2022-04-12

Implementation of a Genome-Wide Survey of Induced Mutations to Identify Agronomically Valuable Variants in Chenopodium quinoa

Andrew Alarcon Parker
Brigham Young University

Implementation of a Genome-Wide Survey of Induced Mutations to Identify Agronomically Valuable Variants in *Chenopodium quinoa*

Andrew Alarcon Parker

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

David Jarvis, Chair
P. Jeff Maughan
Eric N. Jellen

Department of Plant and Wildlife Sciences

Brigham Young University

ABSTRACT

Implementation of a Genome-Wide Survey of Induced Mutations to Identify Agronomically Valuable Variants in *Chenopodium quinoa*

Andrew Alarcon Parker
Department of Plant and Wildlife Sciences, BYU
Master of Science

Quinoa has been utilized for millennia in the Andes region of South America as a nutritious and hardy food crop. In recent years interest in quinoa has grown as need increases for an alternative to traditional cereal crops that can tolerate marginal environments while offering superior nutrition. Growers outside the Andes have experienced several complications adopting quinoa, including undesirable secondary metabolites, poor yield, lodging, and height inconsistency. Unfortunately, access to native ecotypes for crop improvement is limited, and desirable traits are difficult to introduce into available quinoa cultivars because of its allotetraploid genome and tendency to self-pollinate. A genome-wide survey of induced mutations in 244 sequenced M2 families was created from a bank of EMS-treated quinoa seeds and assembled into a library of mutant lineages with predicted variants and their effects on genes to assist in identifying agronomically valuable mutations in target genes as a supplement to crop improvement efforts. Using this library, eight families containing mutations in genes associated with reduced height – GAI1, GA20OX, GID1, and LE – were identified. Several individuals exhibited a shorter than average phenotype; however, because each family contains thousands of EMS-induced mutations, the causative mutation of the reduced height phenotype in each family could not be definitively identified. In one family, absence of the GAI1 mutant allele, but the presence of a mutant CKX3 allele, provided a correlation between a mutation and the short phenotype. Genotyping each generation would be required for a targeted mutant allele to be tracked through selection.

Keywords: Chenopodium, quinoa, TILLING, EMS, WGS, Illumina, phenotypic screen, GAI1, CKX3, SLY1, MAPS, VEP, forward genetics

ACKNOWLEDGEMENTS

I would like to thank David Jarvis who has worked diligently as my advisor as well as P. Jeff Maughan and Eric N. Jellen who graciously inducted me into their circle of genetic knowledge. I also acknowledge Brian Cox who built the foundation of this project and the small army of undergraduates that were essential to its progress. I also thank my parents, especially my mother, who have guided and motivated me to continue my education and seek for excellence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Introduction

Quinoa as a Food Resource

Pseudocereals are excellent candidates for more sustainable staple crops. They are generally defined as "fruits or seeds of non-grass species that are consumed in very similar way as cereals having nutritive value very much competitive to conventional crops" (Das, 2016). *Amaranthus sp*. (amaranth), *Fagopyrum esculentum* (buckwheat), and *Chenopodium quinoa* Willd. (quinoa) have become the primary crop species under this definition. They are particularly valuable candidates for more sustainable staple crops because of their resilience in marginal environments and their excellent nutritive value (Choukr-Allah, 2016; Rodriguez, 2020). Quinoa has gained special focus and praise as a super food thanks in no small part to the International Year of Quinoa announced by the Food and Agriculture Organization (FAO) of the United Nations (UN) in 2013 (Bazile, 2015a). Quinoa has been used in the Andes as a reliable, nutrient rich crop for several thousand years (Risi, 1984). Although other Chenopods like *C. pallidicaule* were also semi-domesticated under the native name cañihua or kañiwa, quinoa was slowly selected as the "mother grain" by the Inca. It requires minimal input to cultivate and has shown a great degree of tolerance to different soils and climates including high saline, low moisture, and low nutrient (Bhargava, 2006a). This makes quinoa an excellent crop for marginal areas and times of food scarcity.

Quinoa has a history that makes it particularly difficult to manipulate but provides a diverse selection of ecotypes. The origin of quinoa is the result of a genome duplication event in which two diploid species of *Chenopodium* crossed to result in a disomic allotetraploid hybrid (Bhargava, 2006b; Jarvis, 2017). The product of this event in quinoa is a combination of two genomes into one: genome A from one progenitor and genome B from another. These homeologous genes in the quinoa genome can mask mutations by compensating for a loss of function in the mutant allele (Ward, 2000). Despite this added complexity many variations in phenotype still exist and there is evidence that certain desirable characteristics were selected for by humans over the millennia: greater seed diameter, thinner and smoother seed coats, and more compact inflorescences (Bruno, 2014). Quinoa's natural tolerance to many adverse conditions comes from adaptation to the diverse environmental pressures it faced evolving in the Andes (Bazile, 2015b). Together natural and human selective pressures have led to a great diversity of physical characteristics including variations in seed protein content, stem diameter, plant height, plant color, and grain color and shape (Rojas, 2015). Unlike other globally adopted crops quinoa has maintained this great degree of agrobiodiversity and was not included in the "green revolution" that facilitated the industrialization of agriculture but reduced the diversity of cultivated varieties, leaving these more widely adopted crops more vulnerable to disease and less resilient to adverse conditions (Khush, 1999; Ruiz, 2014*)*.

Improving Quinoa for Modern Agriculture

Starting in the 1960s modern quinoa breeding programs centered around the countries of quinoa's origin through FAO funding, but have since become more international (Gandarillas, 1979; Bonifacio, 2015). Characteristics were targeted for development including early maturity, improved drought tolerance, improved nutritional quality, improved harvest index, disease resistance (especially to downy mildew to which quinoa is particularly susceptible) and lower saponin concentration (Zurita-Silva, 2014; Bonifacio, 2015). Saponins are a group of glycosides produced by plants as a form of defense against biotic and some abiotic stresses (Szakiel, 2011). Quinoa varieties with high saponin levels show improved resistance to downy mildew but are less palatable and more difficult to digest (Danielsen, 2003; Savage, 2016).

There is a great amount of allelic diversity within quinoa populations cultivated in the Andes, but they are not all available for international use. The value of the quinoa market is expected to continue to double by 2025 making it a lucrative investment for many countries, with total quinoa seed production currently dominated by Peru and Bolivia (FiorMarkets, 2019). The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization arose from the United Nation (UN) Convention on Biological Diversity in 2010 and guarantees that local and indigenous communities can benefit from their genetic resources (Buck & Hamilton, 2011). In practice this has meant that access to certain biological resources is restricted to their countries of origin and in this case safeguards the quinoa production market for the Andean countries while providing a framework for controlled germplasm sharing (Bazile et al., 2016). Researchers attempting to adapt quinoa to different environmental conditions outside of the Andes region must currently do so with access to limited allelic variation.

Increasing allelic variation through mutagenesis to generate varieties with better characteristics is an alternative to searching for and breeding in those characteristics from existing cultivars. Traditional or induced mutagenesis introduces random allelic variation through radiation or chemical mutagens while site-directed mutagenesis uses a nuclease like TALEN or the CRISPR-Cas9 complex to interact with a specific sequence to induce a mutation and has become a more common approach (Wieczorek & Wright, 2012). Although site-directed mutagenesis is a more precise way to introduce a mutation, it requires a greater understanding of the plant's genome and the development of effective transformation and tissue-culture protocols (Lopez-Marques, 2020; Imamura, 2018). The benefit is that mutations can be designed before they are performed, meaning every change is more predictable than the random effects of traditional mutagenesis. In terms of safety, the number of off-target effects using site-directed mutagenesis is much lower than those caused by traditional mutagenesis (Holme, 2019). Traditional mutagenesis, and even traditional breeding, has the benefit of being much less technically complex than site directed mutagenesis but has less control over where variation is introduced. Additionally, once mutations are induced the subsequent generations are selfed and screened for valuable phenotypes which often necessitates massive populations (Jankowicz-Cieslak, 2015).

TILLING for Crop Improvement

TILLING (Targeting Induced Local Lesions IN Genomes) is a hybrid more specific than common approaches to traditional mutagenesis in that once lines of mutants have been established, they are screened for mutations in a gene of interest first and not for a phenotype (McCallum, 2000). This screen is done by amplifying labeled homoduplexed and heteroduplexed target genes in young M2 plants. Using an endonuclease (CEL I) that cleaves mismatched pairs at the 3' end identifies heteroduplexes which can then be sequenced to identify the exact single-nucleotide polymorphism (SNP) in the candidate gene (Barkley, 2008). This is a high-throughput method that has been used successfully for Arabidopsis to generate a "TILLING library" of mutant populations that can be accessed for further analysis and breeding efforts (Henikoff, 2004). An improvement on this method is the removal of the endonuclease selection screen to map mutations and instead, sequencing and analyzing the entire population (Mo, 2017). This method relies more heavily on bioinformatics but omits the need for amplification of every gene in the genome and allows for the selection of multiple genes of interest from within a single mutant family, generally streamlining the process. A TILLING with next-generation sequencing (NGS) approach was used previously by Mestanza et.al., 2018 to induce and identify mutations in the genes encoding acetolactate synthase (AHAS) in quinoa. While the associated herbicide resistance from a mutation in this gene was not observed, it demonstrates that TILLING can be used on quinoa to induce mutations in specific genes.

This current work demonstrates the utility of using a TILLING library generated by NGS to reveal desirable phenotypes based on mutations in targeted genes identified in a population of EMS mutants. The phenotype of interest for this study was dwarfing/reduced plant height. This decision was made with agronomic importance in mind since lodging and lack of height uniformity are both challenges reported by quinoa growers (Lopez-Marques, 2020). In other crops, reduced height has been observed alongside mutations in genes associated with the plant growth hormone gibberellin (Peng et.al., 1999). These genes are Gibberellin Insensitive (GAI1)(Koornneef, 1985) and Gibberellin Insensitive Dwarf (GID)(Griffiths, 2006) which are both involved in the gibberellin response pathway; and Gibberellin-3-Oxidase (GA3OX)(Lester, 1997) and Gibberellin -20-Oxidase (GA20OX)(Rieu, 2008) which are involved in gibberellin biosynthesis.

Methods & Materials

Developing an EMS Library for Quinoa

Quinoa mutant families were previously generated by treating seeds from the quinoa variety QQ74 (PI 614886) with ethane methylsulfonate (EMS) (Cox, 2020). Those seeds that survived the mutagenesis were grown to maturity and their seed was collected and stored. The seed from a total of 5030 families formed a seed bank of EMS families that could be sequenced and included in the genetic resource library.

Generating the Library

Previous sequencing results indicated that an average of 10 mutant alleles could be identified for each of the genes in the quinoa genome by performing whole-genome sequencing of 768 EMS families. These families were randomly selected and grown in groups of 5-6 M2 individuals per family. Plants were grown in 3.5-inch square pots in a standard flat with 12-18 plants per flat. The substrate used was a simple all-purpose potting mix (80-90% sphagnum peat moss and perlite) with little to no supplemental nutrients. All plants were hand watered either from above, by showering them from overhead, or beneath using a capillary mat. During growth, young leaf tissue from each plant was excised and stored at -80° C in 1.7-ml microcentrifuge tubes for sequencing. When the plants were mature, M3 seed was collected from all M2 plants, pooled by family, cleaned, and stored in coin envelopes in a dark, dry location. Tissue harvested from the M2 plants was freeze dried using the Labconco FreeZone Bulk Tray Dryer at -50° C and 0.00 mbar. The dried tissue was pulverized by adding two 3-mm glass beads to each tube and shaking the tubes with the SPEX Sample Prep 2010 Geno/Grinder at 1200 rpm for 45 seconds. The pulverized tissue was then pooled by family, by weight, at approximately 10 mg of tissue per individual.

DNA Extraction and Whole Genome Sequencing

The DNA of these pooled samples was extracted following a revised protocol from Sambrook (Sambrook et.al., 1989) with modifications described by Todd and Vodkin (Todd and Vodkin, 1996; for the complete protocol see Appendix I). Pulverized and pooled tissue was treated in 600 µl of warmed Extraction Buffer (100 mM Tris, 500 mM NaCl, 50 mM EDTA, 9 mM phenanthroline, 28 mM sodium dodecyl sulfate, 8 mM 2(Β)-mercaptoethanol at 65° C) for 25 minutes after being vortexed thoroughly to suspend and hydrate the tissue. 200 µl of 5 M KOAc were added and the samples were incubated on ice for 15-20 minutes. A phenol:chloroform:isoamyl (PCI) extraction was performed using 600 µl of presaturated phenol./chloroform/isoamyl alcohol (25:24:1) followed by a Sevag extraction using 500 µl of Sevag solution (chloroform:isoamyl alcohol; 24:1). 40 µl of 3 M NaOAc was then added and the entire solution was transferred to new tubes containing -20° C isopropanol. This was allowed to rest in a -20° C freezer for at least 30 minutes and up to 24 hrs. The precipitated DNA was pelletized using an Eppendorf 5417 R benchtop centrifuge and washed with 75% ethanol before being dried in a Savant SPD 1010 SpeedVac Concentrator system. The dried pellets were resuspended in 200 µl of 1X Tris-EDTA (10mM Tris, 1mM EDTA) with 0.1 mg/ml RNAse A. DNA samples were left to thoroughly resuspend overnight before purity was tested using Nanodrop and DNA concentration measured using the Qubit Flurorometer Broad-Range test kit. DNA samples were stored at -80° C and then sequenced by Novogene Corporation with Illumina 150-bp paired-end whole-genome sequencing using a seqWell plexWell LP 384 library prep kit to prepare the libraries. The seqWell library prep kit prepares two libraries of 48 individual samples for sequencing which are processed simultaneously.

Bioinformatics Pipeline

The finished reads were passed through the variant-detection pipeline (Figure 1; for full protocol see Appendix II). The raw reads were trimmed using trimmomatic with leading edge (LEADING) and trailing edge (TRAILING) set at 20 with a sliding window (SLIDINGWINDOW) of 5:20 and a minimum length (MINLEN) of 75. Trimmed reads were then mapped to the quinoa var. QQ74 version 2 reference genome with bwa_mem using the paired-end results only. The mapped reads were converted to BAM format using samtools view, and duplicate reads were removed by first filling in mate coordinates with samtools fixmate and then with samtools markdup with options -s to print basic stats and -r to remove duplicates. The BAM files were sorted after each step using samtools sort with option -n.

The sorted BAM files were given to a workflow of custom Python scripts called the Mutation And Polymorphism Survey (MAPS) designed by the Comai Lab at UC Davis for working with EMS-type mutations in TILLING libraries. Within the workflow the BAM files were compiled together using mpileup with map quality cutoff --mapqual of 21, a base quality cutoff --basequal of 21, and a maximum depth cutoff --maxdepth of 4000. The mpileup files were parsed before being sent through maps1 by mpileup-parser. The maps1 script takes the parsed mpileup data and generates a list of possible mutation sites based on differences to the reference – and between samples – in the mpileup file that can be sent to maps2 which uses a more stringent set of parameters to produce the final output of putative mutations (Lieberman & Henry, 2012). The program maps1 was run with a minimum depth -c of 6 and a maximum depth -C of 1000, a minimum percentage for a heterozygous position -b of 80 and a minimum library size -I of 10 with the options -m and -H which set the mode to mutant detection and includes heterozygous positions into the output file. The program maps2 was run with a minimum percentage for a position to be heterozygous of -p 10 and the mutant detection mode -m. All other options were left as default for both programs.

The Ensembl Variant Effect Predictor (VEP) was used to categorize the effects of the identified mutations. VEP compares the mutant position to a reference GFF file to define the affected gene and the predicted effect of the variant (McLaren et.al., 2016). The maps2 output file was modified to fit the VEP input format and then run through VEP. The quinoa var. QQ74 version 2 reference GFF was used to describe mutant positions. VEP uses The Sequence Ontology (SO) website to define variant type. For a full description of VEP variant definitions see:

https://uswest.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

When BAM files were merged from multiple sequencing runs of the same batch of samples the batch was chunked into libraries of 12 families each. This increased the number of false SNPs called by MAPS, leading to the same SNP position getting called in multiple families, which had to be removed manually before getting incorporated into the final library. The final variant library includes the variant position, the gene associated with the variant if relevant (not applicable for intergenic variants), the nucleotide and amino acid changes, the variant type (missense, synonymous, intergenic, etc.), whether the variant is found on the plus or minus strand, the predicted impact, a brief description of the putative gene function (from the GFF), and the EMS family harboring the variant (Table 1).

Finding Mutant Lines with Desirable Characteristics

During the initial development of the mutant detection pipeline 52 EMS families were analyzed using a combination of whole exome sequencing (WES) and whole genome sequencing (WGS) and they were the first families to be included in the library of predicted variants. Using the library, eight families from this original group of 52 were chosen to screen for a short phenotype based on the presence of missense and stop-gained mutations in genes related to the gibberellin biosynthesis and regulatory pathways. A phenotypic screen was started the second week of March 2021 by growing 24 M2 individuals from each candidate family in the BYU Life Science Greenhouse in Provo, UT. Six individuals from each family were grown on a capillary mat as part of a separate investigation, while the remaining 18 were grown in trays. Therefore, plants were labelled as either M# or T# for mat or tray respectively as a separate designation from their generation (labelled Rh, “reduced-height”, to separate them from other M2 plants). All plants were grown in all-purpose potting media with 2/3-oz per gallon of slow-release fertilizer contained in square 3.5 in. pots. Plant heights were measured one-month post-germination (PG) from the base of the stem to the bottom of the panicle using a metric measuring tape. This was done instead of measuring internode length because it offered better resolution between measurements. A selection of plants with a significantly short phenotype was separated from the group to reduce the chance of outcrossing and grown to maturity so their seed could be harvested. One family in particular (EMS88) was progressed further to generate a homozygous line because it demonstrated an extreme phenotype. Short plants from EMS88 were allowed to self for two generations - labelled F1 and F2 to differentiate them from the pooled, un-selected M3 and M4 generations. All plants from the F1 generation that demonstrated shorter growth compared to a wild-type control were selfed and perpetuated to the F2 generation. Tissue from short-phenotype plants in this generation was collected and the gene of interest was amplified using PCR and analyzed using Sanger sequencing to test the presence of the expected mutant allele and to identify the potential causative allele when the expected was absent.

