues conducted a mutation-accumulation study in *Arabidopsis thaliana* that discovered 99 new base substitutions and 17 indels that had accumulated in 5 lineages within 30 generations, and found an overall mutation rate of $\sim 7x10^{-9}$ for point mutations/base/generation and $0.3x10^{-9} - 0.6x10^{-9}$ for insertions and deletions respectively/base/generation (38).

Rather than spend the several plant generations to create mutation-accumulation lines (37-40) and then demanding deep full genome sequencing to identify/confirm any *de novo* mutations, we have chosen to utilize currently available genome data to enrich for *de novo* mutations without any investment of plant growth time or sequencing expense. We have chosen

to undertake this initial study with the rice variety Nipponbare, although the strategy can be applied to any species that has both a reference genome sequence and resequencing data for other lineages within the species. Nipponbare is the rice cultivar that was the target of the first high-quality reference genome sequence for *Oryza sativa* (44,45,51). Because Nipponbare is a unique cultivar, it has accumulated *de novo* mutations in the generations that it has been separate from any other rice germplasm. Recent genome resequencing studies have investigated a great deal of the germplasm of domesticated rice, providing the raw material for genome comparisons (45). In this study, we present a new protocol whereby resequencing data can be used in tandem with a reference genome to analyze *de novo* genomic instability, and we herein identify and confirm thousands of recent mutations in the Nipponbare lineage of *Oryza sativa* ssp. *japonica*.

Results

The Nipponbare genome sequence was generated as part of an international consortium effort to sequence *Oryza sativa* ssp. *japonica* at the highest resolution and quality possible at that time (51). It has since been updated by correction of genomic sequencing errors, extension of previous gap sequences (N chains), and correction and expansion of annotation (44). Nipponbare is currently one of the best annotated and highest sequence quality genome assemblies in the plant genomics world, and as such provides an excellent resource for all sorts of evolutionary, genetic and molecular studies.

In order to discover the nature and relative frequencies of different types of *de novo* mutations in rice, we selected Nipponbare as the target genome. The basic concept is that Nipponbare has had a unique breeding history (as has any unique cultivar or lineage within any species), and that any *de novo* mutations during the descent/creation of Nipponbare would not be found in any other rice variety. Hence, Nipponbare IRGSP 1.0 was broken *in silico* into 50-bp

oligomers (50mers) that cover the entire genome 2X because they overlap by 25 bp. These 50mers were then compared to pools of shotgun sequence data from the resequencing of closely related rice cultivars, all from subspecies *japonica* (45). Any 50mers that had an exact match with any read from the shotgun resequencing data were judged to be not specific to the Nipponbare lineage (and, thus, probably ancestral), and were then removed from further analysis. The "Nipponbare-specific" 50mers that remained could then be investigated to see if they were due to sequencing errors in the original Nipponbare assembly or were truly unique to the Nipponbare cultivar. The usage of overlapping 50mers allowed a precise positioning and confirmation of any Nipponbare-specific mutation, because any change (even a single bp indel or substitution) should affect at least two overlapping 50mers.

Many *Oryza sativa ssp. Japonica* cultivars were resequenced at low redundancy in efforts to study the domestication of Asian rice (45). Out of the publicly available data from this publication, we chose 126 accessions with good quality sequence data to provide a combined ~113.8x coverage. The sequences generated were between 0.5x and 2x coverage for each line, and were generated via the Illumina Genome Analyzer llx platform as paired end reads of size 73 bp (45).

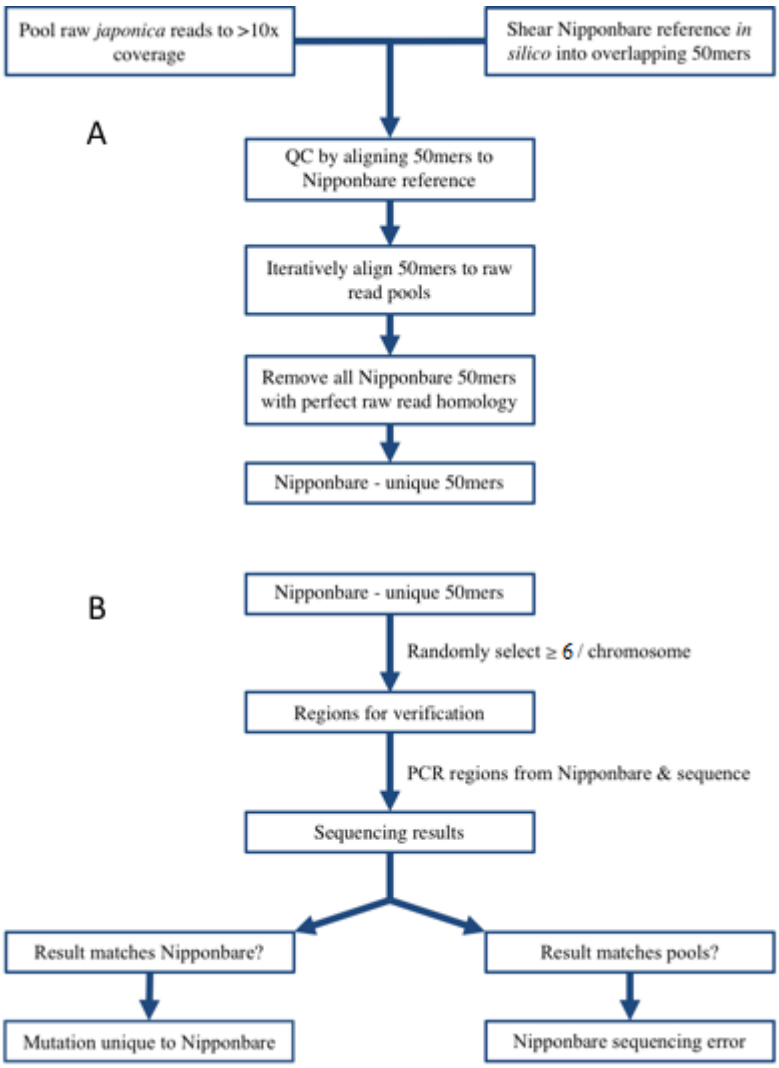


Figure 1: Flow chart depicting the steps taken in finding recent mutations in the Nipponbare line via comparison to pools of other *japonica* rice lines. A: Candidate Nipponbare-specific 50mer discovery. Initial *japonica* accessions were pooled to approximately 10x coverage per pool and Nipponbare was sheared *in silico* to create overlapping 50mers. Iterative alignments were conducted between the Nipponbare 50mers and the *japonica* pools, and all Nipponbare 50mers with perfect homology to the *japonica* pools were removed from consideration, yielding 50mers unique to the Nipponbare line. B: Resolution of candidate Nipponbare-specific 50mers between sequencing error and *de novo* mutation possibilities. Random selection of at least 8 Nipponbare-

specific 50mers per chromosome was conducted to identify potential changes across the genome. Polymerase chain reaction (PCR) and sequencing of PCR products were performed on the selected Nipponbare 50mers, followed by classification of either a mutation unique to Nipponbare or a Nipponbare sequencing error.

Overlapping 50mers from the Nipponbare reference genome were created by a custom PERL script using the repeat-masked Nipponbare genome. We then conducted a quality control step in aligning the 50mers back to the Nipponbare reference genome assembly to ensure 50mer accuracy. A 100% homology between all 50mers and the assembled Nipponbare genome was observed. Then, comparisons were performed of the Nipponbare 50mers to the *japonica* line resequencing data. The *japonica* reads from (45) were combined into "pools" that had a total coverage of approximately 10x per pool, resulting in 11 total pools. The Nipponbare 50mers were then iteratively aligned to each pool using Bowtie2, bringing the resulting total alignment coverage to ~113.8x across all 126 accessions in the data pools. Any 50mers that mapped to the *japonica* pools with perfect homology (50/50 bp match) were removed, leaving only those 50mers that showed a possible difference between Nipponbare and other *japonica* lines to be analyzed.

The genomic coordinates of the candidate missing 50mers were extracted, and some of the candidates were chosen at random to be verified using PCR and sequencing. The PCR reactions were carried out with template data from Nipponbare, and the resulting PCR products were analyzed by direct Sanger sequence analysis of excised PCR bands. If the 50mer that was amplified and sequenced matched the original Nipponbare genomic sequence, the 50mer was considered to be a Nipponbare-specific mutation. If the 50mer sequence did not match Nipponbare, it was considered to be an error in the Nipponbare assembly.

We created ~17 million starting 50mers because each 50mer has an overlapping 50mer at 25 bp intervals in the ~430 Mb Nipponbare genome (44). The "candidates" were 50mers that were not identical to any raw sequence in the 11 *japonica* pools, and thus not removed from our analysis. Our computational analysis comparing the DNA sequences of the Nipponbare and *japonica* pools yielded 17,588 50mers of interest out of the ~17 million starting 50mers. Because each of these 17,588 were covering each sequence novelty 2-fold, due to the overlap, this led to two candidate 50mers representing the same sequence novelty region. Hence, these results indicated 8794 novel sequence candidate sites. These candidates could represent novel recent mutations unique to the Nipponbare lineage or errors in the Nipponbare assembly (Figure 1), so 250 of them were taken at random and investigated via PCR and sequencing.

Table 1: Distribution of verified Nipponbare-specific sequences and Nipponbare sequencing errors organized by chromosome across the Nipponbare genome. The numbers in brackets indicate the number of actual sequence changes within these 194 analyzed sequences.

Chromosome	# of 50mers Analyzed	# of *De Novo* Sequences Verified	# of Nipponbare Sequencing Errors
1	8	6	2
2	9	7	2
3	13	9	4
4	61	4	57
5	17	11	6
6	9	7	2
7	6	5	1
8	13	11	2
9	10	5	5
10	20	11	9
11	11	9	2
12	17	8	9
Overall	194	93 [148]	101 [141]

Primers were designed from regions flanking the candidate *de novo* sequence sites, in order to generate predicted amplification products of 100-200 bp. These primer pairs yielded amplification products ~83% (207/250) of the time. Amplification products were subjected to direct Sanger sequence analysis of the PCR product excised from an agarose gel. Useful sequences were found in ~94% (194/207) of these sequencing attempts.

Overall, there were 93 candidate 50mers with confirmed variants and 101 confirmed errors (Table 1). On most chromosomes, the number of confirmed *de novo* alleles was equal to or greater than the number of detected sequencing errors. The exception was chromosome 4, where the ratio of *de novo* alleles to sequencing errors was 4/57. Not including chromosome 4, 89 sequence changes were true *de novo* alleles and 44 were sequencing errors, suggesting that about two thirds of our candidates are actually *de novo* mutations that are Nipponbare-specific. Our analysis indicated that the IRGSP 1.0 Nipponbare genome sequence includes a predicted 3694 sequencing errors, for an overall accuracy of 99.99914%.

Table 2: Summary of the nature of the Nipponbare-specific variants discovered in the Nipponbare genome, organized by chromosome. The numbers of observed indels, transitions, and transversions are denoted per chromosome.

