# NAVIGATING THROUGH THE UNCERTAINTY OF GENOTYPING-BY-SEQUENCING DATA IN POLYPLOIDS

BY

WITTNEY MAYS

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Bioinformatics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2020

Urbana, Illinois

Master's Committee:

Associate Professor Erik J. Sacks, Chair
Associate Professor Alexander E. Lipka
Professor Ray Ming
Dr. Lindsay V. Clark

Abstract

The development of genotyping-by-sequencing (GBS) methods has facilitated genomics studies in non-model species, including polyploids. Variant and genotype calling methods have been established for autopolyploids but for a species with a complex genome, such as sugarcane, the level of uncertainty within GBS data increases making trait mapping difficult. Furthermore, variant and genotype calling methods remain a challenge for both recent and ancient allopolyploids (e.g. wheat, maize, soybean, *Miscanthus*), particularly where the reference genome contains highly similar paralogous sequences that do not pair at meiosis. Alignment of sequence tags to the appropriate position within highly duplicated reference genomes remains a challenge inadequately addressed by existing alignment software. Although some variant calling pipelines can discriminate a paralogous locus from a Mendelian locus, the detection of these paralogous loci is typically for the purpose of the exclusion of these loci from the downstream analysis of genomic studies. We explore the significance of eliminating paralogous loci in downstream analysis using a newly developed pipeline developed to sort sequence tags to their correct alignment locations based on the novel $H_{ind}/H_E$ statistic. The goal of this study was to evaluate the sorting pipeline's ability to properly align paralogous loci to the correct position with respect to the reference genome. Three studies were conducted with a population of 400 individuals simulated based upon the *Triticum aestivum*, the reanalysis of a previously published genome-wide study of fusarium head blight in 273 wheat breeding lines, and the reanalysis of a previously published genome-wide study of traits associated with yield in a *Miscanthus* diversity panel. Results from the study suggested that the filtering of sequences using the $H_{ind}/H_E$ statistic underlying polyRAD v1.2 may lead differences in the output of sequences. Further comparison of each output suggested that the output of the novel pipeline, polyRAD, was concentrated in

gene-rich regions compared to other standard variant calling pipelines. From this study, we provide recommendations for future users of the polyRAD v1.2 variant calling pipeline. Overall we recommend that polyRAD v1.2 is more useful for populations of outcrossing species.

*I dedicate this thesis to my family and friends.*

Acknowledgments

Firstly, I would like to express my sincere gratitude to my advisor Erik Sacks for his continuous support, patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Masters study. My sincerest appreciation also goes to Dr. Lindsay Clark who has provided support, patience, and encouragement throughout my graduate studies. It is not often that one finds an advisor and colleague that always finds the time for listening to the little problems and roadblocks. Lindsay has been responsive and given valuable suggestions to every small and big roadblock I have encountered. Her technical and course advice was essential to the completion of this thesis. Many thanks also to the other members of my committee, Dr. Alexander Lipka and Dr. Ray Ming, who also provided me the opportunity to grow through their immense knowledge and support. I am also thankful for their guidance, constructive criticism and advice related to this project.

Table of Contents

# Chapter 1: Literature Review

## 1.1 Advantages and barriers to breeding polyploid crop species

Species or populations within species are often characterized by the number of chromosomes they possess. Species that undergo whole-genome duplication are further characterized as polyploid. Genomics studies have revealed that whole-genome duplication has been a major theme of plant evolution (Jiao & Paterson, 2014; Barabaschi et al., 2012; Tang et al., 2008; Moore et al., 1995). Polyploidization has facilitated speciation in plants and has occurred frequently in nature (Osabe et al., 2012; Renny-Byfield & Wendel, 2014, Moghe & Shiu, 2014, Cui et al., 2006). Polyploids are often characterized into two major categories depending upon the mechanism by which the species was formed. Specifically, polyploid-driven speciation derived from interspecific hybridization, or allopolyploidy, has been found by comparative genomic studies to be more common than autopolyploidy, which is caused by chromosomal doubling within a species (Soltis et al., 2015; Barker et al., 2016). Because of this difference in frequency, less is understood about the mechanisms underlying autopolyploids compared to allopolyploids (Ayala et al., 2000; Doyle & Coate, 2019; Spoelhof et al., 2017). Several assumptions about autopolyploids have been supported strongly through research: *i*) autopolyploids form multivalent and/or random bivalents during meiosis resulting in polysomic inheritance; *ii*) autopolyploids have higher levels of heterozygosity than their diploid progenitors leading to higher genetic variability; and *iii*) for some species, natural populations of autopolyploids can survive despite the irregular meiotic phases leading to unbalanced gametes (aneuploidy) and reduction in fertility (Soltis & Soltis, 2000).