Results

The EMS Library and Preliminary Analysis

The seed from a total of 5030 families now form a seed bank of EMS families that can all be sequenced and included in the variant library. From this seed bank, 52 families were sequenced by a preceding investigator, 43 using whole exome sequencing (WES) and 9 using whole genome sequencing (WGS) and were used to determine the WES and WGS mutation rates. Using the WES mutation rate of 11.35/Mb as a minimum and the WGS mutation rate of 21.67/Mb as a maximum approximately 300 of the mutant families would be required to achieve a goal of 4-10 mutant alleles per gene for every gene in the quinoa genome. Using the data in the genetic resource library from WGS, a more informed estimate of 752 families were needed to reach the highest goal of 10 mutant alleles per gene based on an initial average of 994 mutant alleles per family and a 73% chance a mutation in the coding region of the gene will be moderate impact mutation or higher (missense, start lost, stop gained, and stop lost), and the gene product will likely be altered (Cox, 2020). The basic utility of the library was demonstrated by correlating a change in hypocotyl color to a mutation in the betalain biosynthesis gene CYP76AD1-1.

Generating the Library

Cost of Sequencing

Based on sequencing of the first 52 families, it was decided to use WGS for the remaining EMS families for two reasons: first, WGS provides data for all SNPs in the genome including those in potential regulatory elements within the intergenic regions; and second, WGS is cheaper costing only $16.57/Gb of sequencing data compared to $61.67/Gb for WES. The costliest aspect of WGS is library preparation, which costs $100 per sample when performed by Novogene. After some investigation into available library prep kits, the seqWell plexWell LP 384 library prep kit was chosen to use as an alternative. The seqWell kit contains 4 96-well PCR plates to prepare 8 libraries of 48 samples each, or a total of 384 samples, for a per sample cost of $13.80 which is considerably lower than the quoted price at Novogene.

Impact of Increased Duplication Rates

PCR amplification is a necessary part of library preparation for NGS but results in redundant duplicates which reduce true sequencing coverage (Bansal, 2017). Part of the variant detection pipeline is the removal of duplicate reads since the inclusion of duplicates can impair variant detection. The seqWell library prep kit is optimized for 20 million read pairs – with 300 bp per read pair – yielding 6 Gb of total sequencing data. Given the quinoa genome is 1.45 Gb, producing the recommended number of sequencing reads for each sample would only provide approximately 4X coverage, which is half the coverage achieved for the original 52 families.

Comparing the 9 families sequenced by WGS, from the original 52 families, using libraries prepared by Novogene against a batch of 96 families sequenced using the seqWell kit at the same sequencing depth of 10X coverage shows that after duplicate removal seqWell prepped samples have a true average read depth of approximately 4X, ~50% that of the Novogene prepped samples at approximately 8X coverage (Figure 2). One from the original 9 WGS families (EMS2427) was included in the batch of 96. The 4X coverage EMS2427 sample was shown to have roughly the same number of SNPs called as the 8X sample – with 21938 called variants in the 8X and 22862 called variants for the 4X

– with 15876 variants common between them (Figure 3). Comparing the sequencing depth for both 8X and 4X samples of EMS2427 at SNPs called in the 8X sample and SNPs called in the 4X sample revealed that when a SNP was not called in one it was at a lower read depth than the other (Figure 4). These results are likely attributable to differences in sequencing depth between samples for a given site and MAPS read depth cutoffs rejecting SNP sites with low depth. To obtain a complete survey of SNPs in each EMS family sequencing coverage would need to be increased for every base pair in the genome to avoid missing any SNPs.

This would be burdensome to do with over 700 samples and increasing coverage using the seqWell library prep kit was not necessarily a good choice. Coverage could be increased by producing double the number of read pairs and therefore double the sequencing product. SeqWell explained that while increasing the number of read pairs would increase the size of sequencing data returned e.g., 6 Gb of data to 12 Gb of data, the rate of duplicate reads would also double. This means that the actual return in coverage would be less than double, seqWell estimated it to be about 75%. So instead, it was suggested to double the number of libraries per sample by preparing two kits for each batch of 96 families, or using special kits with double the reagents, to double the coverage. Since we already had BAM files from one kit, the simplest option was to run another kit using DNA from the same samples, process out duplicates, and merge the two BAM files together. There was a roughly 50% increase in the number of variants called using two kits instead of one, with an average of 20210 in common (Figure 3). Again, this is likely due to an increased read depth at each site. While this additional information is valuable, because the cost is doubled while the resulting increase in called variants is only 50% it would not be cost-effective to sequence each sample twice since the goal of this project is to generate a survey of induced mutations and not necessarily an exhaustive analysis of each mutant family.

Distribution of Variant Impact on Gene Function

Given these results, we decided to proceed with the sequencing of 768 EMS families (621 families after the first 52 and the first batch of 95) using a single seqWell kit per batch of 96 families, generating approximately 4X coverage for each family. The first two batches of families processed with the seqWell kit had to be chunked into sections of 12 BAM files each to overcome computing memory limitations. This introduced errors in MAPS which compares SNP positions within a population to help determine their veracity. A smaller population leads to some SNPs getting called in multiple families that were separated into different populations through chunking the data. This means that out of 592436 total entries for the first two seqWell batches only 484318 were unique positions. From those 413397 positions had a single entry in a single family, while 51722 had multiple entries in the same family from the way VEP identifies and categorizes variants based on proximity to genomic features in both plus and minus sense. Some positions, 19199, were called in multiple families and are not true SNPs; they were removed from the library.

After the first 244 families had been included in the library, with false SNPs removed, a wide distribution of variants was revealed. An average of 2524 variants were called in each family including what VEP defines as intergenic – or a variant located between genes – with a total of 315559 intergenic variants out of 618612 total variants. An average of 1242 variants were called per family as genic but with a min of 6 variants in a family and a max of 14801 (Figure 5). Many different types of variants were also called in each family with the most abundant types being upstream and downstream variants (Figure 6), which are intergenic variants within 5 kb of a genomic feature. Each variant must be characterized experimentally to truly understand its impact on gene function, but VEP provides an estimate based on variant type (Table 2).

The impact of each non intergenic variant varied greatly (Figure 7) with the majority, 84%, of variants being classed by VEP as MODIFIER because they have an impact that is harder to predict or have no impact on gene function and are in the intron, the 3' and 5' untranslated region (UTR), and upstream/downstream of the gene. LOW impact variants made up ~5% and include mutations and include synonymous variants, splice region variants, and start/stop retained variants. MODERATE impact variants, ~10%, include all missense and missense splice region mutations. HIGH impact variants, ~1%, are frameshift, stop/start gained/lost and their splice region variants, and splice acceptor/donor variants and are the least abundant (Figure 7). VEP defines variant type based on definitions from The Sequence Ontology which provides a controlled vocabulary for the description of DNA sequences annotated by GFF3 while the putative impact is considered a subjective analysis (McLaren et al., 2016).

Having a wide range of mutations per family is desirable because it provides a balance of having many mutations with which to find a desired phenotype while also not having an overabundance of background mutations in too many EMS families. Having too many background mutations may be responsible for potential difficulties in downstream applications like mutant characterization.

Targeting Variants with Desirable Phenotypes

Phenotypic Screen

While the library provides interesting information about what mutations were induced in the EMS families its greatest purpose is the targeted characterization of mutations in specific genes. Reduced height is an agronomically valuable trait in quinoa since it has been reported to increase yield and reduce issues with lodging and pre-harvest sprouting (Katwal, 2020; Khush, 1999). To test the utility of the library for the identification of variants that cause useful phenotypes like reduced height, EMS families with putative mutations in genes related to plant height were included in a phenotypic screen. Genes related to stem elongation and plant growth were searched for in the mutant library to identify EMS families that may have reduced height. In the original 52 families it was found that eight had missense or stop-gained mutations in four genes associated with plant height in the gibberellin biosynthesis and regulatory pathways (Table 3). Once the candidate families were identified 24 M2 individuals from each were grown in the BYU Life Science Greenhouse in 3.5 in. pots.

After one-month post germination (PG) the length of the plant stems was measured, and results were analyzed using a Student's T-test against the wild-type control. All families but EMS1679 contained individuals that were significantly shorter than the average length of the wild-type control group (Figure

8). One family, EMS88 with a missense mutation in the GAI1 gene CQ040570 on chromosome 8 of the A genome, had an extreme phenotype (Figure 9) . Short individuals of EMS88 had deformed leaves, and more compact growth habits. Once fully mature, the shortest individuals from family EMS88 were allowed to self and progress to what was labeled the F1 generation - to avoid confusion with the pooled un-selected generations of M3 and M4. All F2 progeny from one F1 individual were dwarf (Figure 10a), whereas the other F1 individuals either partially or completely lost the dwarf phenotype in their progeny (Figure 10b).

Identifying a Causative Allele

Individuals from family EMS88 had complex genetic backgrounds harboring thousands of mutant alleles that make identifying the causal allele difficult. GAI1 is a DELLA protein that regulates the gibberellin (GA) response by acting as a repressor of the expression of genes associated with stem elongation. When in the presence of the GID1:GA complex GAI1 is removed and ubiquitinated by the SCF:SLY1 complex and the 26S proteosome (Hedden, 2020). If the protein cannot be ubiquitinated due to a mutation in any of the genes in the GA regulatory pathway the genes it represses no longer get expressed and growth is reduced. Sanger sequencing of the GAI1 gene CQ040570 revealed that both tall and short F2 plants lacked the mutant allele (Figure 11) signifying that it did not contribute to the mutant phenotype.

Within the GA regulatory pathway there is another gene that harbors a mutation in EMS88 that is a likely candidate for causing the dwarf phenotype. The SLEEPY1 (SLY1) (McGinnis et al., 2003) gene CQ022076 in chromosome 4 of the B genome has an identified missense mutation in EMS88. SLY1 acts as part of the SCF-SLY1 E3 ubiquitin ligase complex to terminate DELLA repression by destroying DELLA proteins via the ubiquitin-proteasome pathway (Sun, 2011). Loss-of-function *sly-1* mutants can exhibit a *gai-1* like dwarf habit. Sanger sequencing verified the absence of the expected SL1 mutant allele in short EMS88 plants while it was potentially heterozygous in a tall plant (Figure 12).

Since the mutant phenotype loosely resembles mutant CLAVATA1 (CLV1) in *Arabidopsis thaliana,* a gene that maintains the apical meristem, variants associated with genes involved with meristem maintenance were investigated within family EMS88. The library did not contain mutations in any of these genes for family EMS88 but did contain a mutation in the gene Cytokinin Oxidase/Dehydrogenase-3 (CKX3)(Werner et al., 2003) CQ012056 in chromosome 8 of the B genome. CKX3 has been shown to localize to the apical meristem and regulate WUSCHEL (WUS) expression, with excess regional cytokinin and a larger shoot apical meristem associated with *ckx3* loss-of-function mutants (Werner et al., 2003; Bartrina et al.,2011). Short plants were shown to be heterozygous for the mutant allele after Sanger sequencing of the CKX3 gene CQ012056 (Figure 13). Because of the genetic background of EMS88, with 92 high-impact variants and 1185 moderate-impact variants, it was not clear which other variant may be contributing to the dwarf phenotype but there are more candidates that will be discussed in the next section.

Discussion

Generating the Library

Time to Completion

Progression to the M2 generation was the most time-consuming step in the genome wide-survey pipeline. Initial production of M2 plants was not efficient since more space than necessary for both tissue and seed production were given to each plant to observe and isolate interesting phenotypes. Tissue harvesting can occur as soon as the primary leaves are mature, and seed can be harvested more efficiently from plants that are generally smaller from overcrowding (something that was avoided to not contribute to environmental influences on phenotype). DNA extraction is also labor intensive but not necessarily time consuming, with an average of 4-hours to process 30 samples, but processing time could be reduced by increasing the number of workers processing samples simultaneously.

The next bottleneck to analyzing the EMS families is the actual variant detection pipeline. Once the variant-detection pipeline was standardized a sequencing batch of 96 families could be processed in one week. The total time estimated to generate the mutant library therefore depends on the number of families to include and the availability of labor – with the bioinformatics processes only restricted by limitations in computing power and software capabilities. Assuming space and labor are available in sufficient quantities to reduce the time of tissue harvesting and DNA extraction then a library could be produced at a rate equal to the number of families that can be processed by the variant-detection pipeline.

Explaining Variations in SNP Numbers

The number of called variants reached an average of 1242 genic variants per family but the distribution of variant number differed widely (Figure 5). This can be attributed to several factors including poor sequencing coverage, SNP loss, variant type definitions and EMS bias. It has been established that increasing sequencing depth increases the number of variants called, this is partially due to the cutoffs set for maps1 which discards any SNP called without reaching a minimum read depth (Lieberman & Henry, 2012). The library preparation involved diluting DNA samples to an average concentration of 1.7 ng/ul and pooling them together with individual indices. It is possible that variability in DNA concentrations and random loss throughout the process contributed to differences in sequencing depth between samples.

Since average outcrossing rates between closely grown individuals are estimated to be between 0.5 to 17.4% there is a chance every generation for the mutant allele to be lost through outcrossing, which reduces the probability of producing a homozygous recessive individual to about 16-20% from heterozygous parents (Gandarillas, 1979; Silvestri & Gil, 2000). In addition, while seed selection is random (and 6 individuals are planted from each family) those finally included in the library may not be entirely representative. This is especially true in cases where not all 6 individuals could be included in the sequencing samples due to premature die-off or poor tissue quality. However, the greatest variable may come from the beginning of the variant inducing process with EMS mutagenesis.

EMS acts as an alkylation agent that puts an alkyl group at the N7 and O6 guanine causing G/C to A/T transitions (Greene et al., 2003). It has been reported that certain regions of the genome have more allelic variations than average implying that mutagenesis with EMS is not entirely random. In a report by Yan et al. several “hotspots” of regions carrying more than 10 SNPs were identified in EMS treated rice alongside several “coldspots” carrying no SNPs (Yan et al., 2021). Regions with high densities of EMS-type variants were clustered around centromeres and were more often in regions of heterochromatin or regions with high GC content. This has been attributed to the theory that EMS mutations are more likely to be repaired in actively transcribed regions (Burns et al., 1986). Our data supports the observation of EMS bias with 29143 genes in the first 244 families analyzed having fewer than the average number of 6 variants and 23202 having more than the average (Figure 14). The average is greater than 1 partially because VEP will provide annotations for variants within 5 kb upstream or downstream of a gene leading to two entries for the same SNP position, but also from legitimate SNPs in the same gene found among the 244 families (McLaren et al., 2016). Our goal is to raise this average to 10 SNPs per gene.

Targeting Variants with Desirable Phenotypes

Trouble with GAI1

Although the library contains a review of most variants in many EMS families, the impact of any given mutation on plant height was limited for variants in GAI1 genes. From the first 52 families only one harbored a high-impact variant in a gene belonging to GA biosynthesis or regulation (Table 3). The missense mutation in EMS88 had potential to disturb GAI1 function, but to specifically reduce plant height an exact region needs to be affected. GAI1 contains a DELLA amino acid motif that acts as the binding site of the GA3-GID1 complex that initiates its ubiquitination. Other *gai1*-like mutants associated with dwarfing in wheat (*Triticum aestivum*), rice (*Oryza sativa*), and *Arabidopsis thaliana* contain mutations that interrupt the DELLA motif, usually through deletions (Peng et al, 1997; Ikeda et al., 2001; Willige et al., 2007). None of the original 52 EMS families had mutations in GAI1 in this region of the gene and CQ040570 and its homolog on chromosome B do not have identifiable DELLA motifs at all (Table 4).

While disappointing it was not surprising to discover then that the causative mutation for the dwarf phenotype observed was not in GAI1. There were other genes with missense mutations in EMS88 that could potentially cause the observed reduced height phenotype; SLY1, CKX3, FRIGIDA-LIKE 2 and 5 (FRL)(Michaels et al., 1999), and EARLY-FLOWERING 3 and 6 (ELF)(Zagotta et al., 1996). However, finding the causal allele would be challenging. The genetic background of EMS88, with over 1278 identified MODERATE to HIGH impact mutations, is “messy” and it is difficult to attribute a specific phenotype to the presence of a given mutant allele – especially with the presence of many “proteins of unknown function” in the reference genome annotation. The potential causal alleles and the putative genes they affect are described as follows in order of likelihood to cause the observed dwarf phenotype.

SLY1 Variant as a Potential Causative Allele

The plant response to gibberellin is reliant on GID1 to interact with GAI1 and facilitate its ubiquitination by the SCF-SLY complex and the 26S proteosome (Sun, 2011; Nelson & Steber, 2017; Hedden, 2020). Mutations in SLY1 disrupt the ability of the ubiquitinase-proteasome to degrade DELLA proteins resulting in their accumulation (McGinnis et al., 2003). The resulting plants are phenotypically similar to reduced height *gai1* mutants, with mutations in their DELLA motifs, and are partially rescued by knockout mutations of the same DELLA proteins (Fu et al, 2004). It was found through yeast two-hybrid and in vitro pulldown assays that SLY1 interacts with the C terminal GRAS domain of DELLA proteins GAI1 and REPRESSOR OF ga1-3 (RGA) independently of the N-terminal DELLA motif (Dill et al., 2004). Since the function of SLY1 is not impaired in EMS88 dwarf mutants it is not likely the causal allele.