Chromosome	# 50mers Analyzed	Transitions	Transversions	Insertions	Deletions
1	6	5	5	0	0
2	7	2	8	0	0
3	9	5	11	0	0
4	4	1	2	2	0
5	11	6	10	0	1
6	7	6	5	0	0
7	5	2	4	0	0
8	11	7	6	0	0
9	5	2	8	0	0
10	11	6	14	1	0
11	9	6	6	1	0
12	8	9	7	0	0
Total	93	57	86	4	1

Of the 93 50mers confirmed with Nipponbare-specific sequences, many had more than one unique nucleotide per *de novo* 50mer. The specific nucleotide changes were determined by comparing the novel 50mer sequence to the consensus sequence of the *japonica* pools. In most cases, the japonica pools had at least 90% agreement on the consensus sequence, with the exceptions presumably due to actual allelic variation among pool lineages or to sequencing errors in the data generation. Overall, there was a slightly higher number of transversions compared to transitions, and indels were quite rare (Table 2). All of the indels were tiny, involving only 1 or 2 bp.

A comprehensive analysis of the entire dataset of 8794 novel oligos was undertaken to see if any had less than 70% identity in their best hit within the japonica pools. This would be expected if any of the novel oligos were created by the insertion or deletion of a large (>20 bp) fragment of DNA, for instance due to TE activity. No such case was found, indicating that TE insertion and deletion activity had been zero during the unique descent/creation of the cultivar Nipponbare.

Of the confirmed 56 sequence changes associated with genes, 26 were from coding sequences (CDS), 17 were in introns, and 13 were from either 5' or 3' UTRs. The 26 CDS changes were found to have a dN/dS ratio of 1.6 (16/10). Overall, the dN/dS ratios for CDS changes when comparing different *Orzya* species has been found to be 0.28-0.47 (52), in agreement with the fact that most genes are under strong purifying selection. The much higher ratio observed in the *de novo* 50mer dataset is compatible with a 1:1 ratio indicative of random drift, as would be expected for *de novo* mutations that have not yet undergone long periods of selection.

We randomly chose and compared ten 10 kb windows of orthologous genes and flanking regions from Nipponbare and its close relative *Oryza glaberrima.* In these comparisons, we found 31 sequence differences in 15.6 kb of CDS (2/kb), 27 in 5 kb of UTR (5/kb), and 188 in 25.9 kb of introns (7/kb) (Supplementary Table 4). Hence, the frequencies of sequence variation per kb in introns were much higher than other gene components in the comparison of rice with *O. glaberrima*. In contrast, the 26 in 31 kb (CDS), 13 in 19.6 kb (UTRs) and 17 in 57.2 kb (introns) in the *de novo* 50mer dataset were closer to a 1:1:1 ratio per kb, as expected of a *de novo* mutation dataset.

Table 3: Number of predicted sequence differences per 50mer, based on alignment.

Exact # Differences (N) per 50mer	
N =	Differences
1	6116
2	1340
3	724
4	378
5	172
6	60
7	4

The majority of our data suggest that a single mutation on a given 50mer is the most common observed event (Table 3). We observed no 50mers with more than 7 predicted mutations.

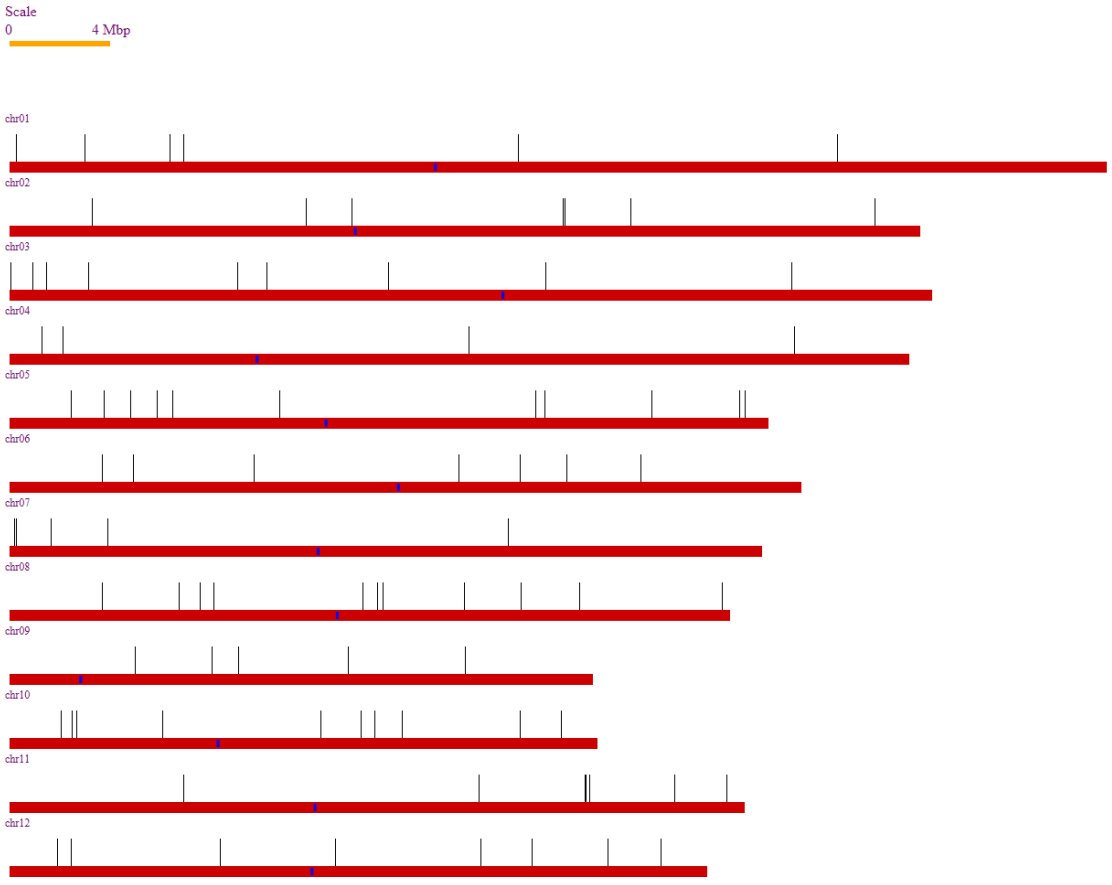


Figure 2: The distribution of mutations verified across all chromosomes of Nipponbare. Centromeres are highlighted on each chromosome in blue, and each vertical line corresponds to a confirmed Nipponbare-specific sequence variant.

The chromosomal locations of all candidate *de novo* 50mers were plotted (Supplementary Figure 3) and also of all confirmed *de novo* sequence changes (Figure 2). Sequence variants were distributed across all chromosomes, with no major clusters of *de novo* or candidate *de novo* sequences obvious, although chromosome 4 exhibited a much higher number of candidate novel 50mers because of the higher rate of sequencing errors (Supplementary Figure 3). The majority of the novel 50mers (67%) were plotted to areas in the Nipponbare genome that do not contain annotated genes.

Materials and Methods

Seed from *Oryza sativa* ssp. *japonica* cultivar Nipponbare (accession GSOR100) were provided by the USDA. PCR investigation of candidate Nipponbare-specific alleles from the *japonica* pools comparison used seed from the original 2005 distribution of GSOR100. Genomic DNA was prepared from pools of multiple plants. DNA isolation, PCR primer design, and PCR amplification were performed as in Vaughn et al. 2014 (15). Primers were designed using the software Primer3 by using 500bp flanking regions surrounding the candidate 50mer, with the target PCR product size being 100-200bp. PCR product purification was carried out as described by the manufacturer with Macrogen kits (Macrogen Corporation, Rockville, MD, USA), and samples were directly sequenced by Sanger technology, then read on an ABI 3730 machine. Inspection of resulting PCR fragment sequences was conducted manually and via the usage of BLASTn (16). As in Figure 1, if the sequencing result matched Nipponbare genomic data, it was

considered a Nipponbare-specific mutation and if it matched the *japonica* pools, it was considered a Nipponbare-specific sequencing error.

The reference genome of rice cultivar Nipponbare, version IRGSP 1.0, used as the basis for this study, was downloaded from the IRGSP 1.0 website (http://rapdb.dna.affrc.go.jp/download/irgsp1.html) on December 4, 2014 (12). *Oryza sativa ssp. japonica* pools of raw reads were obtained on December 5, 2014 from EBI (http://www.ebi.ac.uk/ena/data/view/ERP000106) and were generated by the Bin Han group (13). Nipponbare was sheared *in silico* into 50mers (Figure 1) via a custom PERL script with the resulting overlapping 50mers subsequently iteratively aligned to raw *japonica* reads in pools of approximately ~10x coverage each sorted by geographical location (13) using Bowtie2 (17) to conduct alignments under the "very-sensitive parameters" setting (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml). The 50mers that mapped back to *japonica* pools with perfect homology were removed from further analysis using the SAMTools suite (18). Target candidate Nipponbare-specific 50mers (Figure 1) were then chosen and their genomic coordinates extracted using BLASTn+ (16) against the full Nipponbare genome.

Several apparent Nipponbare-specific 50mers were chosen at random across all chromosomes for verification via PCR from Nipponbare genomic DNA, with at least 6 per chromosome, out of the total of 8794 candidate 50mers. When the Nipponbare PCR fragments were found to be identical in sequence to the candidate Nipponbare-specific 50mer, then this confirmed that the novel sequences were actually a product of Nipponbare-specific mutation during the descent of this cultivar. When the Nipponbare PCR fragments were found to differ in sequence from the candidate Nipponbare-specific sequence, then this was concluded to be

caused by an error in the Nipponbare IRGSP 1.0 sequence assembly, and was not further investigated.

Those true Nipponbare-specific 50mers that were confirmed by PCR were then compared to the most homologous sequences in the *japonica* pools. Sequences chosen as the best *japonica* read candidate by BowTie2 had at least 90% consensus among all *japonica* reads (Supplementary Table 5), and the >90% nucleotide was considered the ancestral nucleotide to the Nipponbare-specific change. Thus, the nature of sequence change from the *japonica* pool sequence to the Nipponbare-specific sequence were extracted from the alignment BAM files created by Bowtie2 through the use of a custom Python script that recorded the types and number of changes and positions they occurred on the Nipponbare-specific 50mers.

To search for large indels like those expected from transposable element insertion or excision, we used the Bowtie2 alignment results and SAMtools to search for any Nipponbare-specific 50mers that had 70% identity or less to their best hit in the *japonica* pools. The result would be expected for 2-4 50mers for an insertion (which would create two novel junction sites) and for 1-2 50mers for an excision (which would create one novel junction site). No such low homology best hits were found with any of the candidate Nipponbare-specific 50mers, indicating a complete absence of large indels that were Nipponbare-specific.

Sequence comparisons performed to calculate changes per kilobase in CDS/UTR/introns between Nipponbare and *O. glabberima* were carried out using BLAST. The data source for *O. glaberrima* was from Zhang et al. (16,19).

Unless otherwise specified, UNIX shell, PERL and Python scripts were written to execute the above analyses. No other external libraries were used.