Polyploids have persisted in nature despite known fertility problems due to chromosome mispairing during meiosis, suggesting that polyploids can have strong competitive advantages over diploids (Murat et al., 2010; Lynch & Conery, 2000; Baduel, Bray, Vallejo-Marin, Kolář, & Yant, 2018). Diversification of gene function and genomic complexity in allopolyploids is thought to provide advantages in environments to which their diploid progenitor species were not adapted (Gottlieb, 1973; Chelaifa, Monnier, & Ainouche, 2010). Genome doubling as a major driver of observed diversification has been studied in the Poaceae, Solanaceae, Fabaceae, and Brassicaceae plant families based on nucleotide diversification rates (Dar & Rehman, 2017; Soltis & Soltis, 2009). Greater genomic plasticity via duplicated gene subfunctionalization or neofunctionalization can enable polyploids to acquire differing morphological and physiological characteristics than their diploid relatives (McCarthy et al., 2019; Sato et al., 2012; Otto, 2007). Despite the complications that may arise from polyploidization, polyploids have many potential advantages, including *i*) increased heterosis or hybrid vigor, which may produce a more adaptive plant, *ii*) increased allelic diversity, and *iii*) gene expression changes (Comai, 2005; Estep et al., 2014; Kashkush et al., 2003). For example, Sánchez Vilas & Pannell et al. (2017) examined a population of *Mercurialis annua* with varying ploidy levels, including diploid, tetraploid and hexaploid, and observed that the *M. annua* with a higher ploidy had higher nutrient levels and higher biomass yields (Sato et al., 2012). Varying ploidy levels expressing different levels of variation has also been observed in other polyploid species (Gao et al., 2017; Todd et al., 2017; Jeyasingh et al., 2015).

Disadvantages that are associated with polyploids include *i*) difficulties during mitosis and meiosis *ii*) epigenetic instability *iii*) negative effects from the changes in gene expression and *iv*) aneuploidy (Comai, 2005; Adams et al., 2003; Ramsey & Schemske, 1999; Song & Chen, 2015;

Matzke et al., 1999). These disadvantages can lead to gene expression levels similar to one parent, gene expression levels lower or higher than both parents, or an unequal contribution of gene expression (Chen, 2007). Investigating gene expression levels from non-additive effects is more difficult for polyploid species especially for recent polyploids that lack genomic resources such as a reference genome than diploids (Chang et al., 2010; Combes et al., 2013; Flagel & Wendel, 2010; Hawkins & Yu, 2018). From a commercial breeding perspective, higher levels of heterozygosity are often exploited, and gene redundancy can lead greater genetic stability but without the knowledge of how the gene expression effects phenotypic values it is difficult to exploit the heterozygosity observed (Sattler et al., 2016). This uncertainty has limited the development of varieties in polyploids (Jansky & Spooner, 2018). However, non-additive effects in the expression levels create unpredictable phenotypic values in offspring compared to species that demonstrate predominantly additive effects (Bouvet et al., 2016).

The impact of polyploidy on different species has been widely studied among grasses because many grasses are polyploids (Levy & Feldman, 2002). Among polyploid grasses, commercial sugarcanes are interspecific hybrids that have a genomic contributions primarily from *S. offinarium* (typically octaploid with x = 10) and to a lesser extent *S. spontaneum* (typically octaploid with x = 8), and are a great example of the genomic barriers that can hinder the progress of polyploid breeding pipelines (Ming et al., 2010). With the recent interest in sugarcane being used for bioenergy, Kandel (2018) addressed some of the challenges of developing sugarcane cultivars, in particular the amount of time and resources required. The selection cycle of commercial sugarcane is typically greater than 10 years using traditional breeding methods, in comparison to 1-4 years for model diploid plant species such as maize in which modern genomic and biotechnology techniques are more widely adopted (Mirajkar et al.,

2019; Chaikam et al., 2019). The amount of variation is high in *Saccharum* spp., which makes understanding the allelic variation of this species challenging (Zhang et al., 2012). Barriers that lead to long breeding cycles in commercial sugarcane include poor synchronization and fertility, and high complexity of the sugarcane genome (Kandel 2018). These barriers mentioned are consistent among other grasses considered polyploids (Matsuoka, 2011; Carnahan & Hill, 1961; Ouyang & Zhang, 2013; Griffin et al., 2011).