CKX3 as a Potential Causative Allele

Since cytokinin has been demonstrated to induce WUS expression in the shoot-apical meristem (SAM) then the loss-of-function of a gene that reduces levels of cytokinin should increase WUS expression (Werner et al., 2003; Gordon et al., 2009). The gene WUS works in tandem with CLV to moderate the size of the meristem (Clark et al., 1997, Fletcher et al., 1999). If WUS is expressed outside of its normal domain, such as in *clv3* mutants, the meristem becomes enlarged usually resulting in an abnormal number of club-shaped flower organs (Clark et al, 1993). While this may affect growth by increasing shoot-apical meristem size and influencing the initiation of nodes it does not directly explain the dwarf growth habit. The overexpression of CKX3 was shown to induce a dwarf growth pattern in tobacco (*Nicotiana tabacum*) indicating that a loss-of-function mutation (like the missense mutation in CQ012056) should produce the reverse and increase plant height (Werner et al, 2001). In *Arabidopsis* it has been found that a pair of loss-of-function mutations in the genes CKX3 and CKX5 both induced a mild dwarf phenotype and increased seed size/yield (Bartrina et al., 2011). It will require more research to prove if the missense mutation in CKX3 is contributing to the dwarf phenotype of EMS88 including several generations of backcrosses to isolate the mutant allele.

Other Potential Causative Alleles

While genes that directly affect growth i.e., stem-elongation, are the most likely candidates to harbor mutations leading to a reduced height phenotype in EMS88 there are other options that could terminate growth in more round-about ways. The FRL genes are responsible for regulating flowering time by repressing transition to flowering until the appropriate environmental queues occur, such as vernalization or transition to short/long days (Gazzani, 2003). The gene family ELF has been shown to affect circadian regulation of flowering in Arabidopsis (Hicks et al., 2001). Mutations in these genes may be stimulating dwarf plants to arrest vertical growth and transition to flowering early in the season, replacing the SAM with the floral meristem (FM). In EMS88 there is an indication that short plants transitioned to flowering earlier than tall plants. F2 plants after one-month post germination had small panicles forming while plants that would grow tall at maturity did not (Figure 15). However, tall plants had begun forming panicles within a week later and would eventually continue growing to an additional 20 to 60 cm in height. While short plants may turn out to harbor mutations in these genes, given no *frl* or *elf* mutant to date has been characterized to cause reduced height, they are unlikely candidates for being sufficient to cause the observed dwarf phenotype although they could possibly contribute to it. A study like that for CKX3 would need to be conducted to isolate the mutant allele and define a phenotype for those variants in quinoa.

Genotypic Screens for Better Precision

Much of this guesswork could be reduced with an approach to screening the candidate EMS families for targeted variants that more resembles TILLING. Instead of focusing efforts on finding the desired phenotype a targeted mutant allele can be selected for through genotyping, traditionally done with the CEL I endonuclease in TILLING, and then characterized. The most common genotyping method in plant breeding is based on marker assisted selection (MAS) (Knapp, 1998; He et al., 2014). MAS simply tracks molecular markers, usually SNPs or simple sequence repeats (SSR), to make selections for multiple generations of crosses and backcrosses. Currently qPCR using fluorescent hybridization probes to identify the presence of specific SNPs is the most common system of tracking SNPs for MAS (Ragoussis, 2006). This system can identify the presence of thousands of SNPs per day and can be performed on a benchtop real-time PCR machine, which pairs well with the high-throughput mutant detection pipeline we are using. Another genotyping option exists with next-generation sequencing known as genotyping by sequencing (GBS) – which is like the qPCR method but uses NGS to track SNPs instead of fluorescent probes (Ragoussis, 2009). Using either method in a genotypic screen may reduce the number of individuals required per generation to find and isolate the mutant allele (identifying it in vitro instead of waiting for an expected phenotype) and could offer multiple genetic backgrounds for a more thorough analysis of gene function. The benefit of a phenotypic screen over a genotypic screen is its simplicity and independence from costly laboratory equipment. This means that the library remains a valuable resource to simpler or more primitive breeding programs.

Conclusion

Quinoa is a nutritious food crop that has the potential to improve food security but is less hardy outside of its native regions in the Andes. Improvement programs are hampered by a lack of available germplasm and allelic diversity. Allelic diversity can be induced through mutagenesis and a TILLING by WGS approach that facilitates the discovery of mutant alleles throughout the entire genome that contribute to agronomically desirable characteristics. An EMS library was developed with an associated library of variants with their predicted effect. Using this variant library to target EMS lines that may have mutations in genes affecting plant height; a dwarf phenotype was observed in several EMS families and a more extreme dwarf phenotype was observed in EMS88. While this line was targeted for a phenotypic screen because of an identified missense mutation in the gene GAI1, Sanger sequencing revealed that this mutant allele was absent from the dwarf plants. Dwarf plants were also shown to be heterozygous for a missense mutation in the gene CKX3 which has not been characterized to have a direct effect on stem length.

Using a phenotypic screen is a simple way to parse a population for individuals with a desirable phenotype but does not guarantee that the phenotype is caused by the targeted mutant alleles. A genotypic screen, while more technically complex and potentially expensive, would maintain focus on the targeted allele whether it caused the predicted phenotype or not. In essence the library can be used in two ways: in forward genetics to reduce the number of EMS families to include in a phenotypic screen, or in reverse genetics to target individual alleles harbored by EMS families in a genotypic screen. Both approaches are capable of introducing the needed genetic changes in quinoa to adapt it to the needs of growers internationally, securing it as a resource into the future.

Figures and Tables

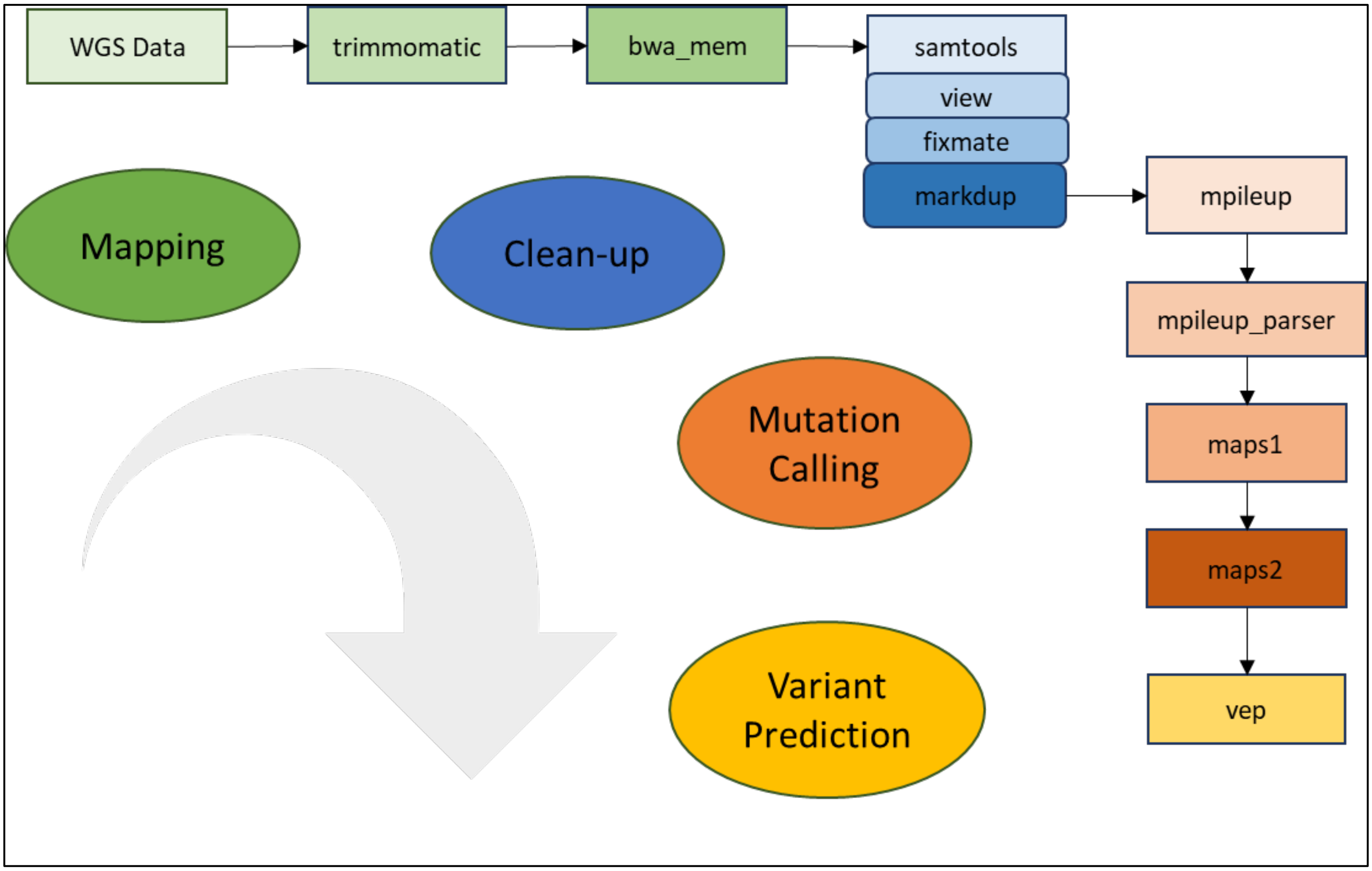

Figure 1 Variant detection pipeline. NG sequencing reads are trimmed with trimmomatic before being mapped to the quinoa v2. reference genome with bwa_mem. The resulting SAM file is cleaned using the samtools suite. First the SAM files are converted to BAM files with view, then mate coordinates were filled with fixmate, and finally duplicate reads were removed with markdup. The cleaned BAM files were sent to the UC Davis Comai Lab's Mutant and Polymorphsim Survey (MAPS) pipeline by first generating an mpileup file of all the BAM files. The mpileup file was then parsed and sent to maps1 and then maps2. The final list of polymorphism sites was sent to the Ensembl Variant Effect Predictor (VEP) program to define the predicted effect of each polymorphism based on a reference GFF file.

Table 1 The format of the library of variants. The library contains information describing the location and nature of the variant in the genome. Gene position is based on the chromosome scale scaffolds in the quinoa reference genome. The nucleotide and amino acid changes are based on the quinoa v2. reference GFF and may not necessarily be in the correct reading frame, and the gene similarities are based on homology alone. This affects the predicted variant type and impact, based on the VEP definitions (https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html). *Actual library places this information into a single tab-delimited line per entry.

Position	**Contig92_pilon:2830721**	**Contig92_pilon:515776**	**Contig92_pilon:1861861**	**Contig1620_pilon:12288**	**Contig1134_pilon:24787**
Gene	CQ000240	CQ000031	CQ00015	CQ057466	-
Nucleotide Change	tgG/tgA	cCt/cTt	aaG/aaA	-	-
Amino Acid Change	W/*	P/L	K	-	-
Variant Type	Stop Gained	Missense Variant	Synonymous Variant	Downstream Gene variant	Intergenic Variant
Strand Sense	-	-	-	-	-
Impact	HIGH	MODERATE	LOW	MODIFIER	MODIFIER
Notes	Similar to LAX2: Protein LAX PANICLE 2 (Oryza subsp. japonica OX%3D39947)	Similar to BAM1: Leucine-rich repeat receptor-like serine/threonine-protein kinase (Arabidopsis thaliana OX%3D3702)	Similar to ORC6: Origin of replication complex subunit (Arabidopsis thaliana OX%3D3702)	Similar to TMN3: Transmembrane 9 superfamily member 3 thaliana OX%3D3702)	

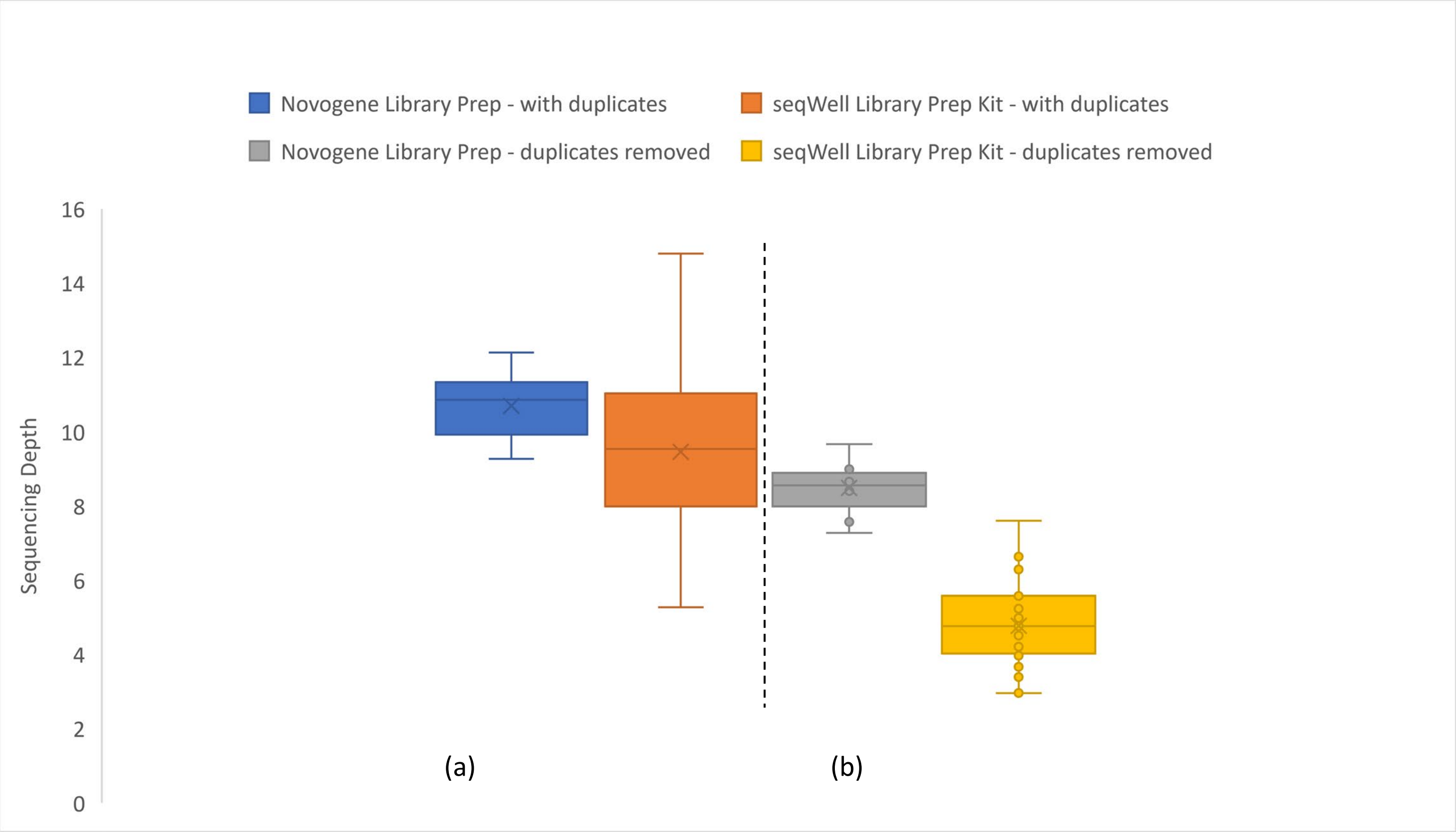

Figure 2 Different sequencing coverage depths between EMS families using WGS and either in-house library prep by Novogene or from using the seqWell plexWell LP-384 library prep kit. Initial read depths were similar (a) but after duplicates were removed, (b) data obtained from using the seqWell library prep kit had roughly ½ the depth of the other data.

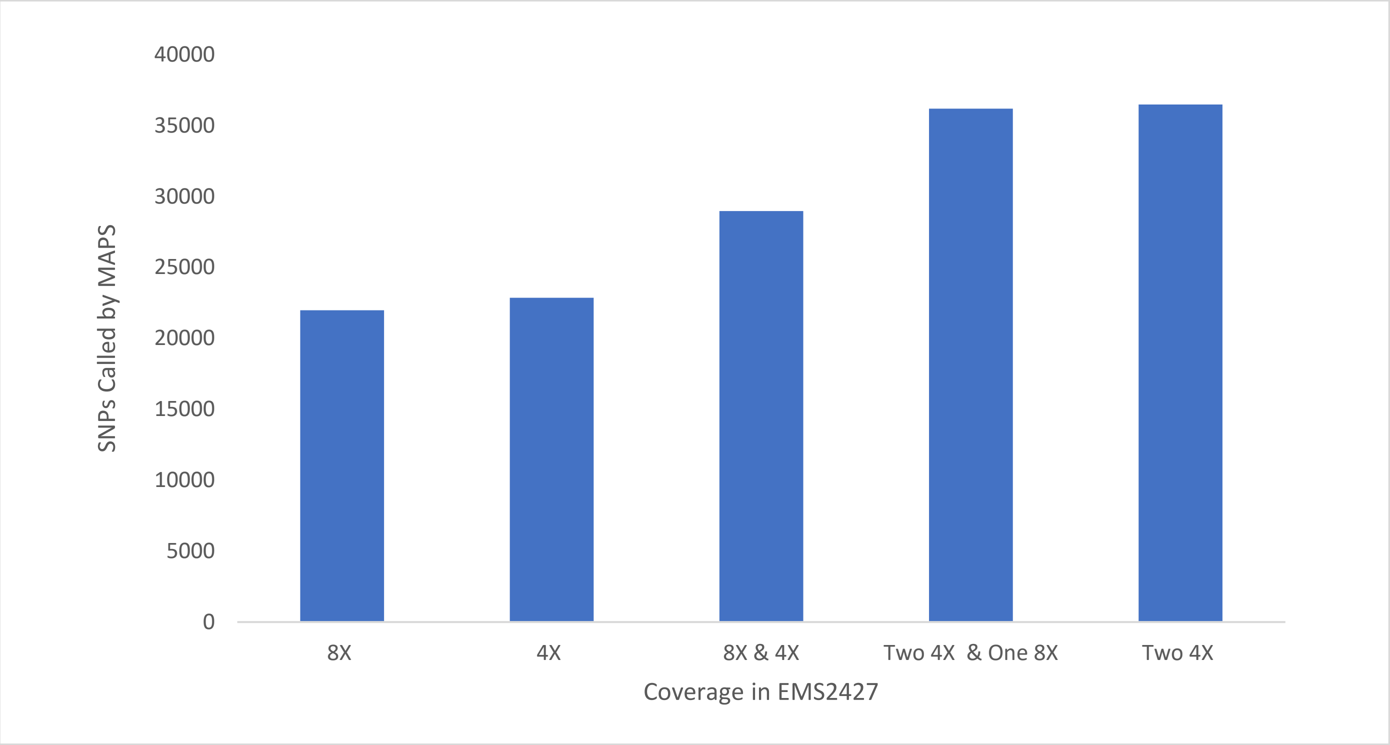

Figure 3 Changes in SNP number called in family EMS2427 based on different library preparation strategies leading to full coverage 8X (Novogene library prep) or half coverage 4X (seqWell library prep kit). 8X and 4X coverage alone are not very different but merging multiple sequencing results increase SNP calls by ~50%. The greatest increase was from merging two 4X coverage samples (two seqWell library prep kits). There was no loss in SNPs called by using a single seqWell library prep kit per batch of 96 families.