Discussion

The abundance of sequencing and resequencing data for multiple organisms presents a novel opportunity for the study of sequence variation. Commonly, these studies use variation across individuals and populations within a species to investigate population histories, especially with respect to geographic origin and population dispersal (40,45,47). However, this wealth of data could also be mined to study the molecular nature and patterns in sequence change by comparing high quality reference genomes to the resequencing data.

All individuals, even biological clones, will differ somewhat in genome sequence from their closest relatives because of the vagaries of mutation (somatic and germinal), segregation and selection. The heritable differences that make each individual unique are an outcome of these genetic, and some epigenetic, alterations. In crop improvement, a subject that inspired Darwin's discovery and elaboration of natural selection (57), each developed variety has novel genetic characteristics that make it particularly productive in a specific agricultural environment. Most of the sequence variations that make each variety unique are presumed to be derived from the segregation of pre-existing variants dispersed in the germplasm pool, but some variation is also generated during the crop breeding process. Hence, the discovery of alleles unique to a single developed variety is expected to enrich for *de novo* mutations among the variety-specific alleles.

Most molecular evolution studies investigate the current status of allelic variation across individuals within a closely related set of taxa, for instance members of the same species. The observed changes are a combined outcome of *de novo* mutation type and rate, plus population history, plus selection. Even with large studies that provide some knowledge of the population histories, it is still difficult or impossible to determine how much of the genetic change is due to

these three contributing factors. One way to begin to sort this out is to investigate the natures and rates of *de novo* change *per se*. However, because mutation is so rare, the expense and time demands of such studies are so great that relatively few have been performed (37,38). We herein develop and describe an inexpensive and rapid alternative to methods such as mutation accumulation studies for discovering *de novo* mutations.

Our strategy utilizes a high quality reference genome sequence for the targeted organism, and then compares that sequence to all other genome sequence data that are available for that organism. The quality of the results depends on the depth of that additional sequence data, and the degree to which the other sequenced genomes are closely related to the targeted (reference sequence) genome. If the other genome sequences are deep and closely related, then the only novel sequences in the targeted reference genome will be ones that arose during the descent of the reference genome sequence lineage. Hence, this is the equivalent of a mutation enrichment study, but with all of the line progression having been done either by breeders (for a crop or domesticated animal) or by the natural process of lineage descent.

We decided to test this strategy on rice (*Oryza sativa*). The Nipponbare rice genome has been proven to be one of the best quality plant genome sequences (51). It has also been much improved over the past ~15 years (44). However, to our knowledge, no studies have been conducted to identify the recent genome sequence changes in Nipponbare rice. By dividing the Nipponbare genome *in silico* into overlapping 50mers, we were able to directly compare the Nipponbare 50mers to ~113-fold depth data for *Oryza sativa ssp. japonica* resequencing accessions generated by the Bin Han group (45). This analysis uncovered 17588 candidate Nipponbare-specific 50mers, of which a predicted ~7400 are false positives derived from sequencing errors in the Nipponbare reference genome IRGSP 1.0, while the other ~10200 are

apparently true Nipponbare-specific 50mers. Since all the candidate 50mers cover each sequence change twice, the number of unique candidate 50mers was 8794.

The primary advantages of this strategy are the immediacy of the analysis (with all data only a database download away), zero cost for data generation and low costs for result confirmation, and the ability to find thousands of candidate *de novo* sequence changes compared to the handful available from traditional mutation enrichment studies. Moreover, this technology can apply to any organism, even those with exceedingly long generation times (like some conifers) or very large genome sizes that make accurate full genome sequencing cost-prohibitive. In addition, this approach can lead to the identification of the errors in a reference genome sequence. There are some limitations, however. First, certain *de novo* changes will be missed. Any nucleotide variation that arose in Nipponbare, but was already present as a sequence polymorphism in one of the pooled japonica cultivars, will be missed as a sequence change. This should be a minor source of false negatives for most sequences, because these Nipponbare pools were near 100% identical at most nucleotide positions in the analyzed data. This is not true for repeats, however, where multiple paralogous variants are already observed, particularly with such hypervariable entities as simple sequence repeats (SSRs). Hence, we expect that our data under-represents both sequencing errors and Nipponbare-originated changes that are within repeat sequences. Because of this category of missed changes, it is not possible to calculate sequence change rates with this approach. Second, it is possible that the large pool of assembled *japonica* varieties still did not contain some of the germplasm found in the ancestors of Nipponbare. Therefore, some Nipponbare-specific 50mers would not be caused by *de novo* change during Nipponbare descent but by transmission of ancestral alleles not found in the japonica pools. If this were a major problem, then we would expect to see major clusters of

apparent Nipponbare-specific alleles caused by linkage drag on transmitted chromosomal segments, but such clusters were not observed. Thus, we feel that the great majority of our confirmed Nipponbare-specific sequences are the result of *de novo* mutation during the breeding of the Nipponbare variety.

Two additional lines of evidence support the conclusion that we have discovered novel Nipponbare alleles generated by recent *de novo* mutation. If many of the verified *de novo* changes that we observed were actually standing variation, then we would expect that those within genes would show strong evidence of purifying selection, both with a dN/dS ratio of <1 and an over-representation in introns and UTRs relative to exons. However, our verified mutations were fairly evenly distributed across the various gene components and exhibited an ~1:1 dN/dS ratio, both results expected of *de novo* variation.

Comparison of *japonica* pools and Nipponbare revealed 8794 Nipponbare-specific 50mers, of which subsequent confirmation analysis indicated that ~52% (101/194) were not actually Nipponbare-specific, but were rather errors in the reference genome sequence. Of 101 confirmed sequencing errors, 76 contained precisely 1 error and the remaining 25 contained 2-4 sequencing errors per 50mer. Hence, the overall sequence accuracy of the Nipponbare reference genome appears to be excellent. The few clusters of candidate de novo 50mers in our analysis are from chromosomes with the highest sequence error rates, so we expect that those clusters may be caused by errors that were particularly difficult to sequence, thus causing regions with overall lower quality.

Although some bases substitutions should be missed by our analysis, primarily because they are identical to some standing variation in the japonica germplasm, our approach to search for best-hit 50mers that were less than 70% identical to any *japonica* pool sequence should find

100% of the indels larger than 30 bp. Because of the vast number of different ways any indel covering a specific nucleotide position on any chromosome can have different indel end points, it is unlikely that any standing variation would be identical to any *de novo* allelic variation, except precise transposable element (TE) excision (which is a rare phenomenon, even for well-studied plant TEs (58,59)). Hence, the fact that we discovered zero cases of TE excision, insertion or other large indel variation indicates a very quiescent genome during the breeding of Nipponbare. Although most angiosperm genome analysis shows a great deal of TE activity in the last few million years in all lineages examined (13,59), including rice (44,45,59), these studies rarely have the power to differentiate events that occurred a million years ago from one that happened in the last 10 or 100 years. This surprising large indel absence in the Nipponbare lineage suggests that none of Nipponbare's improvement is associated with TE-induced genetic change, but is primarily an outcome of the improved combination of standing genetic variation.

Regarding the nature of *de novo* mutations, we confirmed 93 50mers that contained Nipponbare-specific mutations. The investigated 50mers were chosen to represent each chromosome fairly, and their locations on each chromosome were random in their selection. Confirmed *de novo* variants were not clustered on any chromosome. Of the 93 confirmed 50mers, 64 contained exactly 1 mutation and 29 of the 50mers contained more than 1 mutation. A total of 57 transitions, 86 transversions, 4 insertions, and 1 deletion were confirmed in Nipponbare. A higher frequency of substitutions compared to indels was also found in *D. melanogaster*, *A. thaliana,* and human (38,39,42). However, the substitution/indel ratio in our analysis of over 28/1 (143/5) is a bit higher compared to that seen in these other systems, such as 732/60 in *Drosophila* and 99/17 in *Arabidopsis thaliana* (38,42). These differences may be a result of different relative frequencies of mutation types (e.g., chromosome breaks vs. cross-

linking vs. deamination vs. base modification) in each lineage, or differences in the relative efficiencies of different repair processes. Studies in plants have shown that even closely-related species may differ dramatically in their rate of failure in certain types of DNA repair, and that this surprising "repair-efficiency variation" might have some selective value in creating different levels of genetic novelty in lineages that differ in their need to adapt genetically to changing environments (60,61).

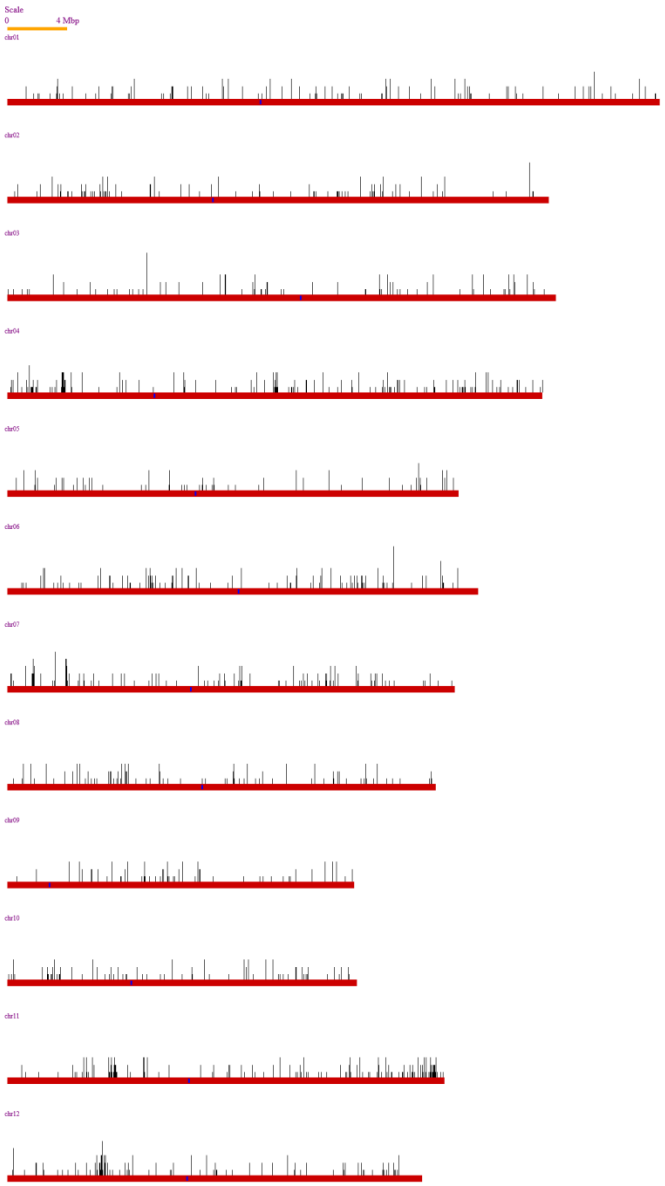
Although thousands of confirmed mutations across a genome provide a wealth of data, these are still rare events that should not be found in multiple types in a single 50 bp region if they were fully independent. Hence, we believe that some mutation processes in the rice genome act on stretches of nearby DNA rather than on single nucleotide positions. Clusters of genomic changes are typically due to repeat-rich regions being more prone to polymerase slippage during repair (3), the higher spontaneous deaminations in methylated DNA regions leading to nearby additional mutations because of repeated short patch repair (26), indels causing frameshifts that are associated with an increased frequency of point mutations around the indel (16), or double strand breaks causing multiple new mutations and becoming "induced" hotspots (17). Many of these clusters are thus likely to be an outcome of repair using less-accurate DNA polymerases (30,31) that are called into action when DNA damage in a region is not easily repaired.