All polyploid crops do not present the same obstacles to breeding. *Triticum aestivum* (common wheat) is an allohexaploid crop that has been less difficult to breed and is one of the most characterized examples of allopolyploids. The duration of the breeding cycles of wheat is comparable to diploid crop species (Curwen-McAdams & Jones, 2017). Outside of yield-related traits, wheat breeding programs also focus on traits greatly affected by abiotic and biotic stressors such as stem rust resistance and drought tolerance (Chen et al., 2019; Zörb et al., 2018; Olivera et al., 2018; Kulkarni et al., 2017). Many of these traits in wheat have large effects and are only influenced by a few genes; therefore, using modern techniques to introgress the genes in wheat and other allopolyploids is achievable (Bernardo, 2003). However, many genes associated with other traits of can be difficult to identify due to the similarities of sequences within homeologs and overall polyploid nature of the genome (Chen et al., 2018; Blumstein et al., 2020; Leal-Bertioli et al., 2018). Small mutations between homeologs can lead to subtle phenotypic effects but distinguishing all variations of genes working together to amplify variation through additional copies is difficult. In wheat, the squamosa-promoter binding protein (SBP)-box genes are an example of homeologs that have diverged in function, impacting flowering, leaf development, plant architecture and grain yield (Zhang et al., 2017). Studies have suggested that understanding the relationship between polyploids and the diploid progenitors has been the most

appropriate way to unravel the complexity of polyploids (Soltis et al., 2016).The improvement of understanding of polyploids has most recently been facilitated through improved tools developed genomic studies (Pérez-de-Castro et al., 2012).

## 1.2 Genotyping using next-generation sequencing in polyploids

Strategies for crop breeding can be categorized as conventional, molecular, or a mix of the two. Compared to the conventional methods (i.e. phenotypic selection), molecular methods (i.e. marker-assisted selection, or genomic selection) have the potential to increase efficiency while reducing cost (Grover et al., 2012; Elshire et al., 2011; Sattler et al., 2016). Recently genotyping-by-sequencing (GBS), a reduced representation method, has become popular for obtaining molecular markers in crop species (He et al., 2014). GBS sequences a fraction of the whole genome by making use of restriction enzymes, which cleave DNA only at specific short (four to eight basepair) recognition sites. The restriction enzymes cut sites are located randomly throughout the genome, and only fragments of a certain length and/or ending in a certain cut site are sequenced, resulting in a random but reproducible fraction of the DNA being assayed. Table 1 lists genotyping-by-sequencing methods currently available. Many of the reduced representation methods listed in Table 1 derive from restriction site-associated DNA sequencing (RADseq) protocol described in Baird et al. (2008) and double-digestion protocol described Peterson et. al (2012). Reduced representation approaches enhance coverage of the gene-rich regions of the genome while minimizing the effort towards sequencing the repetitive genomic regions (Elshire et al., 2011). Reduced representation approaches have also enhanced the knowledge of many species through simultaneous discovery and genotyping of SNPs.

RADseq and GBS have advantages and disadvantages in terms of depth of coverage and distribution of loci in comparison to other reduced representation sequencing methods such as sequence capture. One major drawback of restriction enzyme-based approaches is that genomic regions of interest may fail to be sequenced if they do not possess restriction cut sites at the appropriate spacing (Puritz et al., 2014; Beissinger et al., 2013; Kim et al., 2016). In a comparison study between RAD-seq and sequence capture, two reduced representation methods, the analysis of previously published data revealed that the sequence capture method (Gnirke et al., 2009) provided more information per locus while RADseq provided more informative nucleotide sites (Harvey et al., 2016). Harvey et al. (2016) also highlighted that both methods produce great coverage for single loci, but RADseq had more coverage across individuals and across loci genome-wide. Though it is preferred in some studies due to random distribution of loci across the genome, RAD-seq had greater variation within coverage than sequence capture, which could introduce genotyping errors and make analysis performed with RAD-seq difficult to reproduce. The comparison study was performed on *Xenops minutus*, a diploid species, therefore how these sequencing methods compare with a more complex genome was not addressed in this study (Harvey et al., 2016). Although polyploids require greater read depth than diploids for accurate genotyping, this reduced representation approach can still provide enough coverage of the genome to be useful and permits single nucleotide polymorphism (SNP) discovery in non-model polyploid species (Garvin et al., 2010; Liu et al., 2014; Berthouly-Salazar et al., 2016).