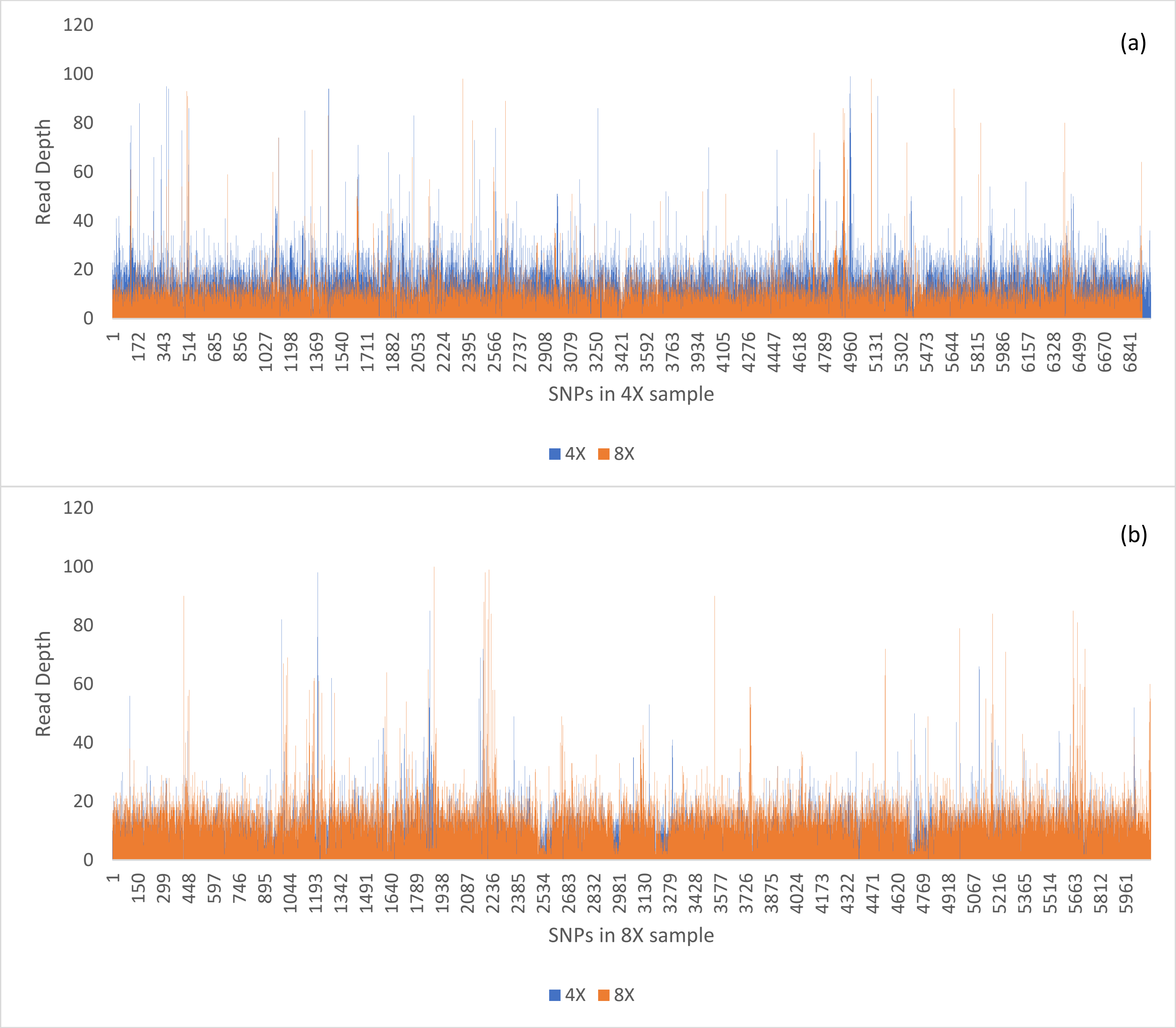

Figure 4 Differences in read depth for each SNP called in the (a) 4X coverage sample and the (b) 8X coverage sample. Most SNP sites called in the 4X coverage sample had a higher read depth in the 4X coverage sample than the 8X coverage sample at those sites. Most SNP sites called in the 8X coverage sample had a higher read depth in the 8X than the 4X at those sites.

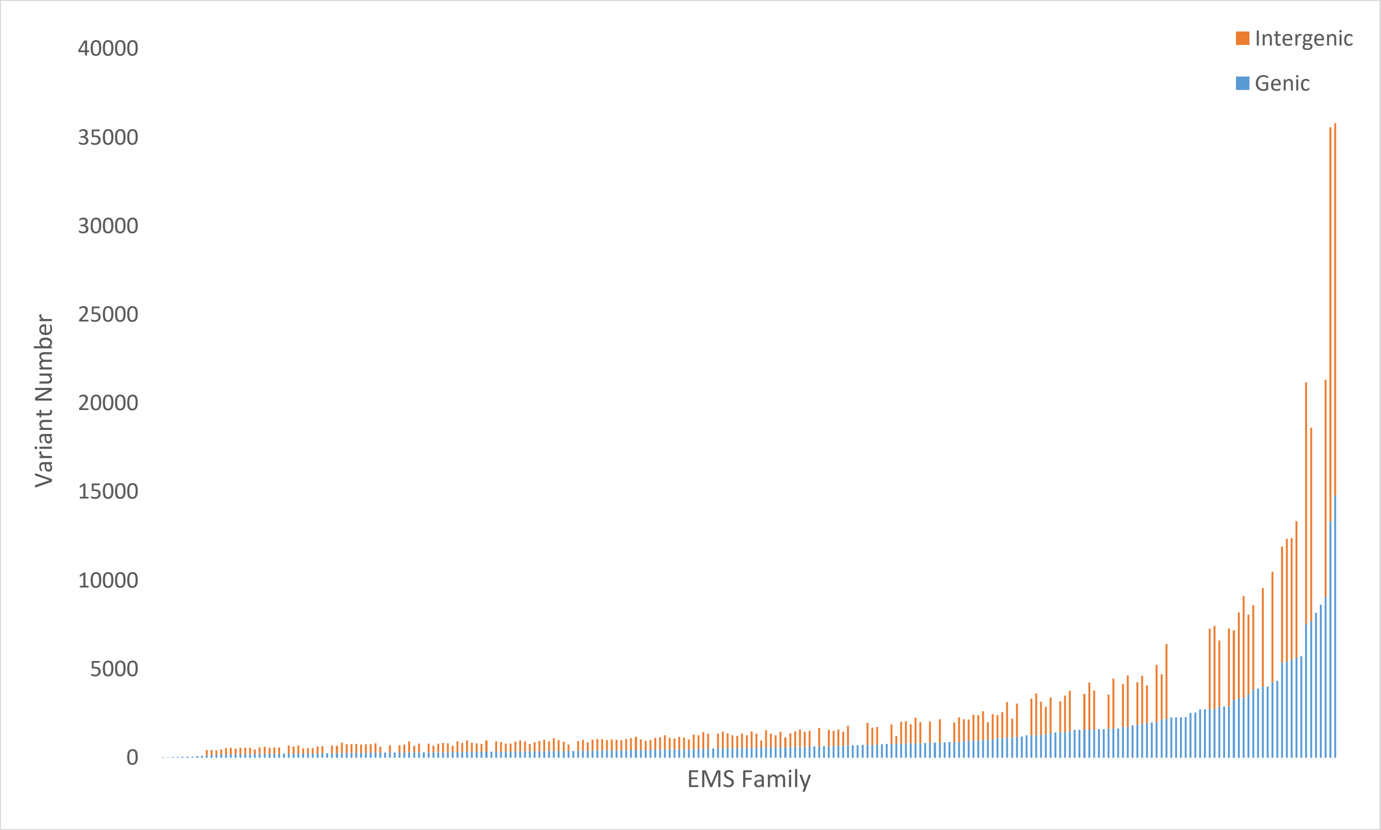

Figure 5 Distribution of variant number called by the variant-detection pipeline per family. Intergenic variants are shown in orange and genic variants are in blue. The largest number of genic variants called was 14801 while the lowest was 6. Variants in intergenic regions made up about half of all variants called in WGS samples; WES samples have no intergenic data so show no orange lines. While this demonstrates that the mutant rate is inconsistent and it has been shown that coverage at any given position varies between sequencing runs, this distribution is satisfactory to provide insight into which allelic variants are present in each EMS family.

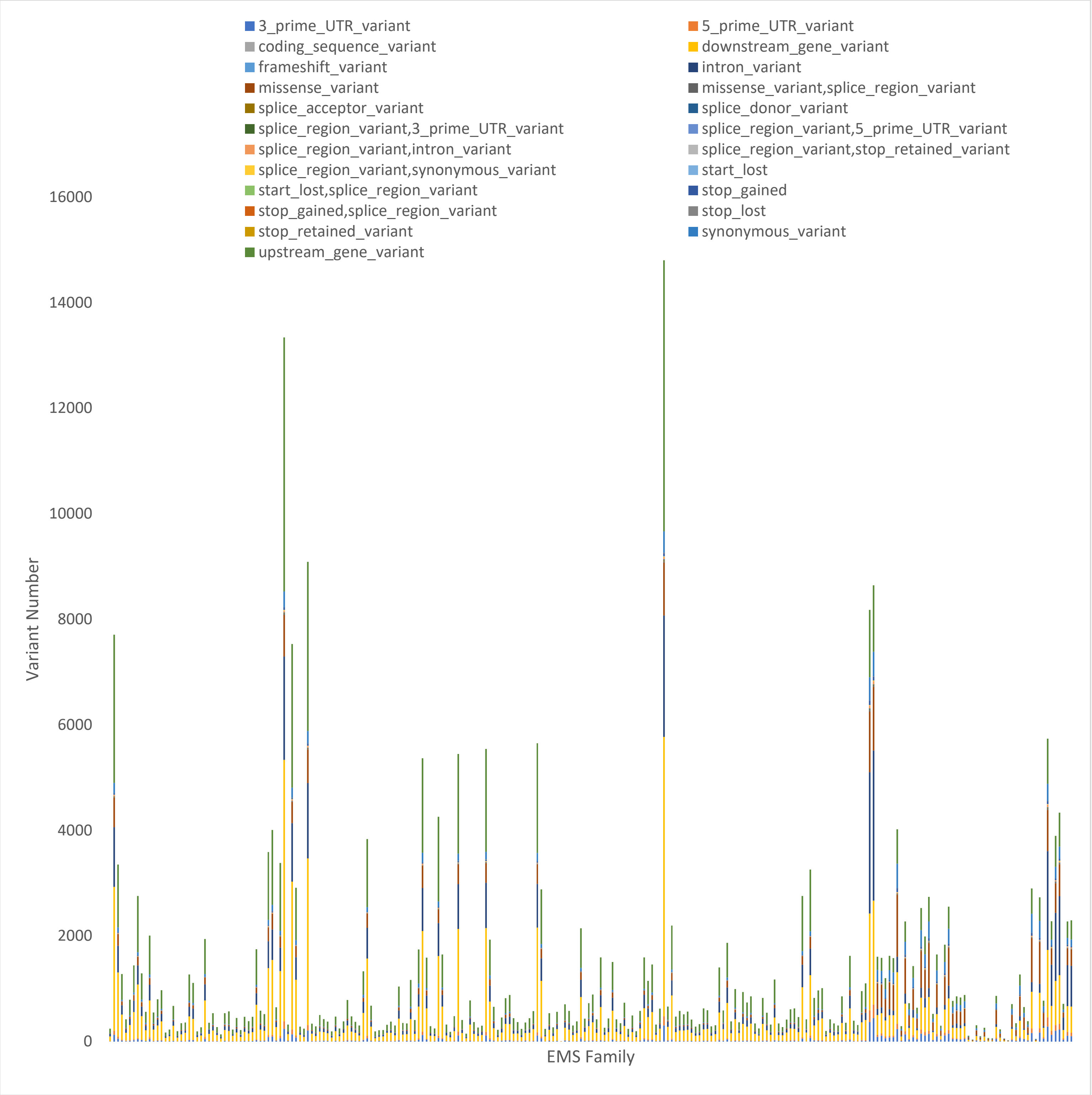

Figure 6 Distribution of variant type by EMS family. Variant types are defined by VEP according to The Sequence Ontology (SO). The most abundant variant types are also in regions less likely to have a major impact on gene function; upstream/downstream of the gene and introns. This is finding is further visualized in Figure 7.

Table 2 The different variant types found within the first 244 EMS families in the variant library. The most abundant variants apart from intergenic are downstream and upstream variants which could be considered intergenic but are within 5 kb of upstream or downstream of a genomic feature, i.e., transcript/gene. The impact designations are considered subjective by Ensembl but are based on potential impact on function.

	IMPACT				
TYPE	TOTAL	MODIFIER	LOW	MODERATE	HIGH
intergenic_variant	315559	*			
downstream_gene_variant	98535	*			
upstream_gene_variant	94010	*			
intron_variant	48121	*			
missense_variant	31105			*	
synonymous_variant	12781		*		
3_prime_UTR_variant	8657	*			
5_prime_UTR_variant	5175	*			
stop_gained	1946				*
splice_region_variant	1355		*		
missense_variant,splice_region_variant	502			*	
splice_acceptor_variant	319				*
splice_donor_variant	288				*
coding_sequence_variant	80	*			
start_lost	52				*
stop_gained,splice_region_variant	45				*
stop_retained_variant	42		*		
frameshift_variant	21				*
stop_lost	4				*
start_retained_variant	3		*		
start_lost,splice_region_variant	1				*

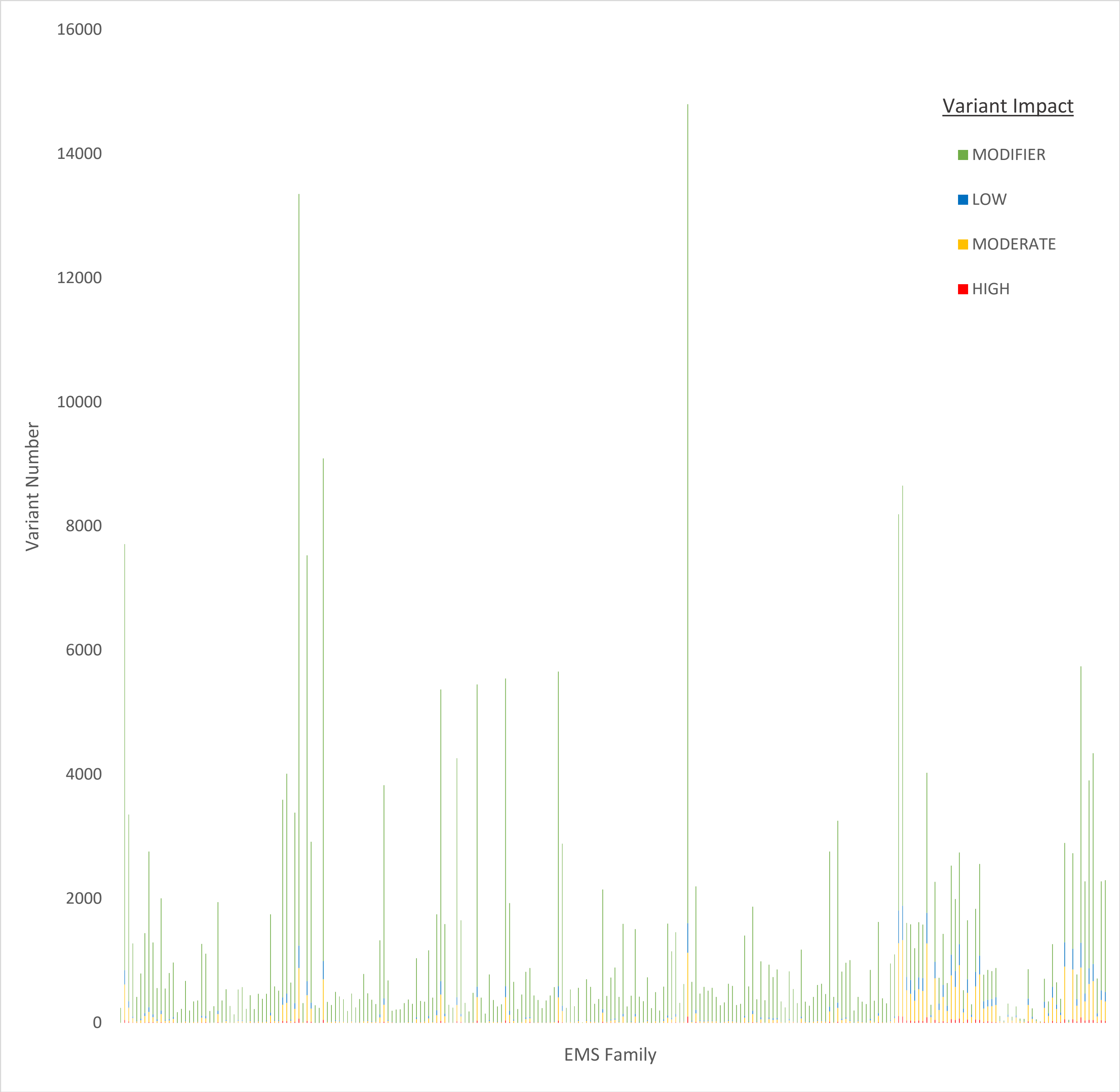

Figure 7 Distribution of variant impact per EMS family in the first 244 processed. High impact variants include stop-gained, stop-lost, splice-region-variant, and start-lost. Moderate impact variants include missense. Low impact variants include 3’ and 5’ UTR, intron, synonymous, stop and start retained, upstream, and downstream. Modifiers are mutations with an unpredicted impact and are usually intergenic. While the true nature of each variant is unknown there is a wide distribution of both overall variant number and impact, with some families having almost exclusively modifiers or mutations in intergenic regions.