We present a new method of analysis that allows the efficient utilization of freely available resequencing data and a reference genome to identify recent mutations and sequencing errors. This technique can be applied to any organism, and future studies could be designed to apply this technique to other plants, animals and unicellular organisms.

Supplementary Materials

Supplementary Table 4: Summary of number of annotated sequence changes found in comparison of ten 10 kb segments of Nipponbare to the orthologous regions of *O. glabberima* or in the novel candidate 50mer data set between Nipponbare and the *japonica* pools.

	Changes/kb CDS	Changes/kb UTR	Changes/kb Intron
Nbare to Glabberima	2	5	7
De Novo Nbare 50mers	0.84	0.66	0.3



Supplementary Figure 3: The distribution of all candidate 50mers across all chromosomes of Nipponbare. Centromeres are highlighted on each chromosome in blue, and each vertical line corresponds to a candidate Nipponbare-specific 50mer.

Supplementary Table 5: Using japonica pool consensus sequences at apparent Nipponbare-specific nucleotides to determine the direction of sequence change. The results indicate that sequence variation at these sites in the japonica pool reads was quite low (always less than 10% attributed to minor alleles). Hence, an ancestral allele could be predicted as the nucleotide that was >90% abundant in other japonica at the site of the Nipponbare-specific change. Although these analyses include all candidate Nipponbare-specific 50mers, the analysis of sequence change in Table 1 in the main text uses this strategy only to analyze our confirmed Nipponbare-specific sequence changes.

japonica pool consensus sequence	
# Candidate 50mers	% Read Consensus
4129	100
2853	95-99
1812	90-95

CHAPTER 3

ANALYSIS OF LINEAGE-SPECIFIC ALLELES IN THE YUGU1 REFERENCE GENOME FOR *SETARIA ITALICA*

Introduction

A frequently explored topic in the life sciences is the study and explanation of how mutations originate in various organisms. Comparative genomic methods are commonly used to delineate differences between and within species on the whole genome scale, although these methods may reveal variations in genome composition resulting from multiple phenomena such as *de novo* mutation, natural selection, or genetic drift. A mutation may be induced by chemical or physical damage to an organism's DNA, replication errors, insertion and shuffling of transposable elements, faulty DNA repair mechanisms, etc. (48,49,62). Despite the existence of many DNA repair mechanisms to minimize errors in replication or transmission of DNA to progeny cells, mutations are generated at measurable levels. Different organisms possess different spontaneous mutation rates, typically ranging from $1x10^{-9}$ to $1x10^{-12}$ per base per generation, depending on the species (38,40,42,50).

Mutation rates not only vary between species, but they also can vary within the same species (3) because of different levels of transposable element activity or different levels of DNA repair accuracy, for instance (3,48). Similarly, the predominant types of mutation may vary between species or within species, although single base substitutions are generally more frequent than large insertions or deletions (indels) (37-41). Routinely, non-coding regions in a genome

will have a higher level of polymorphism when compared to coding regions, at least partly because genes are more often subject to purifying selection (3). In addition, hypermethylated regions in a genome will also display higher mutation rates due to the rapid deamination of 5-methyl cytosine to thymidine (26,39).

Most molecular evolutionary studies that investigate DNA sequences tend to focus on changes which accumulate over long periods of time, with little work being done on the study of *de novo* mutation across various organisms. In *Arabidopsis thaliana*, Ossowski et al. performed a mutation accumulation study that identified 99 substitutions and 17 indels that arose over 30 generations in five independent lineages (38), suggesting a mutation rate of $\sim 7x10^{-9}$ per base per plant generation. Human sperm sequencing has estimated a mutation rate of $\sim 2.4x10^{-8}$ mutations/base/gametic generation (39). Recently, we have developed a technique that enriches for *de novo* mutations, but does not allow calculation of precise mutation rates, and relies only on the pre-existence of a reference genome sequence and sufficient genome resequencing data from a wide diversity panel. This study identified several thousand changes, and confirmed 143 substitutions and 5 indels, that arose during the descent of Nipponbare rice (Chapter 2).

The Yugu1 cultivar of foxtail millet (*Setaria italica*) has been used to generate a full genome reference sequence (46). Because Yugu1 was subjected to a unique breeding history, its genotype is also unique. During the breeding process, it is expected that Yugu1 will have accumulated unique mutations. Jia et al. (47) resequenced several hundred *S. italica* accessions in their study of the domestication and improvement of foxtail millet in China (47). This chapter uses the reference genome sequence and resequencing data to identify Yugu1-specific mutations and assembly errors in the Yugu1 genome sequence.

Results

Since its original communication as a reference genome sequence (46), the Yugu1 genome and its annotation have gone through revisions to generate current version 2.2 (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sitalica). In our analysis, the reference sequence of the Yugu1 genome was compared to 510 resequenced *S. italica* accessions by the protocol described in Chapter 2.

In order to study the domestication and subsequent improvement of *Setaria italica* as a staple cereal in China, the Bin Han group generated several hundred low-coverage sequences for accessions of domesticated *Setaria*, wild *S. viridis*, and other lines (47). The accessions were sequenced at 0.1x to 2.13x coverage, each, and data were generated as 73 bp or 95 bp paired-end reads using the Illumina Genome Analyzer llx and Illumina HiSeq2000 platforms (47). We selected all 510 *Setaria italica* accessions from these materials, yielding raw reads that totaled ~317x coverage, as a dataset to align to the Yugu1 reference genome.

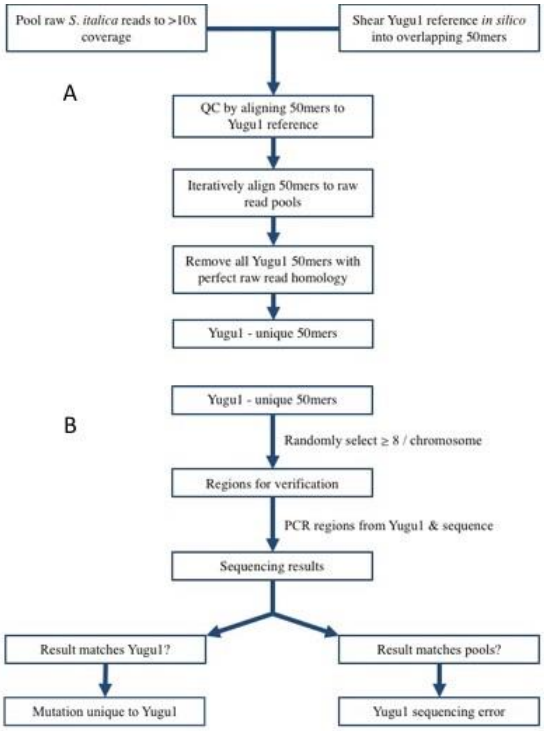


Figure 4: Similar to Chapter 2, this figure is a flowchart delineating steps required for finding novel mutations and sequencing errors in the Yugu1 line through comparative analysis to pools of other *S. italica* genomes. 4A: The *S. italica* raw read accessions from Jia et al. (47) were pooled to roughly 10x coverage per individual pool and the Yugu1 reference genome was cut *in silico* into overlapping 50mers. After, the Yugu1 50mers and *S. italica* pools were aligned in an iterative fashion and any Yugu1 50mers found to possess perfect sequence identity to the *italica* pools were discarded from the data set. We were then left with a data set of candidate Yugu1-specific 50mers. 4B: At least 8 candidate unique sequence sites from each of the 9 chromosomes in *S. italica* were selected at random for verification via polymerase chain reaction and sequencing.

The logic of this process is to find Yugu1-specific 50mers. Any 50mer that is 100% identical to any raw read in the pools is judged to be ancestral, and therefore not of interest for the discovery of *de novo* mutations. By harnessing freely available public sequencing data from prior studies, we can perform this research without the need to generate our own sequencing data or genome assemblies.

The Yugu1 reference genome was divided into overlapping 50mers and the data from Jia et al. (47) were clustered into groups (hereafter referred to as pools) that individually summed to ~10x coverage. There were 30 such pools of *Setaria italica* reads that brought the total coverage to ~317x. Yugu1 50mers were iteratively aligned to each pool using Bowtie2, and any 50mers that aligned to *S. italica* pools with a perfect 50/50 bp sequence identity were removed from the data set. The remainder of the data contained only those 50mers that potentially contained either a Yugu1-specific mutation or Yugu1 sequencing error, and the genomic coordinates for the candidate 50mers of interest were recorded. Several regions containing candidate Yugu1-specific

50mers were randomly selected per chromosome to be verified by PCR. In these verifications, Yugu1 served as the source of the template DNA. If the sequence of the resultant PCR product was found to agree with the Yugu1 assembly, it was considered a Yugu1-specific mutation. Conversely, if it did not agree with the Yugu1 reference genome sequence, but rather with the consensus sequence of the pools (see below), it was labeled a Yugu1 reference genome sequencing error (Chapter 2).

Computational comparison between *S. italica* pools and the Yugu1 reference genome 50mers resulted in 16,870 candidate Yugu1-specific 50mers being discovered. As in Chapter 2, since the 50mers overlap one another by 25 bp, the number of unique candidate regions is 8435.

Table 6: By-chromosome representation of candidates investigated by PCR/sequencing across the Yugu1 genome. The numbers in brackets represent the number of actual sequence variants or sequencing errors found in these 159 candidate Yugu1-specific 50mers.

Chromosome	# of 50mers Analyzed	# of *De Novo* Sequences Verified	# of Yugu1 Sequencing Errors
1	17	7	10
2	19	5	14
3	19	10	9
4	18	8	10
5	16	9	7
6	16	4	12
7	17	10	7
8	19	13	6
9	18	7	11
Overall	159	73 [119]	86 [116]

As in Chapter 2, primers were designed using flanking regions of 500 bp around the candidates, and predicted amplification product sizes were 100-200 bp. The primer pairs produced successful amplification products ~91% (163/180) of the time. PCR products excised

from an agarose gel were subjected to Sanger sequence analysis. Of those sequenced, useful sequences were found in ~97.5% (159/163).

In total, 73 candidate 50mers were found to have at least one Yugu1-specific variant and 86 candidate 50mers were found to have at least one Yugu1 sequencing error (Table 6). These results indicate similar rates of sequencing errors across all chromosomes, as expected for a genome that was sequenced by a whole genome shotgun strategy.