Reduced representation methods also have drawbacks and advantages relative to whole genome sequencing, but generally have been applied with success in polyploid species. However, as with any method, there are many drawbacks to GBS including an increased error rate for sequencing repetitive regions and non-gene-rich regions and the requirement of additional statistical methods

and bioinformatic tools (Heslot et al., 2013; Wickland et al., 2017). For example, GBS data has lower coverage compared to whole-genome sequencing methods leading to a high missing data rate and high error rate (Chen et al., 2014). These drawbacks can lead to a lowered ability to identify rare variants. Despite the disadvantages, GBS is an effective methodology for breeding populations due to the low cost compared to WGS and has been widely used for important non-model polyploids such as cotton and potato (Zhang et al., 2019; Caruana et al., 2019). Examples of the GBS technology used for improvement of key crops include common wheat *(Triticum aestivum*), an allohexaploid, and maize (Alipour et al., 2017; Kadam et al., 2016), an ancient tetraploid. Although sequencing technology has advanced our understanding of polyploids, many of the unresolved issues that create a complex structure within the genome often have residual effect complicating analysis of polyploid species. Thus, the use of molecular methods in breeding programs has been further developed in diploid species than polyploid species. Despite the disadvantages for polyploid species, GBS methods have been used successfully for many polyploid crops (e.g. potato, blueberry, wheat, cotton, sweet potato, strawberry) (Baral et al., 2018; Vining et al., 2017; Shirasawa et al., 2017).

GBS is comparable to SNP microarrays, another genotyping method that has been widely-used for plant breeding and understanding the genomic architecture of many crop species. SNP microarrays provide high genotyping accuracy, in contrast to GBS, but typically requires more upfront cost compared to GBS, causing it to be used most widely in model species (LaFramboise, 2009). SNP microarray technology allows the integration of reliable markers and has been successful in polyploid species where information is known about the genomic architecture such as *T. aestivum* (Wang et al., 2014). In non-model polyploid species, a hybrid method that incorporates both SNP microarray and GBS has been used to increase the accuracy

of allelic variants observed in polyploid species (Manimekalai et al., 2020). Incorporating known SNPs in the downstream analysis of polyploid species may resolve the issue of allelic ratios that do not behave in a Mendelian manner, which is common in GBS studies (Akhunov et al., 2009). Without the use of known markers, GBS allows SNP discovery by genomic resources such as a reference genome, or the reference genome of a related species when a reference genome for the species being studied is not available (Clark et al., 2019a; Kyriakidou et al., 2018).

### 1.3 Challenges of analyzing molecular markers in polyploids

Simple sequence repeats have historically been utilized for the application of molecular markers in polyploids, but recently developed GBS methods that rely on restriction enzymes have the potential for successful application of molecular markers at a fraction the cost (Cordeiro et al., 2000; Wang et al., 2019; Schie et al., 2014; Clevenger et al., 2018; Stafne et al., 2005). SSR markers are reproducible and informative. If markers from reduced representation methods can achieve reproducibility and be informative, then these methods may also lead to the development of molecular markers (Vieira et al., 2016; Mammadov et al., 2012). The identification of SNPs by these methods has been used for overall crop improvement through linkage maps, marker-assisted selection and genome-wide diversity analysis. One major drawback of reduced representation methods is the coverage due to these methods only sequencing a portion of the genome resulting in important regions of the genome not being characterized (Scheben et al., 2017). More importantly reduced representation methods have created a new barrier, data analysis. Despite these barriers, efforts have been placed towards developing genetic markers in important polyploids (e.g. cotton, wheat, potato, sugarcane) (Koebner & Summers, 2003; Hinze et al., 2017; Li et al., 2015; Yu et al., 2020; Balsalobre et al., 2017).