Table 3 Genes associated with gibberellin regulation and biosynthesis that may impact plant height that had identified mutations in EMS families. Target genes were found in eight EMS families from the original 52 families sequenced with a combination of WES and WGS.

Gene	Gene ID	Chromosome	Family	Variant
GAI1	CQ026984	6A	3067	Missense
	CQ040570	8A	88	Missense & Upstream
	CQ001679	6B	549	Missense
	CQ001679	6B	541	Stop Gained
GID1-C	CQ020735	4B	1679	Missense & Downstream
	CQ036017	4A		
GA3OX (LE)	CQ052980	2A	500	Missense & Downstream
	CQ052979	2A	2427	Missense & Downstream
GA20OX-2	CQ006906	5B	170	Missense & Upstream

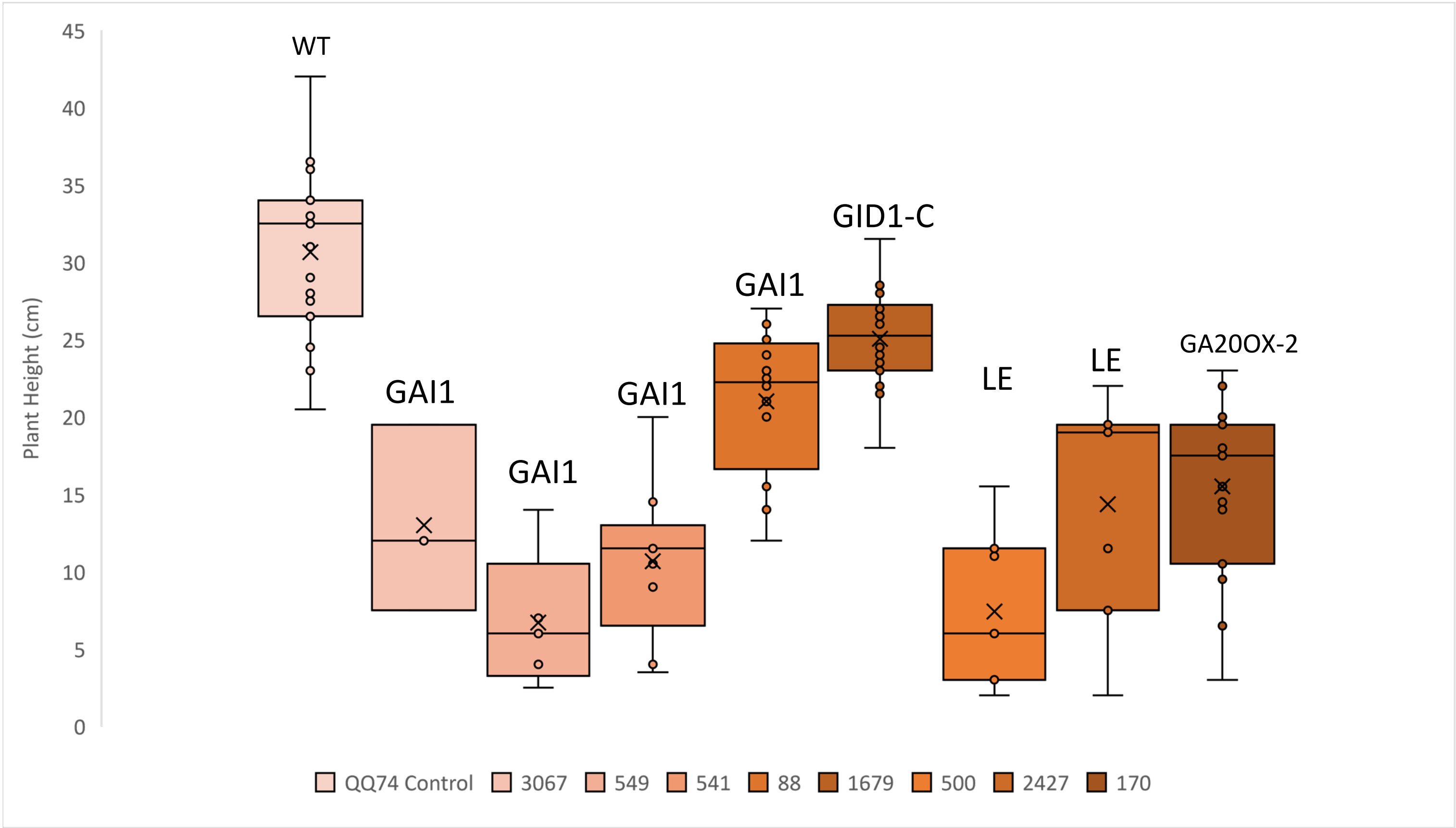

Figure 8 M3 plant heights after one-month post-germination (PG). Every family contains individuals with statistically significant height differences compared to the wild-type QQ74 controls except EMS1679. Family EMS88 contained dwarf plants with an extreme phenotype so they were isolated for a more focused analysis.

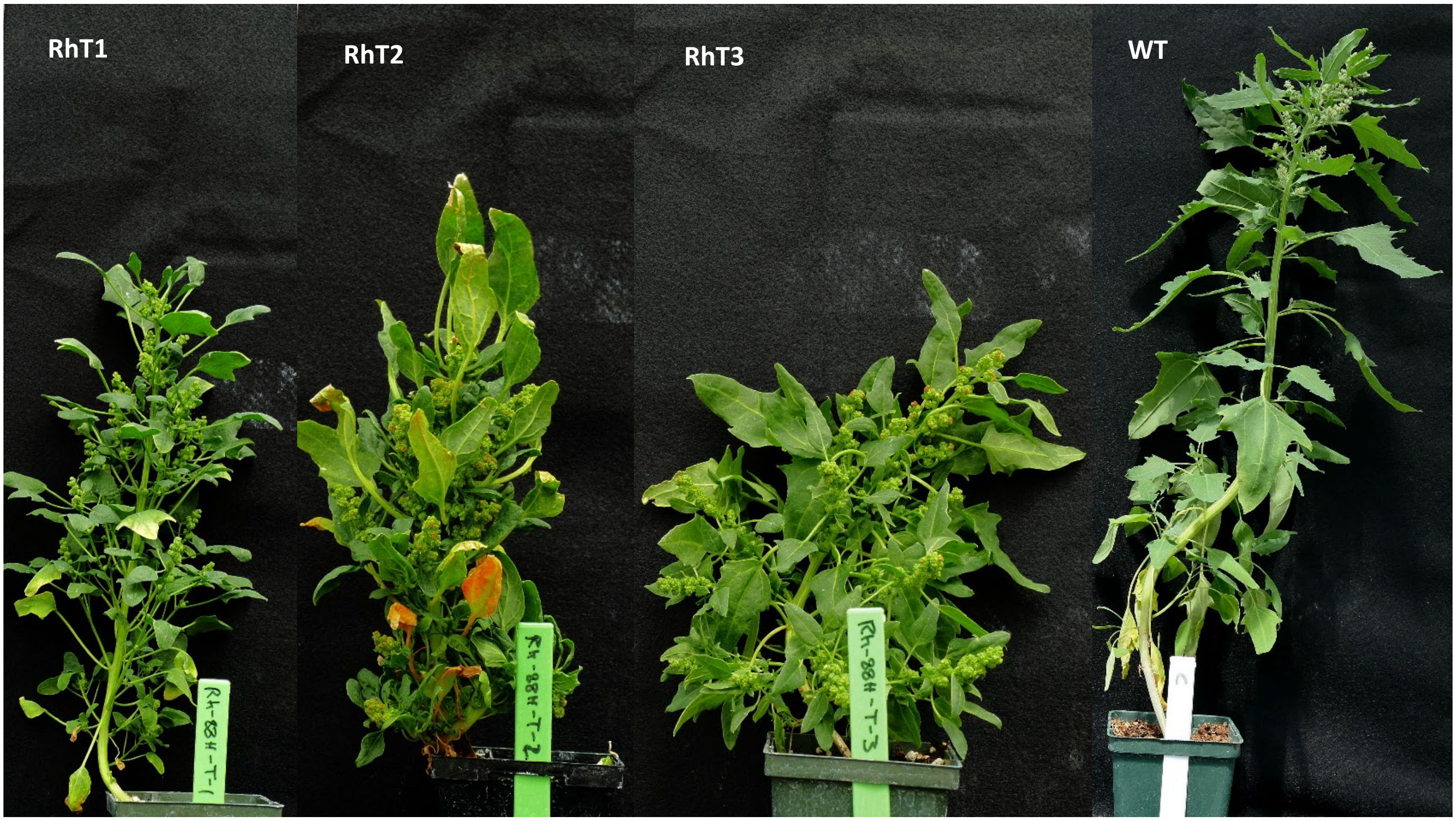

Figure 9 M2 plants from family EMS88 that demonstrated an extreme reduced-height phenotype. Individual RhT2 had short, compact inflorescences while individual RhT3 had a more exaggerated branching habit and individual RhT1 was most similar to a wild-type growth pattern. These plants were allowed to grow to maturity and their seeds were harvested to produce an F1 generation.

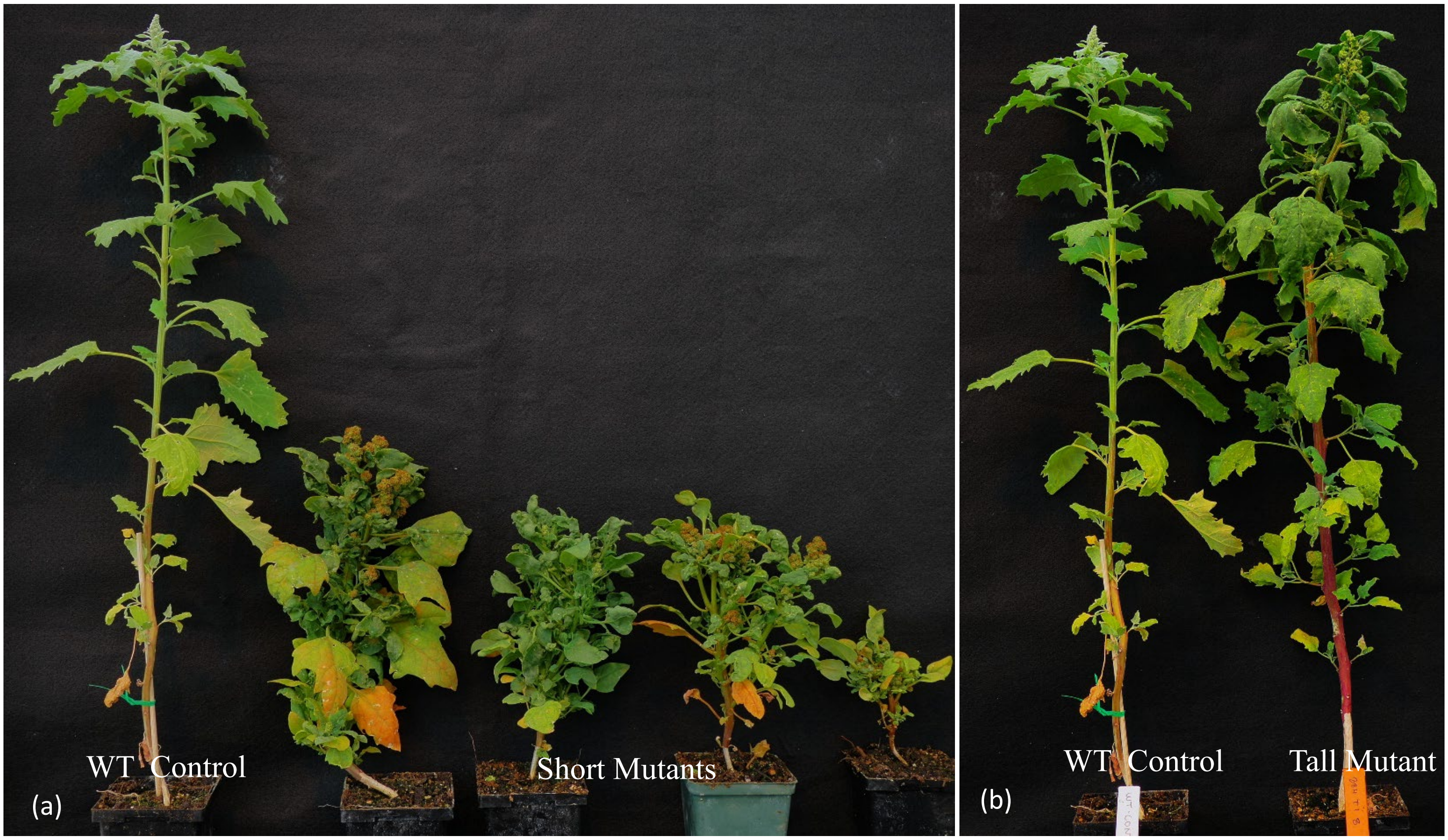

Figure 11 Individuals from the F2 generation of dwarf EMS88 plants. F2 individuals from RhT1 were tall and individuals from RhT2 were short. (a) Short individuals were also branched and compact. (b) Tall individuals lacked the compact-branched phenotype and reached the same height as the wild type.

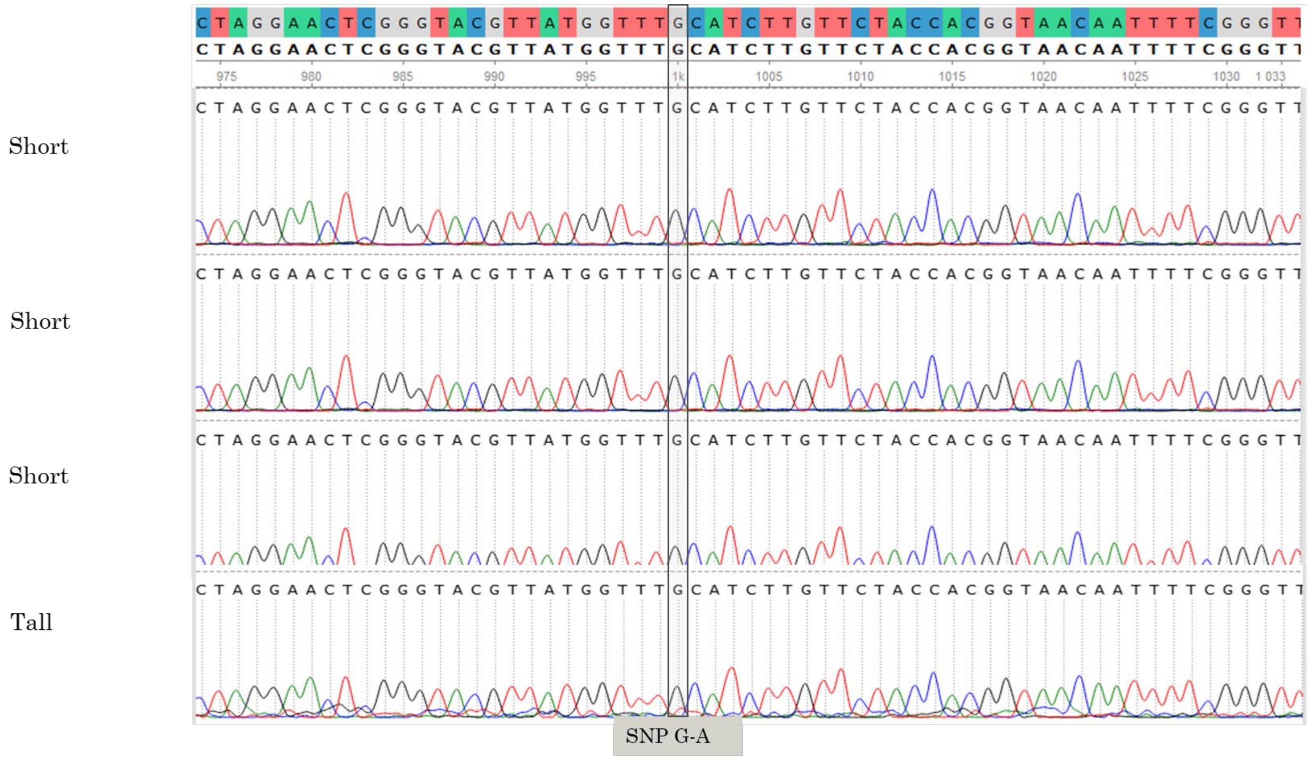

Figure 10 Sanger sequencing results for the GAI1 gene CQ040570 in the short and tall F2 individuals. The highlighted column shows the locus of the identified variant and reveals that neither tall nor short plants harbor the variant.