Table 7: *Setaria italica* consensus sequences were investigated at all sites deemed to be specific to Yugu1. Establishment of this presumed "ancestral sequence" allowed prediction of the nature of sequence change in the Yugu1-specific alleles. The results indicate that sequence variation in the *italica* pools is low, with the proportion of all sequence variations being attributed to minor alleles (<90%) at only 200/8435 (~2.4%).

S. italica pool consensus sequence	
# Candidate 50mers	% Read Consensus
3237	100
2940	95-99
2058	90-95
183	80-89
17	70-79

The 73 candidate 50mers verified via PCR and sequencing displayed an average of 1.6 mutations per 50mer, totaling 119 Yugu1-specific variations. The nature of the sequence change was determined by investigating the raw reads in the *S. italica* pools at each of the novel 50mer sites (Table 7) in order to find the predicted progenitor sequence. The great majority of the predicted progenitor sites in the *S. italica* pools were provided by a single allele, allowing clear assignment of the direction of the Yugu1-specific change.

Table 8: Overview of the types of Yugu1-specific changes uncovered: transitions (S), transversions (V), insertions, and deletions.

Chromosome	# confirmed Yugu1-specific sites analyzed	Transitions	Transversions	Insertions	Deletions
1	7	1	2	4	1
2	5	3	2	1	2
3	10	7	6	2	0
4	8	8	3	0	0
5	9	8	5	0	0
6	4	3	4	0	0
7	10	9	3	2	1
8	13	17	10	0	0
9	7	10	5	0	0
Total	73	66	40	9	4

Most of the substitutions were found to be transitions (Table 8). Each of the 9 insertions and 4 deletions found were single nucleotide events. Within the 20 confirmed sequence changes associated with genic regions, 6 were from coding sequences (CDS), 3 were from 5' or 3' UTRs, and 11 were from introns. Of the 6 CDS changes, the dN/dS ratio was found to be 0.2 (1/5).

As in Chapter 2, the entire dataset of 8435 novel 50mers were analyzed to determine if any candidates had lower than 70% sequence homology in their corresponding best hit to the *S. italica* pools. This would be an expected result if any novel 50mers had been produced by a large (>30 bp) indel, for instance as caused by transposable element activity. We did not find any novel 50mers of this kind, indicating an absence of recent TE excision or insertion activity.

Table 9: Number of predicted mutations per candidate 50mer.

Exact # Mutations (N) per 50mer	
N =	Mutations
1	5705
2	1634
3	717
4	271
5	83
6	15
7	10

The majority of candidate Yugu1-specific 50mers contain a single predicted sequence change (Table 9), but some have multiple variants, suggesting some mutational hotspots. We did not observe any 50mers with more than 7 predicted mutations. If mutations were fully random in their presence per 50mer, then ~6313 estimated sequence changes in 20 million total 50mers leads to a predicted of ~1 change per 333 50mers. Thus, the predicted frequency of 50mers with two changes, if they were independent, would be 1/333 times 1/333. The total number of 50mers with two events caused by random changes would thus be $1/333^2$ times the total number of 50mers, or 180. Because 1634 is much greater than 180, this provides proof that the clusters of sequence change observed are indicative of local hotspots for this change. This point is made even more dramatically by the number of observed 50mers with 3 or more mutations, which would all be calculated to be less than one in this dataset if the mutations were independent.

In Table 10, we compare the numbers of transitions, transversions, and indels found in our *Oryza sativa* ssp. *japonica* and *Setaria italica* studies to those observed in the previous mutation accumulation study conducted with *Arabidopsis thaliana.* In each case, the total number of events is low, but there were major differences in the ratios of transitions to transversions and substitutions to indels in several of these comparisons.

Table 10: Comparison of number of novel mutations found in *Arabidopsis thaliana* (4), *O. sativa* ssp. *japonica* (12), and *Setaria italica*.

Organism	Transitions (S)	Transversions (V)	S/V ratio	Insertions	Deletions	Total Mutations
A. thaliana	70	29	2.41	5	12	116
O. sativa japonica	57	86	0.66	4	1	148
S. italica	66	40	1.65	9	4	119

Figure 5 shows the distribution of the 8435 regions that did not show any 100% identity hits with the resequencing data from the pools of 510 *S. italica* accessions. The results indicate some areas completely lacking in any novel candidates and others that were quite rich in such candidates. Similar, but less detailed, unevenness was observed in the much smaller dataset derived from the confirmed Yugu1-specific alleles (data not shown).

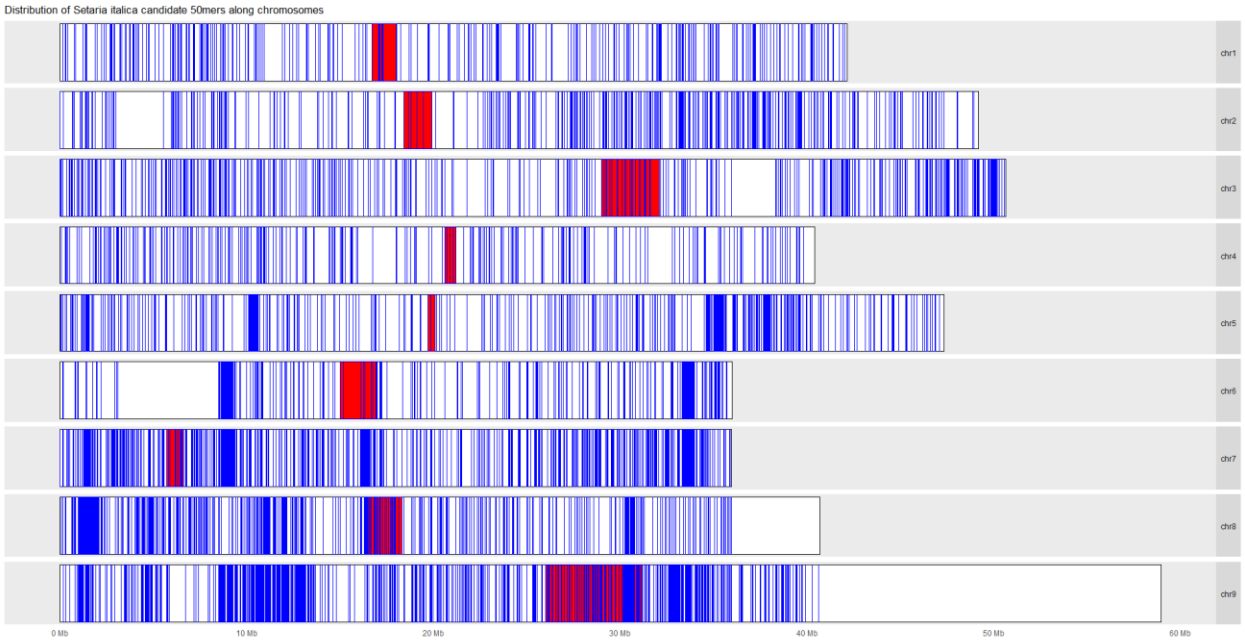


Figure 5: Karyogram of all candidate Yugu1-specific regions, shown as vertical blue lines, distributed across all foxtail millet chromosomes. Predicted centromere locations, derived from Bennetzen et al. (46), are shown in red.

Materials and Methods

Setaria italica cv. Yugu1 seedlings were pooled for DNA isolation. The experiment used seeds from the identical seed source used to generate DNA for the Yugu1 genome sequence assembly (46). DNA isolation, polymerase chain reaction (PCR) primer design, and PCR

amplification were conducted as in Chapter 2. PCR product purification was performed with Macrogen kits (Macrogen Corporation, Rockville, MD, USA) following the manufacturer's instruction. DNA sequencing was on an ABI 3730 machine. Subsequent examinations of sequences generated by Macrogen were carried out using manual verification.

The *Setaria italica cv.* Yugu1 reference genome Yugu1 was downloaded from the JGI Phytozome website (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sitalica) on June 16, 2016 and used as the foundation for this work. Pools of raw reads from *Setaria italica* accessions (47) were obtained from EBI (http://www.ebi.ac.uk/ena/data/view/PRJEB1234) on June 16, 2016. The Yugu1 reference genome was trimmed *in silico* via a custom PERL script to simulate raw reads by creating overlapping sets of 50mers along the entire genome (Figure 4), and these 50mers were iteratively aligned to raw reads from the *Setaria italica* accession pools using Bowtie2 (55) using very sensitive global alignment parameters (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml). Candidate Yugu1-specific 50mers (Figure 4) were identified as those lacking any perfect hits in the resequencing pools and were mapped back to the reference genome using BLASTn+ (54). Each 50mer with 100% identity to *S. italica* pools was excluded from the analysis via the use of SAMTools and custom scripts (56).

UNIX shell scripts, PERL, and Python scripts were written to execute the described analyses with no other external libraries or software used unless specifically mentioned above.

Discussion

Through our recently-developed protocol (Chapter 2), comparison of a high quality reference genome and resequencing low-coverage accessions such as those from Jia et al. (47) enables the prediction of reference-specific mutations and assembly errors for any given organism. As stated previously, our method harnesses freely available data from public DNA

repositories in an alternative fashion to more traditional methods of detecting recent mutation like mutation accumulation studies (37-43). The initial analysis is rapid, and only a small portion of the identified candidate 50mers need to be sequence verified in order to make robust predictions of overall mutation type patterns and genome sequence error frequencies. While the Yugu1 resequencing data were used to study domestication history and crop improvement (47), we are aware of no study that aimed to find Yugu1-unique mutations over recent evolutionary time. The shearing of the Yugu1 genome into overlying 50mers facilitated a direct comparison of Yugu1 to raw sequencing reads from other members of the *Setaria italica* species and the documentation of variants that appear to be specific to the Yugu1 lineage.

Of the 12,782 apparent sequence changes detected, ~51% were found to be errors in the Yugu1 reference genome sequence, predicting ~6,469 sequencing errors. This indicates an overall error rate in this ~500 Mb genome of 0.0013%, or 99.9987% accuracy. Hence, this independent analysis confirms the high quality of the assembly (46) and can also assist in further sequence improvement. There are limitations to this analysis, however, in that any sequence changes generated during Yugu1 descent would not be detected by our approach if they were changes to variation already present in the resequenced germplasm. This will be particularly likely within any repeats, because paralogues of genes, TEs, SSRs, and the like are all likely to contain more than one allele for any given sequence. Hence, we expect that our approach will underestimate the number of changes that originate inside repeat sequences.

A search for Yugu1-specific 50mers that had no homolog in the *S. italica* pools that was >70% identical was undertaken to identify large indels, particularly those generated by transposable element insertions or deletions. Interestingly, we did not find any such 50mers. They would be expected if any novel Yugu1-specific 50mers had been created by a large indel

(size >30bp) that inserted or removed a large portion of DNA. The lack of such a result suggests that the Yugu1 cultivar had zero TE insertion or deletion activity during its unique descent.