The interpretation and application of GBS in polyploid species are heavily predicated upon the estimation of genotypes and allele frequencies. The inability to detect differences and over estimation of genetic effects among individuals and populations of polyploid species can largely be attributed to poor genomic coverage and missing data (Dufresne et al., 2014). Approaches to polyploid genetic marker analysis can be categorized as those that simplify the data and analyze it as if the species were diploid, and those that estimate and utilize allele dosages. Treating polyploid genotypic information as if it were from a diploid has been a common approach to enable the use of analyses and software designed for diploids, despite the loss of allele dosage information (Grandke et al., 2017). For example, for genetic mapping in polyploid biparental populations, it is common to use markers that are heterozygous in only one parent at a time (Crawford et al., 2016; Adhikari et al., 2018). Recently, to circumvent the many limitations of analyzing polyploids, software with underlying Bayesian statistical methods have been developed to estimate the allele dosages of polyploids (Gerard et al., 2018; Blischak et al., 2018; Clark et al., 2019b). The Bayesian statistical approaches to the estimation of allele dosages allow the use of recently developed downstream software that does not require the diploidization of genotype information. There are over twenty softwares frequently used for interpreting polyploids (Table 2). Newer developed softwares that cater to polyploid species are tested on simulated datasets, but the testing using empirical populations could better facilitate improvements to newly developed software (Bourke et al., 2018a; Mollinari & Garcia, 2018).

With a plethora of available software to process RAD-seq data, many recommendations across multiple studies evaluating these methods have been published, and the software chosen by the researcher has an impact on the outcome (Stift et al., 2019; Pereira et al., 2018; Larsen et al., 2018; Tinker et al., 2016; Nielsen et al., 2011; Peng et al., 2020). The most appropriate strategy

is ultimately based on the population type, species, sequence read depth, and the missing data rate being below the recommended threshold. A hybrid strategy between the reduced-representation method and SNP microarray technology can potentially be used to increase the power of downstream analyses by combining high-quality genotypes with broad coverage free of ascertainment bias (Koren et al., 2012). In non-model polyploid species, suggestions such as simulating variants from the data, performing the analysis on the simulated datasets, and determining the best software to use based on the population type, species and other parameters set within the software (Gompert & Mock, 2017; Gao et al., 2015). Genome resources such as a reference genome may not be available for non-model species, thus the use of a reference genome for the study species or a close relative is particularly important for maximizing the number and quality of SNPs identified (Payá-Milans et al., 2018).

1.4 Tables

Table 1. The name of 14 reduced representation methods accompanied by a description of the novel aspect of the method that distinguishes the method from others and the citation describing the protocol in depth.

| Name of reduced representation method | Citation | Novel description |
|---|---|---|
| Complexity Reduction of Polymorphic Sequences (CRoPs) | Orsouw et al., 2007 | CRoPs approach for polymorphism discovery combines the power of amplified fragment length polymorphism (AFLP) markers with the novel Genome Sequencer (GS) 20/GS FLX next-generation sequencing technology. |
| Restriction site-associated DNA sequencing | Baird et al., 2008 | Restriction-site associated DNA (RAD) tags are paired with sequencing technology to discover novel SNP markers and simultaneously genotype individuals. |
| Reduced representation library | Van Tassell et al., 2008 | Single-step method for SNP discovery using reduced representation libraries. |
| Multiplexed shotgun genotyping | Andolfatto et al., 2011 | Restriction enzyme digestion of genomic DNA that does not require shearing and repair of DNA prior to adapter ligation. |
| Genotyping-by-sequencing method | Elshire et al., 2011 | Methylation-sensitive restriction enzymes are paired with sequencing technology to discover novel SNP markers while avoiding the digestion of repetitive regions. |

*Table 1 (cont.)*

| | | |
|---|---|---|
| Sequence based genotyping | Truong et al., 2012 | A technology that allows the simultaneous marker discovery and co-dominant scoring of individual species. |
| Double-digest RAD sequencing | Peterson et al., 2012 | The use of two enzymes simultaneously, double digestion, that results in a cost reduction in library production. |
| Two-enzyme genotype-by-sequencing approach | Poland et al., 2012 | The use of two enzymes and a Y-adapter to generate "uniform" GBS libraries. |
| 2b-RAD | Wang et al., 2012 | The method allows for nearly every restriction site in the genome to be screened and genotyped in parallel. |
| ezRAD | Toonen et al., 2013 | Library preparation requires very little technical expertise or laboratory equipment to complete. The library preparation is directly compatible with companies that render sequencing services. |
| Modified GBS | Sonah et al., 2013 | Optimized GBS method through the use of selective primers for the library preparation to increase genome coverage. |
| RESTseq | Stolle & Moritz, 2013 | Optimized for SNP discovery and genotyping through small scale sequencing platforms. |
| SLAF-seq | Sun et al., 2013 | Use of pre-designed barcode system for locus-specific amplification to optimize SNP discovery. |
| RAD Capture (Rapture) | Ali et al., 2016 | Improved RAD protocol that recovers more unique) RAD fragments. |