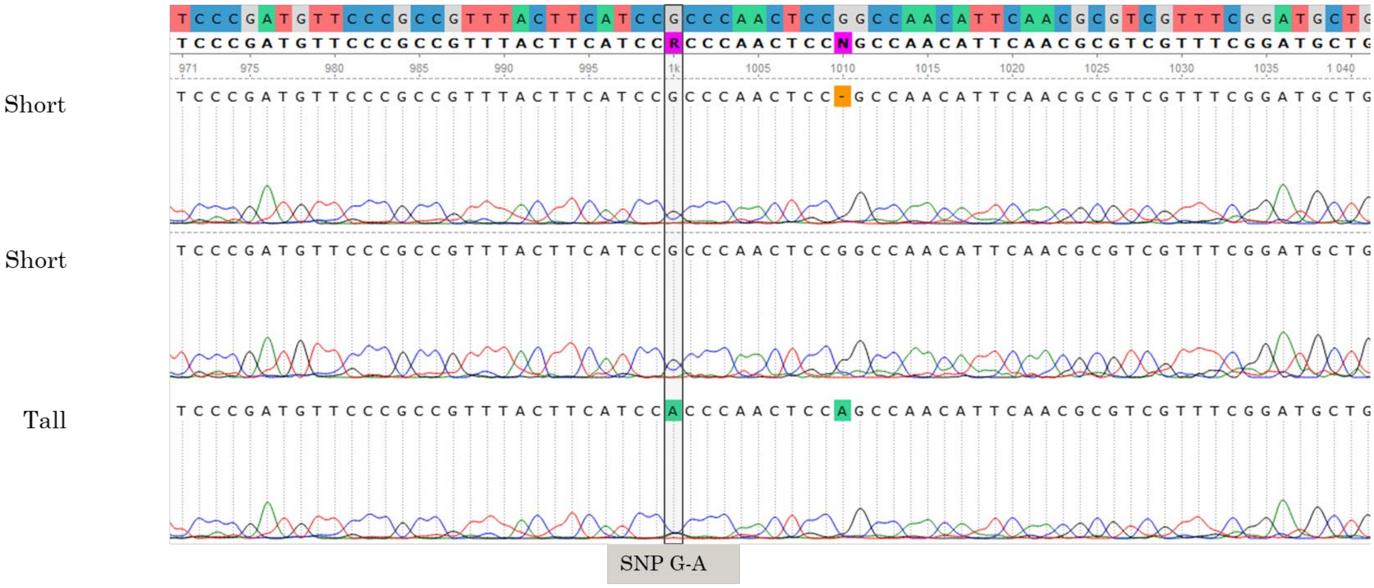

Figure 13 Sanger sequencing results for the SLY1 gene CQ022076 in short and tall F2 individuals. The highlighted column shows the locus of the identified variant and reveals that short plants do not harbor the variant and the tall plant may be heterozygous for the variant.

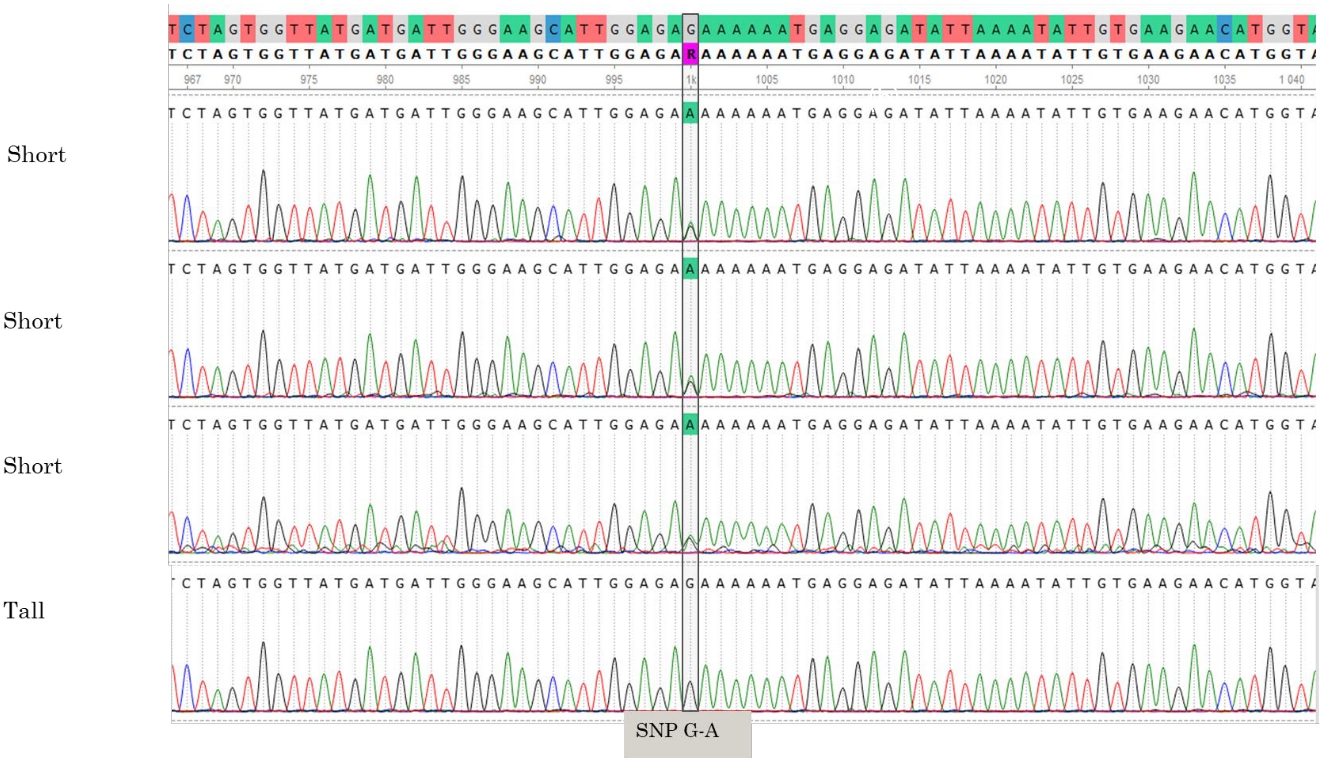

Figure 12 Sanger sequencing results for the CKX3 gene CQ012056 in the short and tall F2 individuals. The highlighted column shows the locus of the identified variant and reveals that short plants are heterozygous for the mutant allele while the tall plants are homozygous for the wild-type allele.

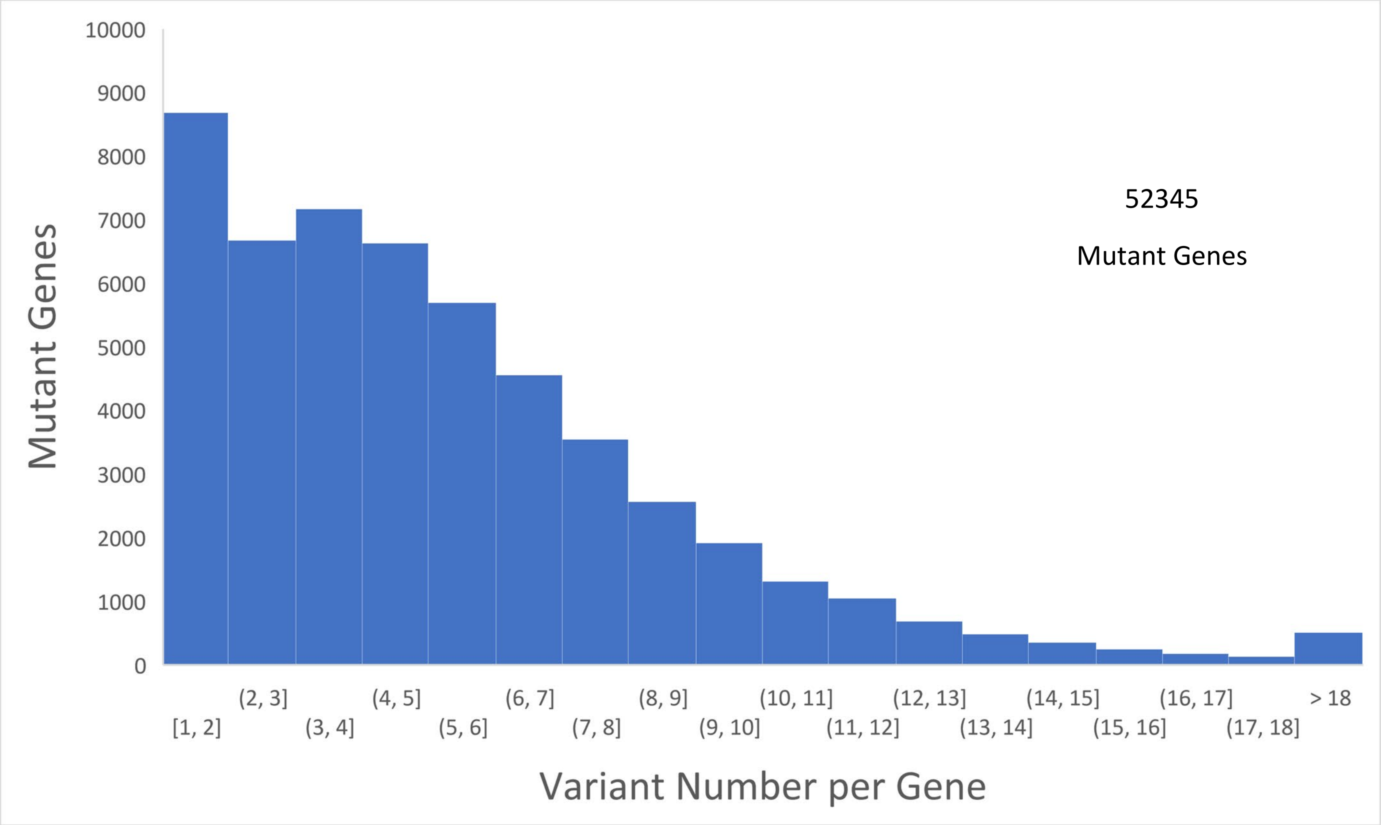

Figure 14 The representation of identified SNP per gene based on an average of 6 SNP per gene in 244 samples. The average of 6 SNPs comes from VEP identifying variants on both positive and negative strands at a position and from legitimate SNPs in distinct positions in genes. From a total of 52345 genes, genes with the highest number of SNP were identified as regions with long-repeats. Some may also have large deletions that MAPS calls as multiple SNPs.

Table 4 The amino acid sequences for CQ040570 on chromosome 8A and CQ012679 on chromosome 8B. Neither putative DELLA protein contains a DELLA motif. The homologous gene CQ026984 on chromosome 6A does have a DELLA motif. Each sequence is in the same reading frame as their NCBI protein Reference Sequence XP_021741625.1, XP_021764513.1, and XP_021727292.1 respectively.

CQ040570	No DELLA motif	Chromosome 8A
	MMSSYELSCVVSHPVSDMDELIGYDRVTSWVNTFVPEHILGPGS GLGLSYVDPVPNELMKVFGDTWTDQIDGQQLSGESQLVCDMEEDVGIRLVHALLTCAD ALQRGEFQLANSLIGEMSNGMMTRVNTTCGIGKVAGYFIDALSRRLCQPSLGVSVAGP GSVSNEVLYHHFYEACPYLKFAHFTANQAILEAFEGQAHVHIIDFNLMHGLQWPALIQ ALALRPGGPPSLRLTAISPSTSNVHDFFHETRMRLAQLARSINVRFSFRVVTTLRLED IKPWMFQTSPEEVIAVNSILQFHRLLNTNINQVLESIKSLNPKIVTVVEQDANHNVPE FLVRFTEALHYYSAMFDSLEACQLEAAKPLAETFMQREICNIICCEGSARIERHEPLT KWQARLIQAGFKALHMGRNAFKQANMLLSLCSGEGFSVQECDGCLTLGWHNRPLIAAS AWQVRTDKDGISLTNDGSTSSSSSS	
CQ012679	No DELLA motif	Chromosome 8B
	MRSYDLYGVSHPVSDMDELIGYDRVTSWVNTFVPEHVSLPTTLG PGSGPGLSYIDPVPNELMKVIGDTWTDQIDEQQLSGESQLVCDMEEDVGIRLVHVLLT CADALQRGEFQLANSLIGEMTNGMMTRVNTACGIGKVAGYFIDALSRRLCHPGLTGPG SVSNEVLYHHFYEACPYLKFAHFTANQAILEAFEGQAHVHVIDFNLMHGLQWPALIQA LALRPGGPPSLRLTAISPSTSNGHDFFHETGMRLAQLARSVNVKFSFRVVTTSQLEDI KPWMFQTSPEEAIAVNFILQLHRLLITNINQFLESIKSLNPKIVTVVEQDANHNVPDF LVRFTETLHYYSAMFDSLEACQLEAAKPLAETYMQREICNIICCEGSARIERHEPLIK WQGRLIQAGFKALHMGRNAFKQANMLLSLCSGEGYSVQESDGCLTLGWHNRPLIAASA WQVRTDKDGISLTNDGSTSSSSSS	
CQ026984	DELLA motif	Chromosome 6A
	KRELPFTNPEPSTGGKPPVSGKSKMWEDQNDAGVDELLAVLGYKVRSSDMAEVAQKLEQL EQAMGNVREDGLSQLASETVHYNPSDLSTWLESMLSEFNPCTTFDDDSSSSQILLPSPIG SVVDPIVNPTPQSSIRFSSDPYSDYDLKAIPGKAILTPPSPSSSSASLVNNNDNNNNNSS SCNSLVSNTTSSSSSREAKRLKASNYTAASATATATVSNPSSSKLTANSAVSRPVVLVDS QENGVRLVHTLMACAEAIEQQNMGLAEALLKQIGFLAASQVGSMRKVATYFAEALARRVY KLCPDVPYDGNLSDMLQMHFYETCPYLKFAHFTANQAILEAFSGKKKVHVIDFSMKEGMQ WPALMQALALRREGPPAFRLTGIGPPAPDNSDRLQEVGWKLAQFADSIQIQFEYRGFVAN SLTDLESALLDLRPETEVVAVNSVFELHRLLARPGAAEKVMGLMKEVNPVIVTVVEQEAN HNGPAFLDRFNESLHYYSTLFDSLESCVDSQDKMMSEVYLGRQICNVVACEGVDRVERHE TLAQWRNRFGSAGFAPVHIGSNAFKQASVLLDYFAGGDGYGVEENDGCLMLGWHSRPLIT TSAWQLAKNNQNSTTRL*	

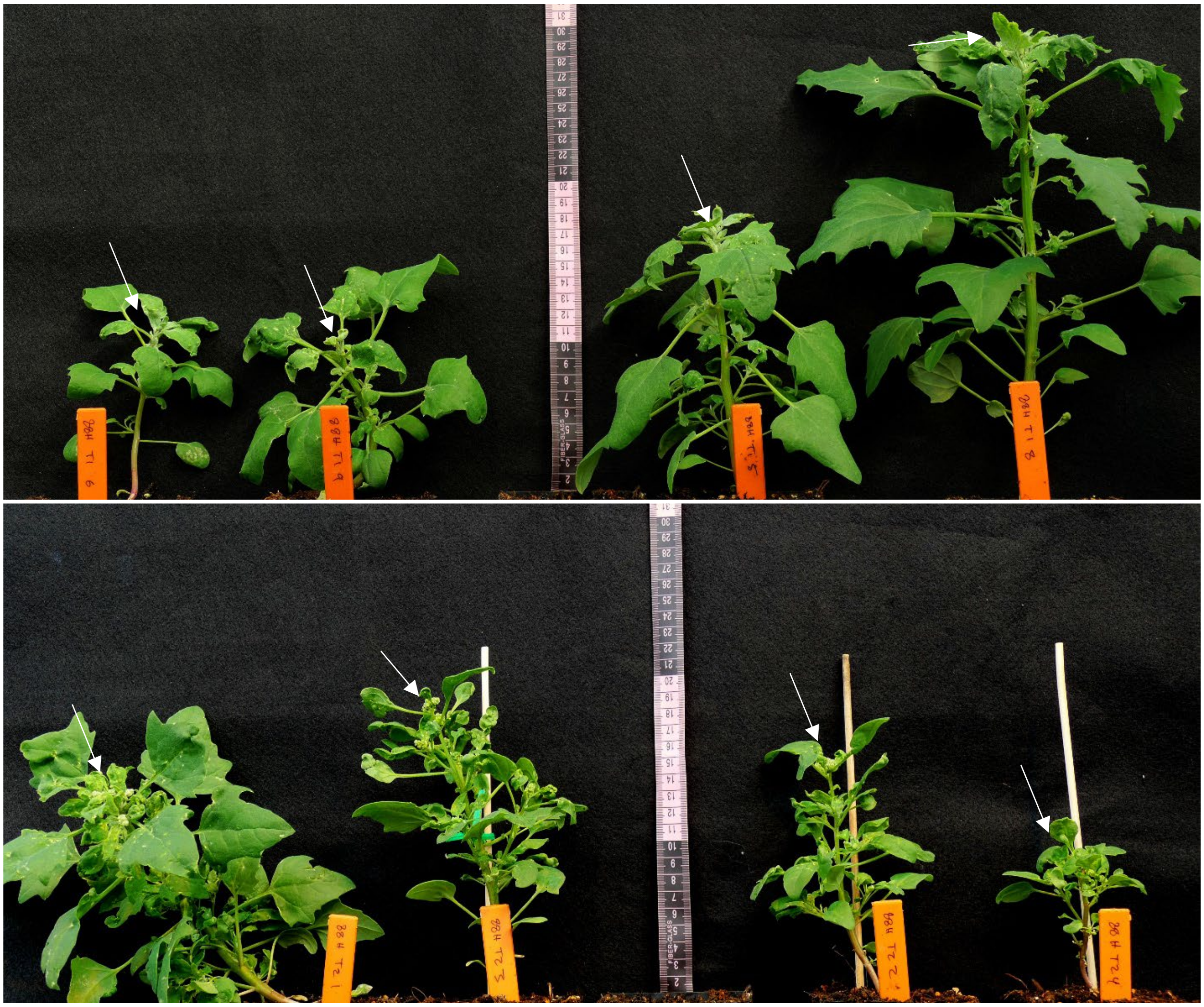

Figure 15 EMS88 F2 plants that would eventually grow tall (Top) and eventually grow short (Bottom) after one-month post-germination. Short plants have small panicles forming while tall plants do not. It is unlikely that this is the sole reason for the reduced-height as tall plants began to developed panicles shortly after and continued to grow up to 35 cm more before maturity.