One possible limitation to this strategy is the degree of relatedness of the resequenced germplasm to Yugu1. If the sequences in the resequenced germplasm are very closely related to Yugu1, for instance siblings or progenitors, then insufficient time for the generation of variation would have elapsed during the derivation of Yugu1, so Yugu1-specific alleles might not exist. We find this to be more of an issue with this Yugu1 data than with the Nipponbare rice investigated in Chapter 2, as exemplified by the candidate Yugu1-specific 50mer distribution (see below). This is predicted to be completely a function of germplasm choice by the resequencing scientists. If, for instance, they had included Yugu1 in their resequencing data, then one expects that zero alleles unique to the reference genome would have been detected. If, on the other hand, the genomes chosen were all distantly related to Yugu1 (for instance, by only sharing common ancestors many millions of years ago), then the great majority of detected variants would be rare alleles under selection and the vagaries of segregation that were mostly standing variation or changes generated millions of years ago. Directly identifying the *de novo* mutations generated within the last few years in the Yugu1 lineage would be impossible in such a situation.

If the pooled *italica* accessions did not include some of the germplasm from Yugu1's ancestors, our method would detect the transmission of ancestral alleles as opposed to true *de novo* genome change. The evidence for this would be clusters of alleles unique to Yugu1 that may be a result of linkage drag on transmitted portions of chromosomes. Because we do, in fact, observe many clusters or hot spots of predicted changes in the Yugu1-specific sequence, we believe there is a nearly complete or complete lack of much of the progenitor Yugu1 germplasm

in our data set. Hence, the Bin Han resequencing study seems to have sampled only a small subset of the S. *italica* germplasm that gave rise to Yugu1 and perhaps to many other foxtail millet cultivars, despite sampling over 500 *S. italica* accessions. In contrast, the rice resequencing project by this same laboratory described in Chapter 2 seems to have chosen a much broader set of japonica rice germplasm, despite only sampling 126 accessions. Perhaps this was feasible because of the much greater pre-existing knowledge regarding rice germplasm (63,64) compared to what was known for foxtail millet.

For this *S. italica* study, the issue of whether de novo mutations were detected or not is key to the entire project, and thus deserves this more detailed and somewhat repetitive re-emphasis. In our previous study on Nipponbare rice (Chapter 2), the distribution of mutations across the genome, the dN/dS ratios of the variants inside genes, and their relative distribution in coding versus non-coding regions of genes all indicated that the great majority of the newly identified allelic variants were likely to be the outcome of *de novo* mutation during descent of the Nipponbare lineage. The distribution of candidate Yugu1-specific alleles, however, unlike the earlier Nipponbare research, showed an extreme lack of uniformity. As we predicted (but did not see) in our earlier research (Chapter 2), this is an expected outcome if the resequenced pools were very uneven in their breadth of relatedness to the reference genome species. That is, if one or more of the resequenced varieties was a very close sibling or progenitor to Yugu1, then one would expect to see few to no variants in regions of the Yugu1 genome that were passed down from the common progenitor. Conversely, if none of the resequenced varieties were closely related to one or more of the progenitors of Yugu1, then regions with large quantities of pre-existing variations would cluster in those chromosome segments derived from those progenitors. Both variant-free and variant-rich clusters are clear in the candidate Yugu1-specific allele

distribution, indicating that both some progenitors were missing from the pool and that some very closely related sibling lineages were present. These results indicate that many of the variants identified in this study, perhaps the majority, are not *de novo* mutations that occurred during the descent of Yugu1.

Although many of the ~6313 predicted Yugu1-specific sequence changes are likely to be reflections of standing variation in Yugu1's progenitors, there still should be enrichment for *de novo* mutations compared to a standard analysis of sequence variation across *S. italica* lineages. These Yugu1-specific mutations would generate rare alleles that may account for some of the properties that make Yugu1 a uniquely productive crop variety. In Nipponbare rice, we reported a dN/dS ratio of 16/10 (which is not significantly different from the 1:1 expected for *de novo* mutation) and a relatively evenly distributed amount of novel sequence changes among CDS, introns, and UTRs (Chapter 2). Current literature suggests an expected dN/dS value of between 0.28-0.47 for *Oryza* species (52), suggesting that we had enriched for *de novo* events in Nipponbare. However, in Yugu1, we found a dN/dS ratio of 1/5, suggestive of purifying selection within genes. We also observed a noted increase relative to Chapter 2 of the number of confirmed Yugu1-specific variants in intron regions, with 11 variants originating from 24.1 kb of introns (0.45/kb), 3 from 4.8kb of UTRs (0.625/kb), and 6 from 22.6 kb of CDS (0.27/kb). These two observations together suggest a bias towards the detection of standing observation in the comparison of Yugu1 to *S. italica* germplasm as opposed to the largely *de novo* genome changes that were detected in Nipponbare rice.

Of the 119 confirmed Yugu1-specific alleles, only 20 (~17%) were from regions annotated as containing genes. However, only ~14% of the Yugu1 genome (~35,000 genes, at ~2 kb/gene, divided by ~500 Mb genome) is predicted to contain genes, so this may appear to be

something of an enrichment in genic areas. Nonetheless, because repetitive DNAs will contain more variation in paralogous repeat copies, it is unlikely that Yugu1-specific changes would as often be identified in these chromosome segments compared to the lower-copy-number regions that account for most genic space.

On average, a given candidate 50mer was found to contain ~1.6 mutations, similar to the averages estimated in Nipponbare rice of ~1.6 mutations/candidate 50mer (Chapter 2). Our results indicated that 50mers with more than one change were vastly more frequent than was to be expected from an independent origin model. This indicates that origins of standing variation and *de novo* mutation involve processes that often create clusters of mutation in single events. Error-prone repair across severely damaged regions has been shown as one origin of this type of clustered mutation (30,31). The similarity of this observation for mostly standing variation in foxtail millet and mostly *de novo* mutation in rice suggests that clusters of mutation are a routine phenomenon in grass genomes, and perhaps caused by similar mechanisms occurring at similar rates.

In contrast to the similar study on Nipponbare, the observed transition to tranversion ratio in foxtail millet was greater than one, as was also observed in the Arabidopsis mutation enrichment studies (38,43). We cannot determine whether these differences are an outcome of true biological differences or of differences in the strategy for *de novo* mutation detection. The low frequencies and tiny sizes, of indels in all four studies suggest that these are routine features of mutation in many angiosperm lineages (38,43,52, Chapter 2).

This analysis extends our investigation of the nature of mutational change, and the power of our novel strategy, for assessment of plant genome variation. The depth and diversity of the resequenced accessions is apparently a key factor in the power of this strategy. The analysis

indicates that plotting the distribution of candidate variants is a powerful way to assess these characteristics of the resequenced genomes and their relationship to the reference genome source. In cases such as that for Nipponbare rice, mechanisms and types of *de novo* changes can be investigated, with thousands of *de novo* alleles available for inspection. In cases like that seen here with Yugu1 foxtail millet, rare alleles can be identified, as can segments of the genome derived from rare or unknown progenitors. In both cases, thousands of alleles are made available for further investigation, without any need for initial data generation by the investigator.

CHAPTER 4

DISCUSSION

The widespread availability of reference-genome quality sequencing data and subsequently developed lower-coverage resequencing data for many organisms permits the development and application of a new type of strategy for the study of sequence variation. Resequencing studies generally exploit and study sequence variation among populations or individuals to examine levels of change over time genetically, one example of which is crop domestication and origin studies in plant systems (45,47,52). Genetic deviation of populations or individual lines and later genetic dispersal are often projected using resequencing data in efforts to understand how modern cultivars descended. Often, resequencing data are published to public repositories and are thus available for the use of any researcher at no monetary cost. We suggest that the data could be employed for an alternative purpose: the study of the types and patterns of sequence change between higher quality reference genomes and resequencing data of a given organism. This strategy is proposed as an alternative to traditionally performed methods of detection of recent genome change like mutation accumulation studies (37-43). Our strategy is relatively fast, takes less time than mutation accumulation methods as there is no requisite time necessary to produce multiple generations of an organism for study, uses freely available data, and only a small subset of the thousands of generated results need to be confirmed via PCR and sequence verification to make strong extrapolations about an organism's overall mutation nature and errors in the reference genome sequence.

Due to heritable differences via mutation, natural selection, drift, and genetic segregation, any given member of a population will have a distinct genome sequence that no other member of

that population has. Selection is thought to act to give an advantage to particular genotypes in a particular environment. Since the first studies of segregation performed by Mendel (65), plants have been widely used to study genetic phenomena. Although some genetic variation can be generated during the process of crop breeding, the majority of sequence variations that make a given variety unique are thought to be derived from segregation. Then, these variations must already be in extant in a population. We therefore expect that any variation specific to a given cultivar (for example, which may only exist in one distinct lineage of a given crop) would be expected to contain some *de novo* mutations.

Any noted alterations between individuals or populations can be ascribed to an amalgamation of natural selection, population history, and *de novo* mutation rates and spectra. Many studies have been conducted in order to attempt to estimate the *de novo* mutation rates in some species. This is usually done by the use of a method called mutation accumulation study. Nonetheless, *de novo* mutation is infrequent due to many factors, including DNA repair pathways, and may take multiple generations to detect across a given number of lineages of a species. In addition, these studies typically involve the creation and maintenance and ensuing genomic sequencing of several generations of separate lineages, which may take years in the case of some plants (37-43). In contrast, our method can be more rapidly applied to the study of any organism with a reference genome and resequencing data for no additional cost, even if the organism has long generational times or needs multiple lines to draw accurate conclusions.

The strategy we propose requires a target organism that has a high quality reference genome sequence available and multiple resequencing accessions of the target organism to compare it against. Overall, the strength of this analysis is highly reliant on both the coverage depth and the breadth of the resequencing data, or how closely- or distantly-related the

resequencing data germplasm are relative to the reference genome sequence. We expect that if the resequencing genomic sequences are closely related to the target reference genome and deeply sequenced, we would observe only sequence changes in the reference genome that may be attributed to the unique descent of the reference genome lineage. On the other hand, if the resequencing data are from accessions that are only distantly related to the variety that was used to generate the reference genome sequence, we expect to see vast numbers of predicted changes across the genome. As such, a wider and more varied germplasm would be best for the resequencing lines.

Our strategy was used in Japanese rice (*Oryza sativa* ssp. *japonica*). The rice reference genome, Nipponbare, is among the top reference genome sequences currently extant in the plant kingdom (51). Since its original incarnation, it has undergone multiple revisions up until its current iteration of IRGSP 1.0 (44). Despite this, we know of no prior studies that aimed to find *de novo* genomic changes in the Nipponbare lineage. As highlighted in Chapter 2, our method of shearing the Nipponbare genome into overlapping 50mers enables the comparison of these Nipponbare 50mers to *Oryza sativa* ssp. *japonica* resequencing accessions (45). We discovered 17,588 candidate 50mers unique to the Nipponbare lineage which include a projected ~7400 sequencing errors in IRGSP 1.0 and ~10,200 50mers distinct to Nipponbare. This number can be halved as all candidate 50mers covered any given predicted novel sequence twice, bringing the unique number of candidates to 8,794.