Literature Cited

Bansal, V. (2017). A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments. BMC Bioinformatics, 18(Suppl 3), 43. doi:10.1186/s12859-017-1471-9

Barkley, N. A., & Wang, M. L. (2008). Application of TILLING and EcoTILLING as reverse genetic approaches to elucidate the function of genes in plants and animals. Current Genomics, 9(4), 212-226. doi:10.2174/138920208784533656

Bartrina, I., Otto, E., Strnad, M., Werner, T., & Schmulling, T. (2011). Cytokinin regulates the activity of reproductive meristems, flower organ size, ovule formation, and thus seed yield in arabidopsis thaliana. The Plant Cell, 23(1), 69-80. doi:10.1105/tpc.110.079079

Bazile, D., Bertero, H. D., & Nieto, C. (2015). State of the art report on quinoa around the world in 2013 FAO.

Bazile, D., Jacobsen, S., & Verniau, A. (2016). The global expansion of quinoa: Trends and limits. Frontiers in Plant Science, 7 doi:10.3389/fpls.2016.00622

Bazile, D., & Santivañez, T. (2015). Introduction to the state of the art report on quinoa around the world FAO.

Bhargava, A., Bhargava, A., Shukla, S., Shukla, S., Ohri, D., & Ohri, D. (2006). Karyotypic studies on some cultivated and wild species of chenopodium (chenopodiaceae). Genetic Resources and Crop Evolution, 53(7), 1309-1320. doi:10.1007/s10722-005-3879-8

Bhargava, A., Shukla, S., & Ohri, D. (2006). Chenopodium quinoa—An indian perspective. Industrial Crops and Products, 23(1), 73-87. doi:https://doi.org/10.1016/j.indcrop.2005.04.002

Bonifacio, A. (2015). Quinoa breeding and modern variety development FAO.

Bruno, M. C. (2014). Quinoa: Origins and development. Encyclopedia of global archaeology (pp. 6215-6220). New York, NY: Springer New York. doi:10.1007/978-1-4419-0465-2_2184

Buck, M., & Hamilton, C. (2011). The Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the convention on biological diversity. Review of European Community & International Environmental Law, 20(1), 47-61. doi:10.1111/j.1467-9388.2011.00703.x

Burns, P. A., Allen, F. L., & Glickman, B. W. (1986). DNA sequence analysis of mutagenicity and site specificity of ethyl methanesulfonate in uvr+ and UvrB- strains of Escherichia coli. Genetics; (United States), 4 Retrieved from https://www.osti.gov/biblio/5002225

Choukr-Allah, R., Rao, N. K., Hirich, A., Shahid, M., Alshankiti, A., Toderich, K., et al. (2016). Quinoa for marginal environments: Toward future food and nutritional security in MENA and central Asia regions. Frontiers in Plant Science, 7, 346. doi:10.3389/fpls.2016.00346

Clark, S. E., Williams, R. W., & Meyerowitz, E. M. (1997). The CLAVATA1 gene encodes a putative receptor kinase that controls shoot and floral meristem size in arabidopsis. Cell, 89(4), 575-585. doi:10.1016/S0092-8674(00)80239-1

Clark, S. E., Running, M. P., & Meyerowitz, E. M. (1993). CLAVATA1, a regulator of meristem and flower development in arabidopsis. Development (Cambridge), 119(2), 397-418. Retrieved from MEDLINE database. Retrieved from http://dev.biologists.org/content/119/2/397.abstract

Cox, B. EMS mutagenesis in quinoa: Developing a genetic resource. Brigham Young University).

Danielsen, S., Bonifacio, A., & Ames, T. (2003). Diseases of quinoa (chenopodium quinoa). Food Reviews International, 19(1-2), 43-59. doi:10.1081/FRI-120018867

Das, S. (2016). Pseudocereals: An efficient food supplement. Amaranthus: A promising crop of future (pp. 5-11). Singapore: Springer Singapore.

Dill, A., Thomas, S. G., Hu, J., Steber, C. M., & Sun, T. (2004). The arabidopsis F-box protein SLEEPY1 targets gibberellin signaling repressors for gibberellin-induced degradation. The Plant Cell, 16(6), 1392-1405. doi:10.1105/tpc.020958

Fletcher, J. C., Brand, U., Running, M. P., Simon, R., & Meyerowitz, E. M. (1999). Signaling of cell fate decisions by CLAVATA3 in arabidopsis shoot meristems. Science, 283(5409), 1911-1914. doi:10.1126/science.283.5409.1911

Fu, X., Richards, D. E., Fleck, B., Xie, D., Burton, N., & Harberd, N. P. (2004). The arabidopsis mutant sleepy1gar2-1 protein promotes plant growth by increasing the affinity of the SCFSLY1 E3 ubiquitin ligase for DELLA protein substrates. The Plant Cell, 16(6), 1406-1418. doi:10.1105/tpc.021386

Gandarillas, H. (1979). Investigaciones agrícolas. Universo, Boletín Experimental, 34, 35.

Gazzani, S., Gendall, A. R., Lister, C., & Dean, C. (2003). Analysis of the molecular basis of flowering time variation in arabidopsis accessions. Plant Physiology, 132(2), 1107-1114. doi:10.1104/pp.103.021212

Global quinoa seeds market is expected to reach USD 1334.74 million by 2025: Fior markets.(2020, Mar 16,). NASDAQ OMX's News Release Distribution Channel, Retrieved from Business Premium Collection database. Retrieved from https://search.proquest.com/docview/2377423541

Gordon, S. P., Chickarmane, V. S., Ohno, C., & Meyerowitz, E. M. (2009). Multiple feedback loops through cytokinin signaling control stem cell number within the arabidopsis shoot meristem. Proceedings of the National Academy of Sciences - PNAS, 106(38), 16529-16534. doi:10.1073/pnas.0908122106

Greene, E. A., Codomo, C. A., Taylor, N. E., Henikoff, J. G., Till, B. J., Reynolds, S. H., et al. (2003). Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in arabidopsis. Genetics, 164(2), 731-740. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462604/

Griffiths, J., Murase, K., Rieu, I., Zentella, R., Zhang, Z., Powers, S. J., et al. (2006). Genetic characterization and functional analysis of the GID1 gibberellin receptors in arabidopsis. The Plant Cell, 18(12), 3399-3414. doi:10.1105/tpc.106.047415

He, J., Zhao, X., Laroche, A., Lu, Z., Liu, H., & Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Frontiers in Plant Science, 5, 484. doi:10.3389/fpls.2014.00484

Hedden, P. (2020). The current status of research on gibberellin biosynthesis. Plant and Cell Physiology, 61(11), 1832-1849. doi:10.1093/pcp/pcaa092

Henikoff, S., Till, B. J., & Comai, L. (2004). TILLING. traditional mutagenesis meets functional genomics. Plant Physiology, 135(2), 630-636. doi:10.1104/pp.104.041061

Hicks, K. A., Albertson, T. M., & Wagner, D. R. (2001). EARLY FLOWERING3 encodes a novel protein that regulates circadian clock function and flowering in arabidopsis. The Plant Cell, 13(6), 1281-1292. doi:10.1105/tpc.13.6.1281

Holme, I. B., Gregersen, P. L., & Brinch-Pedersen, H. (2019). Induced genetic variation in crop plants by random or targeted mutagenesis: Convergence and differences. Frontiers in Plant Science, 10 doi:10.3389/fpls.2019.01468

Ikeda, A., Ueguchi-Tanaka, M., Sonoda, Y., Kitano, H., Koshioka, M., Futsuhara, Y., et al. (2001). Slender rice, a constitutive gibberellin response mutant, is caused by a null mutation of the SLR1 gene, an ortholog of the height-regulating gene GAI/RGA/RHT/D8. The Plant Cell, 13(5), 999-1010. doi:10.1105/tpc.13.5.999

Imamura, T., Takagi, H., Miyazato, A., Ohki, S., Mizukoshi, H., & Mori, M. (2018). Isolation and characterization of the betalain biosynthesis gene involved in hypocotyl pigmentation of the allotetraploid chenopodium quinoa. Biochemical and Biophysical Research Communications, 496(2), 280-286. doi:10.1016/j.bbrc.2018.01.041

Jankowicz-Cieslak, J., & Till, B. J. Forward and reverse genetics in crop breeding. Advances in plant breeding strategies: Breeding, biotechnology and molecular tools (pp. 215-240). Cham: Springer International Publishing.

Jarvis, D. E., Shwen Ho, Y., Lightfoot, D. J., Schmöckel, S. M., Li, B., Borm, T. J. A., et al. (2017). The genome of chenopodium quinoa. Nature (London), 542(7641), 307-312. doi:10.1038/nature21370

Katwal, T. B., & Bazile, D. (2020). First adaptation of quinoa in the Bhutanese mountain agriculture systems. Plos One, 15(1), e0219804. doi:10.1371/journal.pone.0219804

Khush, G. S. (1999). Green revolution: Preparing for the 21st century. Genome, 42(4), 646-655.

Knapp, S. J. (1998). Marker-Assisted selection as a strategy for increasing the probability of selecting superior genotypes. Crop Science, 38(5), 1164-1174. doi:10.2135/cropsci1998.0011183X003800050009x

Koornneef, M., Elgersma, A., Hanhart, C. J., Loenen-Martinet, v., E.P, Rijn, v., L, & Zeevaart, J. A. D. (1985). A gibberellin insensitive mutant of arabidopsis thaliana. Physiologia Plantarum, 65(1), 33-39. doi:10.1111/j.1399-3054.1985.tb02355.x

Lester, D. R., Ross, J. J., Davies, P. J., & Reid, J. B. (1997). Mendel's stem length gene (le) encodes a gibberellin 3 beta-hydroxylase. The Plant Cell, 9(8), 1435-1443. doi:10.1105/tpc.9.8.1435

Lieberman, M., & Henry, I. (2015). MAPS, mutation and polymorphism survey . Retrieved 02/12/, 2021, from http://comailab.genomecenter.ucdavis.edu/index.php/MAPS

López-Marqués, R. L., Nørrevang, A. F., Ache, P., Moog, M., Visintainer, D., Wendt, T., et al. (2020). Prospects for the accelerated improvement of the resilient crop quinoa. Journal of Experimental Botany, 71(18), 5333-5347. doi:10.1093/jxb/eraa285

McCallum, C. M., Comai, L., Greene, E. A., & Henikoff, S. (2000). Targeting induced local lesions IN genomes (TILLING) for plant functional genomics. Plant Physiology, 123(2), 439-442. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1539256/

McGinnis, K. M., Thomas, S. G., Soule, J. D., Strader, L. C., Zale, J. M., Sun, T., et al. (2003). The arabidopsis SLEEPY1 gene encodes a putative F-box subunit of an SCF E3 ubiquitin ligase. The Plant Cell, 15(5), 1120-1130. doi:10.1105/tpc.010827

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., et al. (2016). The Ensembl variant effect predictor. Genome Biology, 17(1), 122. doi:10.1186/s13059-016-0974-4

Mestanza, C., Riegel, R., Vásquez, S. C., Veliz, D., Cruz-Rosero, N., Canchignia, H., et al. (2018). Discovery of mutations in chenopodium quinoa willd through EMS mutagenesis and mutation screening using pre-selection phenotypic data and next-generation sequencing. The Journal of Agricultural Science, 156(10), 1196-1204. doi:10.1017/S0021859619000182

Michaels, S. D., & Amasino, R. M. (1999). FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. The Plant Cell, 11(5), 949-956. doi:10.1105/tpc.11.5.949

Mo, Y., Howell, T., Vasquez-Gross, H., de Haro, L. A., Dubcovsky, J., & Pearce, S. (2018). Mapping causal mutations by exome sequencing in a wheat TILLING population: A tall mutant case study. Molecular Genetics and Genomics, 293(2), 463-477. doi:10.1007/s00438-017-1401-6

Nelson, S. K., & Steber, C. M. (2017). Transcriptional mechanisms associated with seed dormancy and dormancy loss in the gibberellin-insensitive sly1-2 mutant of arabidopsis thaliana. PloS One, 12(6), e0179143. doi:10.1371/journal.pone.0179143

Peng, J., Carol, P., Richards, D. E., King, K. E., Cowling, R. J., Murphy, G. P., et al. (1997). The arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses. Genes & Development, 11(23), 3194-3205. doi:10.1101/gad.11.23.3194

Ragoussis, J. (2006). Genotyping technologies for all. Drug Discovery Today: Technologies, 3(2), 115-122. doi:https://doi.org/10.1016/j.ddtec.2006.06.013

Ragoussis, J. (2009). Genotyping technologies for genetic research. Annual Review of Genomics and Human Genetics, 10(1), 117-133. doi:10.1146/annurev-genom-082908-150116

Rieu, I., Ruiz-Rivero, O., Fernandez-Garcia, N., Griffiths, J., Powers, S. J., Gong, F., et al. (2008). The gibberellin biosynthetic genes AtGA20ox1 and AtGA20ox2 act, partially redundantly, to promote growth and development throughout the arabidopsis life cycle. The Plant Journal : For Cell and Molecular Biology, 53(3), 488-504. doi:10.1111/j.1365-313X.2007.03356.x

Risi C, J., & Galwey, N. W. (1984). Chenopodium grains of the andes: Inca crops for modern agriculture. Advances in Applied Biology, 10, 145-216. Retrieved from https://agris.fao.org/agris-search/search.do?recordID=US201301421301

Rodríguez, J. P., Rahman, H., Thushar, S., & Singh, R. K. (2020). Healthy and resilient cereals and pseudo-cereals for marginal agriculture: Molecular advances for improving nutrient bioavailability. Frontiers in Genetics, 11, 49. doi:10.3389/fgene.2020.00049

Rojas, W., Pinto, M., Alanoca, C., Gomez Pando, L., Leon-Lobos, P., Alercia, A., et al. (2015). Quinoa genetic resources and ex situ conservation FAO.

Ruiz, K., Ruiz, K., Biondi, S., Biondi, S., Oses, R., Oses, R., et al. (2014). Quinoa biodiversity and sustainability for food security under climate change. A review. Agronomy for Sustainable Development, 34(2), 349-359. doi:10.1007/s13593-013-0195-0

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989). Molecular cloning. A laboratory manual, 2nd ed (2nd ed.). New York: Cold spring Harbor.

Savage, G. P. (2016). Saponins☆. In B. Caballero, P. M. Finglas & F. Toldrá (Eds.), Encyclopedia of food and health (pp. 714-716). Oxford: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780123849472006103

Silvestri, V., & Gil, F. (2000).