We report thousands of *de novo* sequence changes in Nipponbare, a higher number compared to the data available from traditional mutation accumulation studies performed in *Arabidopsis thaliana* (38,43). In addition, all resequencing data used were available from a data repository at no cost, and the confirmation strategy for our results can be done at low cost. We

also report many cases of identified reference genome errors that could be corrected. Still, our method is not without its weaknesses. Some *de novo* genome changes will be overlooked as any potential sequence change in Nipponbare that was a polymorphism present in the *japonica* resequencing data would not be reported as a candidate 50mer as the two would have perfect sequence homology. We believe the number of these possible false negatives to be relatively low as the *japonica* resequencing pooled data had near 100% consensus on most candidate nucleotide locations in the data. In the case of repeat-rich regions, many sequence differences will be in the data and thus potentially missed. Consequently, our approach will underestimate repeat sequence variations and Nipponbare sequencing errors. If the *japonica* resequencing data was missing some of the ancestral germplasm from Nipponbare, we would expect that some of the Nipponbare-unique 50mers would result from ancestral alleles not present in the resequencing data as opposed to genuine *de novo* mutation. Because we do not observe large clusters of candidate Nipponbare-specific changes when we analyze where the predicted changes occurred over the entire genome, we do not believe this to be the case for most of our data. This lends credence to the idea that the majority of the verified Nipponbare-unique sequences are in fact resultant from *de novo* mutation that transpired during the descent of the Nipponbare line. The similar distribution of our confirmed Nipponbare mutations among gene parts (CDS, UTRs, and introns) also backs the idea that we discovered recent mutations in Nipponbare, as does our reported dN/dS ratio of 16/10.

A search for 50mers with a best hit below 70% sequence identity to the *japonica* resequencing pools was carried out to find all indels of size >30bp. These larger indels would be indicative of TE content in the unique Nipponbare cultivar. The presence of TEs would not be unexpected as it is known that most grasses do have high levels of TE activity (58,59,66).

However, we did not find any cases of TE excision or any large indels, signifying that the events that led to the descent of the Nipponbare genome did not include any recent TE activity.

The comparison of Nipponbare to the *japonica* resequencing data returned 8,794 unique candidate sites. Analysis done to verify some of these findings showed 101/194 (~52%) were errors in the Nipponbare reference sequence. The 101 50mers that had sequencing errors included 76 with exactly 1 sequencing error and 25 with between 2 to 4 errors per 50mer. These low numbers of sequence errors indicate exceptional overall genome accuracy in Nipponbare.

We confirmed 93/194 candidate 50mers to have Nipponbare-specific mutations. We report 93 verified loci, with 64 containing 1 mutation and 29 containing more than 1 mutation. We verified a total of 86 transversions, 57 transitions, 1 deletion, and 4 insertions in Nipponbare. All of the 194 50mers chosen for verification were selected randomly from the 8794 candidates to represent each chromosome in Nipponbare. The confirmed *de novo* changes found specific to Nipponbare were not observed to be intensely clustered on any chromosome, but a few of the errors were found to cluster on chromosomes that we found had the highest error rates. We attribute this to regions of the genome that were particularly difficult to sequence (44), with certain chromosomal regions having lower relative sequence quality. We detect an increased base substitution to indel ratio of 143/5 in rice relative to the numbers observed in other systems, like 99/17 in *A. thaliana* and 732/60 in *D. melanogaster* (38,42). The potential reasons for this observation will be discussed in further detail below along with comparisons from our second dataset.

We also utilized our strategy in foxtail millet (*Setaria italica*) using the reference genome Yugu1 (46). As highlighted in Chapter 3, the Yugu1 genome has been studied for domestication and historical information by the Bin Han group (47). However, we again do not know of any

study which analyzed recent Yugu1-specific genomic changes. In the same style as the analysis carried out on Nipponbare (Chapter 2), Yugu1 was split *in silico* into overlapping 50mers so that it could be directly compared with resequencing data generated from other *S. italica* lines to find Yugu1-specific variants.

We identified 12,782 unique sequence differences in Yugu1. Approximately ~51% (or 6,469) candidates were predicted to be sequencing errors in the Yugu1 reference sequence, implying 99.9987% accuracy in the Yugu1 genome. As in Nipponbare, our method is not without its faults. Any unique variations created during the descent of Yugu1 would not be found by this strategy if they already existed in the studied *S. italica* lineages. We would expect this to happen in areas with repetitive DNA, and our method would underrepresent the true number of changes from repeat regions.

Similar to Chapter 2, we looked for any possible unique Yugu1 50mers with a best hit to the *S. italica* pools of less than 70% sequence identity to elucidate any large indels (>30bp) that may be present. Such a query would find any and all evidence of TE insertions or deletions. However, we did not find any such Yugu1-specific 50mers. The absence of such Yugu1-specific 50mers implies that Yugu1 also had no TE activity through its origin.

A potential weakness to this strategy is the how dependent it is upon the breadth of the resequencing data relative to Yugu1. For example, highly related lineages such as progenitors or cousins of Yugu1 composing the majority of the germplasm may eliminate the potential for variation without enough time having passed since Yugu1's inception. If this were the case, we would expect to not discover any alleles unique to the Yugu1 line. As seen in the karyogram of the locations of candidate Yugu1-specific 50mers in Chapter 3, this is indeed more an issue in Yugu1 than in Nipponbare rice. In this way, the accuracy in potentially identifying genuine *de*

novo changes in a given target organism is reliant on a mixed-breadth germplasm. If certain lines were included or removed from the resequencing data, we would witness different outcomes. For instance, had the Yugu1 lineage somehow gotten into resequencing data, our method would reveal no variations unique to the Yugu1 line. A selection of solely distantly related *S. italica* lines would lead to most sequence changes actually being rare alleles and not *de novo* variation. We observe many clusters of predicted changes in the Yugu1 line (Chapter 3), providing evidence for the lack of most of the progenitor Yugu1 germplasm in our resequencing dataset. The resequencing study by Jia et al. (47) appears to have only utilized a small subsection of the germplasm in *S. italica* even though the depth of the coverage was around ~317x. Surprisingly, in Nipponbare rice, the Bin Han group generated a more comprehensive *japonica* germplasm in fewer sequence accessions (45).

We believe that there should be more *de novo* events specific to Yugu1 when compared to other techniques despite the fact that many of the ~6313 predicted Yugu1-specific variants are predicted to be originated from standing variation in *S. italica* germplasm. We found in Yugu1 a dN/dS ratio of 1/5, in contrast to our Nipponbare results of a dN/dS ratio of 16/10, indicating the occurrence of purifying selection within genic regions. While the Nipponbare data suggested an even assortment of predicted unique changes in gene components, in Yugu1, we see a skewing toward introns. Combining these observations supports the detection of standing variation in *S. italica* germplasm.

We report 119 verified Yugu1-specific alleles, with 20 being from gene regions. The candidate 50mers from Yugu1 had an average of ~1.6 mutations/candidate 50mer, in agreement with our results from Nipponbare of ~1.6 mutations/candidate 50mer. We also report that candidate loci with >1 predicted sequence change were noted at a higher number than what an

independent origin model would suggest. We would thus expect to see small clusters of mutations near one another derived from either processes responsible for *de novo* mutation or standing variation (16-17,26,30-31). As we detect this both in *de novo* events in rice and the rare alleles in foxtail millet, this phenomenon appears to be common to grass genomes. Generally, clusters of genomic changes can occur due to repetitive regions being prone to polymerase slippage during repair (19), indels causing frameshift mutations and an increased frequency of point mutations surrounding the indel (16), double strand breaks acting as induced hotspots (17), or high amounts of spontaneous deamination in methylated DNA regions (26).

In Yugu1, we witnessed a transition to transversion ratio of 1.65, a value that starkly contrasts with the ratio from Nipponbare of 0.66. The ratio from Yugu1 is more in line with those predicted from the literature, like the 2.4 ratio observed in *A. thaliana* (38). We observe more than double the amount of confirmed indels in Yugu1 than in Nipponbare (13 in Yugu1 to 5 in Nipponbare), with *Arabidopsis* having more than both at 17 indels confirmed. The reported numbers in plant systems are lower still relative to the 732 substitutions and 60 indels found in *D. melanogaster* (42). The differences in substitution to indel ratios in these species may be due to different relative efficiencies of DNA repair machinery or differing mutation types. As it pertains to plants, studies have shown that species may possess radically different rates of DNA repair efficiency, which may lead to novel genetic events in certain lineages (60,61). This is also the case in some human cancers (27). The lower relative amounts of smaller indels in angiosperms appear to suggest this may be a common property of angiosperm lineages (38,43,44,46).

Though our strategy is highly dependent on both the depth and breadth of the germplasm of the resequencing lineages, we present it as an alternative to traditional mutation accumulation

techniques. As observed in Chapter 2 with Nipponbare rice, if the germplasm of the resequencing data is "wide" or broad, a researcher using our strategy can successfully identify *de novo* mutations and reference genome sequencing errors at high confidence. The only associated costs are encountered when a small subset of the data is verified via polymerase chain reaction and subsequent sequencing of the PCR products, costs which are dramatically less than those associated with the creation of multiple generations and lines of a given organism for the mutation accumulation approach. The *japonica* reads sequencing depth of ~110x appears to be adequate as indicated by the results in Chapter 2, but the success of our method is highly dependent upon the resequencing data having diverse germplasm to include both related sequences and distant sequences relative to the reference genome. Our results in Yugu1 indicate that an increased depth of ~317x coverage does not guarantee successful identification of *de novo* reference-specific changes. Both variant-rice and variant-free regions are observed in the distribution of candidate 50mers from Yugu1 when plotted against the chromosomes, suggesting that only a small portion of the total extant *Setaria italica* germplasm was present in the 510 accessions used in this study. This suggests that successful usage of this strategy can be guaranteed even before variant or error validation begins by simply plotting the predicted locations of the candidate 50mers across the genome. If the plot resembles that in Chapter 3 with dramatic hot spots and interspersed cold spots, it is like a litmus test that the resequencing data being utilized are not sufficiently diverse. On the other hand, if the plot resembles the Nipponbare candidate distribution where only small clusters are observed mostly on chromosomes known to have lower sequence quality, the germplasm is indeed diverse and will be rich in accessions with a moderate degree of relatedness to the reference genome. Even with the narrower germplasm encountered in the *S. italica* lineages, several thousand unique Yugu1-

specific variants were identified as well as several thousand Yugu1 sequencing errors. With a proper resequencing dataset, the odds of successfully identifying recent genome change appear to be much higher, with thousands of *de novo* sequence changes and thousands of Nipponbare sequencing errors having been uncovered in the Nipponbare lineage. Although several thousand events of either novel variants or sequencing errors may seem high, it is vital to remember that these numbers are out of genomes with sizes of approximately 430 and 500 Mb (Nipponbare and Yugu1 respectively), indicating that these novel events are still quite rare and that both reference genomes do indeed have outstanding sequence accuracy despite the differences in how both were generated.

Unlike traditional mutation accumulation studies, we do not know exactly how far apart our target reference genomes are from their corresponding germplasm. In a mutation accumulation study like the work done by Ossowski and colleagues (38), generation-by-generation measurement of genomic difference is performed among all lineages being monitored. Typically, these lines are inbred and products of self-crossing, a limitation of the method in plant systems that is a bit of a double-edged sword in that it does allow a specific date to be placed on any given *de novo* change that arises but it would not work in systems where selfing is not permitted. Mutation accumulation studies have set a high standard for guaranteed results that are 100% accurate with the ability to pinpoint precisely which line at which generation gave rise to a novel mutation. The downside to these studies is, of course, the additional time and monetary cost associated with data generation. It may take years to generate and maintain mutation accumulation lines depending on the species being studied, and each needs to be sequenced and any changes catalogued. These studies typically yield fewer results than what we would expect from our strategy, with Ossowski et al. identifying 121 total

mutations in *Arabidopsis thaliana* in a study that required 30 generations of *A. thaliana* lines to be studied. In contrast, our strategy is able to identify several thousand potential *de novo* mutations and similar amounts of errors remaining in the reference genome sequence we target. We do not require the maintenance and creation of multiple lineages over multiple generations for our data as we use publicly available data from repositories, reducing our data generation cost to zero. As was the case for Nipponbare study, where the resequencing data were sufficiently broad and included a good overview of the germplasm for the target species, we are able to identify and confirm recent genomic change. Even in the less desirable case of Yugu1, when data are narrower and do not include a diverse range of germplasm for the target species, we are still able to identify rare alleles and reference genome sequencing errors at zero data generation costs. Additionally, our amount of verified results for recent sequence change are comparable to the numbers from the Ossowski et al. work in Nipponbare for example, with 148 confirmed *de novo* variants identified. We also uncover several thousand more candidates that could be verified as well, suggesting our method may yield higher numbers of predicted changes using a strategy that takes less time to perform. The only traditional laboratory work that needs to be done using our strategy is the verification of candidate changes via PCR and resequencing after they have been identified computationally. Thus, we believe our method is still justifiable and can be an attractive alternative to other methods given sufficient germplasm and a high quality reference genome sequence for a target organism.

In summation, we present a novel strategy of genome sequence analysis which allows the usage of a reference genome and resequencing data for a target organism to identify potential novel sequence mutations, rare alleles, and reference genome sequence errors at zero data generation cost. We do not expect that either Nipponbare or Yugu1 are average results of this

strategy as they appear to be near polar opposites of one another, indicating they may be extreme cases rather than the norm. In the analysis for Nipponbare, just the right amount of sequence and lineage diversity existed in the *japonica* resequencing pools and we were able to definitely discover and confirm recent genomic variation. In stark contrast, the presence of strong hot/cold spots in the Yugu1 to *S. italica* pools comparison indicate that our Yugu1-specific data may have not been *de novo* change but instead rare alleles due to a very narrow and similar (albeit deep) germplasm in the *italica* reads. A future study employing our strategy could prove this to be the case by generating a dataset of diverse germplasm sequence accessions and an intentionally less diverse germplasm dataset for the same organism and compare the two results. An ideal candidate for a future application of this method would be maize (*Zea mays*) as it has been bred and studied in many locations across the globe. This fact coupled with the high amount of wild germplasm make it comparable to rice and would suggest that maize germplasm would have sufficient sequence breadth to successfully identify *de novo* gene changes in its B73 reference genome (67). Future studies could also be carried out on animals or microbes because our technique can be applied to any system.

REFERENCES

1. Loewe L. (2008). Genetic mutation. Nature Education 1:113-117.

2. Pierce BA. (2000). Genetics: a conceptual approach. Freeman, New York.

3. Drake JW et al. (1998). Rates of spontaneous mutation. Genetics 148:1667–1686.

4. Eyre-Walker A and Keightley PD. (2007). The distribution of fitness effects of new mutations. Nature Reviews Genetics 8:610–618.

5. Haag-Liautard C et al. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. Nature 445:82–85.

6. Lynch M et al. (1999). Perspective: Spontaneous deleterious mutation. Evolution 53:645–663.

7. Orr HA. (2005). The genetic theory of adaptation: A brief history. Nature Review Genetics 6:119–127.

8. Loewe L and Charlesworth B. (2006). Inferring the distribution of mutational effects on fitness in *Drosophila*. Biology Letters 2:426–430.

9. Clancy S. (2008). Genetic mutation. Nature Education 1:187-191.

10. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. Nature 409:860–921.

11. Twyman R. (2003). Mutation or polymorphism? Wellcome Trust website, http://genome.wellcome.ac.uk/doc_WTD020780.html

12. Sandelin A et al. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5:99-110.

13. Ma J and Bennetzen JL. (2004). Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of the Sciences 101:12404-12410.

14. Viguera E et al. (2001). Replication slippage involves DNA polymerase pausing and dissociation. EMBO Journal 20:2587–2595.

15. Pearson CE et al. (2005). Repeat instability: mechanisms of dynamic mutations. Nature Reviews Genetics 6:729–742.

16. Tian D et al. (2008). Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455:105-108.

17. Shee C et al. (2012). Two mechanisms produce mutation hotspots at DNA breaks in *Escherichia coli*. Cell Reports 2:714-721.

18. Seidl H et al. (2001). Ultraviolet exposure as the main initiator of P53 mutations in basal cell carcinomas from psoralen and ultraviolet A-treated patients with psoriasis. Journal of Investigative Dermatology 117:365–370.

19. Clancy S. (2008). DNA damage & repair: mechanisms for maintaining DNA integrity. Nature Education 1:103-109.

20. Lodish H et al. (2004). Molecular biology of the cell, 5th ed. New York, Freeman.

21. Greenblatt MS et al. (1994). Mutations in the P53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. Cancer Research 54:4855–4878.

22. Sinha RP, Häder DP. (2002). UV-induced DNA damage and repair: a review. Photochemical and Photobiological Sciences 1:225–236.

23. Griffiths AJF et al. (2000). An introduction to genetic analysis, 7th Edition. New York, Freeman.

24. Denissenko MF et al. (1996). Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. Science 274:430–432.

25. Sitruk LS et al. (2010). Unknown gonadotoxicity chemotherapy and preservation of fertility: example of temozolomide. Gynecologie Obstetrique & Fertilite 38:660-662.

26. Walser JC, Furano A. (2010). The mutational spectrum of non-CpG dna varies with CpG content. Genome Research 20:875–882.

27. Pray L. (2008). DNA replication and causes of mutation. Nature Education 1:214-220.

28. Branze D, Foiani M. (2008). Regulation of DNA repair throughout the cell cycle. Nature Reviews Molecular Cell Biology 9:297–308.

29. Gorbunova V et al. (2007). Changes in DNA repair during aging. Nucleic Acids Research 35:7466–7474.

30. Waters LS et al. (2009). Eukaryotic translesion polymerases and their roles and regulation in DNA damage tolerance. Microbiology and Molecular Biology Reviews 73:134–154.

31. Colis C et al. (2008). Mutational specificity of gamma-radiation-induced guanine-thymine and thymine-guanine intrastrand cross-links in mammalian cells and translesion synthesis past the guanine-thymine lesion by human DNA polymerase eta. Biochemistry. 47:8070–8079.

32. Mulligan LM et al. (1993). Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A. Nature 363:458–460.

33. Mimitou EP, Symington LS. (2009). Nucleases and helicases take center stage in homologous recombination. Trends in Biochemical Sciences. 34:264–272.

34. Guirouilh-Barbat J et al. (2004). Impact of the KU80 pathway on NHEJ-induced genome rearrangements in mammalian cells. Molecular Cell 14:611–623.

35. McVey M, Lee SE (2008). MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends in Genetics 24:529–538.

36. Decottignies A. (2013). Alternative end-joining mechanisms: a historical perspective. Frontiers in Genetics 4:1-16.

37. Behringer MG, Hall DW. (2016). The repeatability of genome-wide mutation rate and spectrum estimates. Current Genetics 62:507-512.

38. Ossowski S et al. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science 327:92-94.

39. Wang J et al. (2012). Genome-wide single-cell analysis of recombination activity and *de novo* mutation rates in human sperm. Cell 150:402-412.

40. Lee H et al. (2012). Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proceedings of the National Academy of the Sciences 109:2774-2783.

41. Keightley PD et al. (2014). Estimation of the spontaneous mutation rate per nucleotide site in a *Drosophila melanogaster* full-sib family. Genetics 196:313-320.

42. Schrider DR et al. (2013). Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. Genetics 194:937-954.

43. Jiang C et al. (2014). Environmentally responsive genome-wide accumulation of *de novo Arabidopsis thaliana* mutations and epimutations. Genome Research 24:1821–1829.

44. Kawahara Y et al. (2013). Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice 6:1-10.

45. Huang X et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501.

46. Bennetzen JL et al. (2012). Reference genome sequence of the model plant *Setaria*. Nature Biotechnology 30:555-561.

47. Jia G et al. (2013). A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). Nature Genetics 45:957-961.

48. Aminetzach YT et al. (2005). Pesticide Resistance via Transposition-Mediated Adaptive Gene Truncation in *Drosophila.* Science 309:764–767.

49. Burrus V, Waldor MK. (2004). Shaping bacterial genomes with integrative and conjugative elements. Research in Microbiology 155:376–386.

50. Roach JC et al. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328:636–639.

51. International Rice Genome Sequencing Project. (2005). The map-based sequence of the rice genome. Nature 436:793-800.

52. Zhang Q et al. (2014). Rapid diversification of five *Oryza* AA genomes associated with rice adaptation. PNAS 111:4954-4962.

53. Vaughn JN et al. (2014). Whole plastome sequences from five ginger species facilitate marker development and define limits to barcode methodology. PLoS ONE 9:e108581.

54. Camacho C et al. (2009). BLAST+: Architecture and applications. BMC Bioinformatics 10:421.

55. Langmead B, Salzberg SL. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.

56. Li H et al. (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

57. Darwin C. (1868). The variation of animals and plants under domestication. John Murray 1.

58. Rinehart TA et al. (1997). Comparative analysis of non-random DNA repair following Ac transposition excision in maize and Arabidopsis. Plant J 12:1419–1427.

59. Bennetzen JL. (2007). Patterns in grass genome evolution. Curr. Opin. Plant Biol. 10:176-181.

60. Vitte C and Bennetzen JL. (2006). Analysis of retrotransposon diversity uncovers properties and propensities in angiosperm genome evolution. Proc. Natl. Acad. Sci. USA 103:17638-17643.

61. Bennetzen JL and Wang H. (2014). The contributions of transposable elements to the structure, function and evolution of plant genomes. Ann. Rev. Plant Biol. 65:505-530.

62. Bertram JS. (2000). The molecular biology of cancer. Molecular Aspects of Medicine 21:167–223.

63. Caicedo AL et al. (2007). Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 3:e163.

64. Molina J. et al. (2011).Molecular evidence for a single evolutionary origin of domesticated rice. PNAS 108:8351–8356.

65. Iltis, H. (1958). Gregor Mendel and his work (1943). Shapley, H. et al. A Treasury of Science. New York: Harper.

66. Ma J, Bennetzen JL. (2004). Rapid recent growth and divergence of rice nuclear genomes. PNAS 101:12404-12410.

67. Schnable P et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112-1115.