Allogamy in quinoa. rate in mendoza (argentina). Revista De La Facultad De Ciencias Agrarias, Universidad Nacional De Cuyo, 32(1), 71-76. Retrieved from https://www.cabdirect.org/cabdirect/abstract/20001617384

Sun, T. (2011). The molecular mechanism and evolution of the GA–GID1–DELLA signaling module in plants. Current Biology, 21(9), R338-R345. doi:10.1016/j.cub.2011.02.036

Szakiel, A., Szakiel, A., Pączkowski, C., Pączkowski, C., Henry, M., & Henry, M. (2011). Influence of environmental abiotic factors on the content of saponins in plants. Phytochemistry Reviews, 10(4), 471-491. doi:10.1007/s11101-010-9177-x

Todd, J. J., & Vodkin, L. O. (1996). Duplications that suppress and deletions that restore expression from a chalcone synthase multigene family. The Plant Cell, 8(4), 687-699. doi:10.1105/tpc.8.4.687

Ward, S. (2000). Allotetraploid segregation for single-gene morphological characters in quinoa (chenopodium quinoa} willd.). Euphytica, 116(1), 11-16. doi:10.1023/A:1004070517808

Werner, T., Motyka, V., Strnad, M., & Schmülling, T. (2001). Regulation of plant growth by cytokinin. Proceedings of the National Academy of Sciences - PNAS, 98(18), 10487-10492. doi:10.1073/pnas.171304098

Werner, T., Motyka, V., Laucou, V., Smets, R., Van Onckelen, H., & Schmulling, T. (2003a). Cytokinin-deficient transgenic arabidopsis plants show multiple developmental alterations indicating opposite functions of cytokinins in the regulation of shoot and root meristem activity. The Plant Cell, 15(11), 2532-2550. doi:10.1105/tpc.014928

Werner, T., Motyka, V., Laucou, V., Smets, R., Van Onckelen, H., & Schmulling, T. (2003b). Cytokinin-deficient transgenic arabidopsis plants show multiple developmental alterations indicating opposite functions of cytokinins in the regulation of shoot and root meristem activity. The Plant Cell, 15(11), 2532-2550. doi:10.1105/tpc.014928

Wieczorek, A., & Wright, M. (2012). History of agricultural biotechnology: How crop development has evolved. Nature Education Knowledge, 3(10) Retrieved from https://www.nature.com/scitable/knowledge/library/history-of-agricultural-biotechnology-how-crop-development-25885295/

Willige, B. C., Ghosh, S., Nill, C., Zourelidou, M., Dohmann, E. M. N., Maier, A., et al. (2007). The DELLA domain of GA INSENSITIVE mediates the interaction with the GA INSENSITIVE DWARF1A gibberellin receptor of arabidopsis. The Plant Cell, 19(4), 1209-1220. doi:10.1105/tpc.107.051441

Yan, W., Deng, X. W., Yang, C., & Tang, X. (2021). The genome-wide EMS mutagenesis bias correlates with sequence context and chromatin structure in rice. Frontiers in Plant Science, 12, 579675. doi:10.3389/fpls.2021.579675

Zagotta, M. T., Hicks, K. A., Jacobs, C. I., Young, J. C., Hangarter, R. P., & Meeks-Wagner, D. R. (1996). The arabidopsis ELF3 gene regulates vegetative photomorphogenesis and the photoperiodic induction of flowering. The Plant Journal : For Cell and Molecular Biology, 10(4), 691-702. doi:10.1046/j.1365-313X.1996.10040691.x

Zurita-Silva, A., Fuentes, F., Zamora, P., Jacobsen, S., & Schwember, A. (2014). Breeding quinoa (chenopodium quinoa willd.): Potential and perspectives. Molecular Breeding, 34(1), 13-30. doi:10.1007/s11032-014-0023-5

Supplemental Material

DNA EXTRACTION

EXTRACTION PROTOCOL

(Turn on heat bath to 65˚ C ~1.5 hours, or to speed things up you can boil di water and add it to the water bath, before starting extraction and make the extraction buffer ~10 minutes before beginning this protocol. Don't attempt to do too many samples at once – we do 10 - 20 samples at a time)

1. *Samples should be freeze-dried and ground to powder before beginning protocol.*
2. Add 600 µl of complete, warmed Extraction Buffer (**make sure you added the 2-mercaptoethanol)** to each tube and mix well by inversion.
3. Cap the tube and vortex the sample for 10 seconds.
4. Immediately place tube in 65˚ C water bath. After 15 minutes of heating, invert tube and shake slurry to the bottom of the tube. Return tube to water bath for an additional 10 minutes.
    a. Label two new 1.5 µl tubes during this time. It should have the sample name along with #1 and #2 respectively.
5. Add ⅓ volume (about 200 µl) of 5 M KOAc to tube. Invert to mix well and place on ice for a minimum of 15-20 minutes (don't exceed 30 minutes).
    a. While the sample is on ice, add 400 µl of cold isopropanol to another new, labeled 1.5 µl centrifuge tube and place in the –20˚ C freezer. This should be the final DNA tube.
6. Add an equal amount (600 µl) of presaturated phenol:chloroform solution. (In the phenol:chloroform there are two fluids—be sure to take from the lower fluid. The *top fluid* is a buffer protecting the phenol**. Do not take this fluid**).
7. Invert the tube vigorously by hand (about 15 times) and centrifuge at 12000 rpm for 10 minutes.
8. Carefully transfer the upper aqueous phase containing the DNA into the empty, labeled #1 tube (use 500 µl pipette setting or lower)

    **It is much better to lose 100 µl of DNA than to get all of the DNA and accidentally pipette up some of the middle or lower phase of the tube.

    ****Be very gentle with the DNA after this point. Invert, don't shake.**
9. Perform a Sevag extraction by adding an equal volume of **Sevag solution** (~500 µl of 24:1 chloroform:isoamyl alcohol) to the DNA solution.
10. Invert vigorously by hand (20 times) and centrifuge for 2 minutes at 12000 rpm.
11. Very carefully transfer the upper aqueous phase (use 450 µl or less) to the labeled #2 tube.
12. Add 40 µl of 3 M NaOAc to the #2 tube and gently invert a few times (3 times)
13. Transfer the solution into the tube containing the isopropanol. Cap the tube and invert several times (10 times) to precipitate out the DNA.
14. Ensure that the DNA is fully precipitated before centrifuging by putting it in the -20 C freezer for 30 minutes then invert and swirl the tube repeatedly.
15. Centrifuge the tube for 5 minutes at 12000 rpm. Gently pour off supernatant in the aqueous disposal and place the tube upside-down on a paper towel to wick off remaining supernatant.

16. Add 200 µl of 75% EtOH and gently swish the tube to wash DNA. *If the pellet becomes dislodged centrifuge for 2 minutes at 12000 rpm.*
17. Decant the ethanol and again invert the tube on a paper-towel.
18. Dry the DNA pellet for 10 minutes in the Speed-Vac at 5.1 pressure.
19. Add 200 µl of TeR and allow the sample to sit for 10 minutes.
20. Flick sample to dislodge pellet and place in 37˚ C incubator/shaker for 1 hour at 200 speed. Alternatively, leave the sample on the bench overnight.
21. Ensure the DNA is resuspended by flicking and place in a properly labeled box in the -80 C freezer.

WORKING SOLUTION

Complete DNA extraction buffer (made prior to use):

Ingredients	10 samples	50 samples	100 samples
Phenanthroline	16 mg	80 mg	160 mg
100% Ethanol (EtOH)	80 µl	400 µl	800 µl
Salts Buffer	8 ml	40 ml	80 ml
SDS (Sodium Lauryl Sulfate)	80 mg	400mg	800mg
2(β)-Mercaptoethanol	5.6 µl	28 µl	56 µl
Total volume	~10 ml	~40 ml	~80 ml

Directions:

1. Weigh phenanthroline (Sigma 9375) in a microfuge tube and dissolve in the ethanol (phenanthroline will only dissolve in ethanol).
2. When dissolved add it all to the appropriate amount of salts buffer (in a small beaker, flask, or in a centrifuge tube).
3. Add the SDS, cover and heat to 65˚ C either on a hot plate or in a water bath. If a water bath is used, heat for 5-10 minutes before use.
4. Immediately before use, add 2-Mercaptoethanol to buffer in the fume hood.

DNA ASSESSMENT

If the DNA is frozen, leave it to thaw slowly on ice first then room temperature for best quality retention.

You should be wearing gloves.

Nanodrop:

1. Flick the sample several times to thoroughly suspend the DNA
2. Make sure the Nanodrop is on by pressing the screen (if necessary it will tell you to lower the arm, just do what it says).
3. You will do an assay for dsDNA

4. Make sure you have fresh **pink pipette tips, Kimwipes**, and a sample of **elution buffer** (TeR) that is **nearly identical to the one in the sample**!
5. Simply follow the instructions it gives you using 2 µl of fluid for sample size.
6. **Wipe down the sensor and arm after every sample with a damp (di water only) and dry Kimwipe.**
7. Ignore the concentration and record the 260/280 and the 260/230 readings for each sample.
8. When you are done, press end experiment.

Qubit Fluorometer:

1. Obtain enough 500 µl **thin walled PCR tubes** or clear .65 mL microcentrifuge tubes for all your samples, and an appropriately sized (n * 200 µl) tube for your working solution.
2. Label each sample tube appropriately, **only on the top.**
3. Add (n * 199 µl) of dsDNA BR Assay Buffer Solution to the large tube along with (n * 1 µl) of dsDNA BR Assay Reagent
4. Put 198 – 196 µl of working solution into each sample tube
5. Add 2 – 4 µl of sample to its sample tube, cap, and flick gently.
6. Make sure the Qubit reader is on by pressing the screen. Choose the Broad Range dsDNA assay.
7. Place your sample in the well, close the door and follow the instructions.

VARIANT DETECTION PIPELINE SCRIPTS

1: Raw reads are stored in the fslg_chenopodium workgroup. Within the working directory fslg_jarvis/compute/EMS_project/WGS a directory with the batch name should contain the following sub-directories: trimmed_reads, mapped_reads, mut_calling, vep_ready, and vep_output

2: Within the fslg_chenopodium workgroup run trim.sh in the batch raw reads or raw data folder to trim the raw fq.gz files and change the script to direct the output to the trimmed_reads sub-directory you made earlier within the EMS_project/WGS/batchname working directory. You can use this for loop to run the job script:

```
#for i in *1.fq.gz; do prefix=${i/1.fq.gz}; echo "Processing file $i with the prefix $prefix"; sbatch trim.sh $prefix; done
```

3: Run bwa_mem.sh on the trimmed reads, the script should automatically send the output to mapped_reads without changing the script. You can use this for loop to run the job script:

```
#for i in *_1P.fq.gz;do prefix=${i/_1P.fq.gz};echo "processing $prefix"; sbatch bwa_map.sh $prefix; done
```

4: Within the mapped_reads subdirectory run clean.sh to convert from SAM to BAM and remove duplicates. You can use this for loop to run the job script:

```
#for i in *.bwa_mem.sam; do prefix=${i/.sam}; echo "processing $prefix"; sbatch clean.sh $prefix;done
```

***At this point it may be helpful to break the BAM files into different groups to make the computing load smaller. To do this you could create several subdirectories within mut_calling labeled batchname_sec_{1..4} or {1..8}depending on how many files you want

to work on at once. Then move the files into the sub_directories. This can be done using a while loop after a series of text files have been made with the name of the subdirectory containing the prefix names of the files you want to include.

Example: batch5_sec1.list

EMS123

EMS456

EMS567

EMS789

Use this while loop within the mapped_reads subdirectory to move the files over:

```
#for i in batch*.list; do echo $i; while read line;do echo $line; mv $line.bwa_mem.sort.bam
../mut_calling/${i/.list};done<$i;done
```

***WARNING if you continue with too few BAM files in one group it may affect the results of the MAPS workflow! Try to use 12 files or more per group.

5: Copy mpileup.sh and 1.beta-run-mpileup.py into the mapped_reads directory (or each of the mut_calling/batch*_sec subdirectories) and run mpileup.sh. *** If running in the mapped_reads directory change the mpileup.sh script to direct the output to ../mutcalling

6: Copy mpileup_parser.sh, maps1.sh, maps2.sh, 2.beta-mpileup-parser_vMEM-O.py,3.beta-maps1-v2.py, and 4.mutation-genotyping-pipeline-part-2_v2.py to the mut_calling directory or each batch*_sec subdirectory along with the quinoa reference genome .fasta and index file .fasta.fai. Run mpileup_parser.sh.

7: Run maps1.sh

8: Run maps2.sh

9: Copy vep_prep.sh into the directories containing the final maps2_batch* outputs. Run vep_prep.sh, it isn't necessary to do it in a bash script since it is a simple program that only takes a few seconds.

10: Copy vep.sh to the directories containing the vep.txt files (the vep_prep.sh output) and run the script for each vep.txt file.

11: The final vep output just needs to include the annotation notes from the quinoa reference genome .gff file. this can be done with the join command.

trim.sh

```
#!/bin/bash

#SBATCH --time=12:00:00   # walltime

#SBATCH --ntasks=16   # number of processor cores (i.e. tasks)

#SBATCH --nodes=1   # number of nodes

#SBATCH --mem-per-cpu=8096M   # memory per CPU core

#SBATCH --mail-user=  # email address

#SBATCH --mail-type=END

#SBATCH --mail-type=FAIL

#this is how I ran this script:

#for i in *1.fq.gz; do prefix=${i/1.fq.gz}; echo "Processing file $i with the prefix $prefix"; sbatch trim.sh $prefix; done

module purge

module load conda-pws

module load trimmomatic

trimmomatic PE -threads 16 -basein ${1}1.fq.gz -baseout ../trimmed_reads/$1.fq.gz LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:75
```

bwa_mem.sh

```
#!/bin/bash

#SBATCH --time=24:00:00   # walltime

#SBATCH --ntasks=8   # number of processor cores (i.e. tasks)

#SBATCH --nodes=1   # number of nodes

#SBATCH --mem-per-cpu=8192M    # memory per CPU core

#SBATCH --mail-user=   # email address

#SBATCH --mail-type=END

#SBATCH --mail-type=FAIL

#The job was sent using this script to send one job per sample/family

#for i in *_1P.fq.gz;do prefix=${i/_1P.fq.gz};echo "processing $prefix"; sbatch bwa_mem.sh $prefix; done
```

```
module purge

module load bwa

REFERENCE=quinoa_pb_chicago-2-final_PBJELLY_pilon.fasta

FORWARD_TRIMMED_READS=$1_1P.fq.gz

REVERSE_TRIMMED_READS=$1_2P.fq.gz

OUTPUT_DIRECTORY=../mapped_reads/$1.bwa_mem.sam

bwa mem -M -t 8 $REFERENCE $FORWARD_TRIMMED_READS $REVERSE_TRIMMED_READS
> $OUTPUT_DIRECTORY
```

clean.sh

```
#!/bin/bash

#SBATCH --time=12:00:00   # walltime

#SBATCH --nodes=1

#SBATCH --ntasks=16   # number of processor cores (i.e. tasks)

#SBATCH --mem-per-cpu=8096M   # memory per CPU core

#SBATCH -C rhel7

#SBATCH --mail-user= # email address

#SBATCH --mail-type=END

#SBATCH --mail-type=FAIL

#This script was run in the terminal to run the job for each .sam file

#for i in *.bwa_mem.sam; do prefix=${i/.sam}; echo "processing $prefix"; sbatch clean.sh $prefix;done

#to chunk the data use this do-while loop; while IFS= read -r line || [[ -n "$line" ]];do sbatch clean.sh
$line.w2.bwa_mem; done < batch3_sec_1.list

module load samtools/1.9

samtools view -Sb --threads 8 $1.sam > $1.bam &&

samtools sort -n $1.bam -o $1.sorted.bam &&

samtools fixmate -m $1.sorted.bam $1.fixmate.bam &&

samtools sort $1.fixmate.bam -o $1.sortmate.bam &&

samtools markdup -s -r $1.sortmate.bam $1.markdup.bam &&

samtools sort $1.markdup.bam -o $1.markdup.sort.bam &&
rm $1.sam $1.bam $1.sorted.bam $1.fixmate.bam $1.sortmate.bam $1.markdup.bam
```

mpileup.sh

```
#!/bin/bash
#SBATCH --time=48:00:00   # walltime
#SBATCH --nodes=1
#SBATCH --ntasks=16   # number of processor cores (i.e. tasks)
#SBATCH --mem-per-cpu=50000M    # memory per CPU core
#SBATCH --mail-user=   # email address
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL
module purge
module load python/2.7
module load samtools/1.9
python 1.beta-run-mpileup.py --samtools /apps/samtools/1.9/bin/samtools  --reference_file quinoa_pb_chicago-2-final_PBJELLY_pilon.fasta --output_file EMS$1.mpileup.txt --mapqual 21 --basequal 21 --maxdepth 4000 --bamname .bwa_mem.markdup.sort.bam --threads 16
```

mpileup_parser.sh

```
#!/bin/bash
#SBATCH --time=24:00:00   # walltime
#SBATCH --nodes=1
#SBATCH --ntasks=24   # number of processor cores (i.e. tasks)
#SBATCH --mem-per-cpu=53000M   # memory per CPU core
#SBATCH --mail-user=   # email address
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL
#SBATCH --qos=paulbryf
#SBATCH -C rhel7
```

```
module purge
module load python/2.7
python 2.beta-mpileup-parser_vMEM-O.py  --thread 24 --mpileup_file EMS$1.mpileup.txt
```

maps1.sh

```
#!/bin/bash
#SBATCH --time=72:00:00   # walltime
#SBATCH --ntasks=12   # number of processor cores (i.e. tasks)
#SBATCH --nodes=1   # number of nodes
#SBATCH --mem-per-cpu=50920M  # memory per CPU core
#SBATCH --mail-user=   # email address
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL
module purge
module load python/2.7
python 3.beta-maps1-v2.py -f parsed_EMS$1.mpileup.txt -o maps1_EMS_$1.txt -t 12 -c 6 -C 10000 -b
80 -i 10 -m -H
```

maps2.sh

```
#!/bin/bash
#SBATCH --time=72:00:00   # walltime
#SBATCH --ntasks=1   # number of processor cores (i.e. tasks)
#SBATCH --nodes=1   # number of nodes
#SBATCH --mem-per-cpu=27164M   # memory per CPU core
#SBATCH --mail-user=   # email address
#SBATCH --mail-type=END
#SBATCH --mail-type=FAIL
#SBATCH -C rhel7
module purge
module load python/2.7

python  4.mutation-genotyping-pipeline-part-2_v2.py -f maps1_EMS_$1.txt -o maps2_EMS_$1.txt -p 10
-m
```

vep_prep.sh

```
grep -v "Chrom/Scaffold" $1|awk '{print $1"\t"$2"\t"$2"\t"$5"/"$6"\t""+""\t"$7}'|sed
's/.bwa//'>${1#maps2_}.vep.txt
```

vep.sh

```
#!/bin/bash -l

#SBATCH --time=04:00:00   # walltime

#SBATCH --ntasks=1   # number of processor cores (i.e. tasks)

#SBATCH --nodes=1   # number of nodes

#SBATCH --mem-per-cpu=4024M   # memory per CPU core

#SBATCH --mail-user=   # email address

#SBATCH --mail-type=END

#SBATCH --mail-type=FAIL

conda deactivate

module purge

module load conda-pws

conda activate ensemblvep

prefix=${1/.txt.vep.txt}

vep --format ensembl --custom
/panfs/pan.fsl.byu.edu/scr/grp/fslg_jarvis/EMS_project/WES/mapping_to_gDNA/reference_genome/cust
om_annotations_for_VEP/quinoa_pb_chicago-2-
final_PBJELLY_pilon_renamed_sorted_allscaffolds.functional_blast_sorted_2.gff.gz,quinoa,gff

--fasta
/panfs/pan.fsl.byu.edu/scr/grp/fslg_jarvis/EMS_project/WES/mapping_to_gDNA/reference_genome/quin
oa_pb_chicago-2-final_PBJELLY_pilon.fasta --force_overwrite --input_file $1 --output_file
../vep_output/${prefix}_vep_output.txt --stats_file ../vep_output/${prefix}_vep_output.stats
```