Table 2. The name of twenty-one softwares developed specifically for polyploid genomic studies. The name of each software is accompanied by the citation describing the novel software. The software tools included are used to assign marker genotypes, assemble haplotypes, generate linkage maps, identify the mode of inheritance, and simulate polyploid populations.

| Analytic Process | Name of software | Citation |
|---|---|---|
| *Genotyping* | | |
| | polysegRatioMM | Baker et al., 2010 |
| | ClusterCall | Schmitz Carley et al., 2017 |
| | updog | Gerard et al., 2018 |
| | SuperMASSA | Pereira et al., 2018 |
| | polyRAD | Clark et al., 2019b |
| | FitTetra | Zych et al., 2019 |
| *Haplotype Assembly* | | |
| | SATlotyper | Neigenfind et al., 2008 |
| | HapCompass | Aguiar & Istrail, 2012 |
| | HapTree | Berger et al., 2014 |
| | SDhaP | Das & Vikalo, 2015 |
| | SHEsisplus | Shen et al., 2016 |

*Table 2 (cont.)*

| | | |
|---|---|---|
| *Linkage Mapping* | | |
| | TetraploidMap | Hackett et al., 2007 |
| | PERGOLA | Grandke et al., 2017 |
| | TetraploidSNPMap | Hackett et al., 2017 |
| | PolyGembler | Zhou et al., 2017 |
| | polymapR | Bourke et al., 2018b |
| | MAPpoly | da Silva Pereira et al., 2020 |
| *Mode of Inheritance* | | |
| | TetraOrgin | Zheng et al., 2016 |
| *Simulation* | | |
| | polySegratio | Baker et al., 2010 |
| | PedigreeSim | Voorrips & Maliepaard, 2012 |
| | HaploSim | Motazedi et al., 2018 |

1.5 Literature Cited

Adams, K. L., Cronn, R., Percifield, R., & Wendel, J. F. (2003). Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(8), 4649–4654. https://doi.org/10.1073/pnas.0630618100

Adhikari, L., Lindstrom, O. M., Markham, J., & Missaoui, A. M. (2018). Dissecting Key Adaptation Traits in the Polyploid Perennial Medicago sativa Using GBS-SNP Mapping. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00934

Aguiar, D., & Istrail, S. (2012). HapCompass: A Fast Cycle Basis Algorithm for Accurate Haplotype Assembly of Sequence Data. *Journal of Computational Biology*, *19*(6), 577–590. https://doi.org/10.1089/cmb.2012.0084

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., & Miller, M. R. (2016). RAD Capture (Rapture): Flexible and Efficient Sequence-Based Genotyping. *Genetics*, *202*(2), 389–400. https://doi.org/10.1534/genetics.115.183665

Alipour, H., Bihamta, M. R., Mohammadi, V., Peyghambari, S. A., Bai, G., & Zhang, G. (2017). Genotyping-by-Sequencing (GBS) Revealed Molecular Genetic Diversity of Iranian Wheat Landraces and Cultivars. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.01293

Akhunov, E., Nicolet, C., & Dvorak, J. (2009). Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. *TAG. Theoretical and Applied*

*Genetics. Theoretische Und Angewandte Genetik*, *119*(3), 507–517. https://doi.org/10.1007/s00122-009-1059-5

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, *21*(4), 610–617. https://doi.org/10.1101/gr.115402.110

Ayala, F. J., Fitch, W. M., & Clegg, M. T. (2000). Variation and evolution in plants and microorganisms: Toward a new synthesis 50 years after Stebbins. *Proceedings of the National Academy of Sciences*, *97*(13), 6941–6944. https://doi.org/10.1073/pnas.97.13.6941

Baduel, P., Bray, S., Vallejo-Marin, M., Kolář, F., & Yant, L. (2018). The "Polyploid Hop": Shifting Challenges and Opportunities Over the Evolutionary Lifespan of Genome Duplications. *Frontiers in Ecology and Evolution*, *6*. https://doi.org/10.3389/fevo.2018.00117

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., & Johnson, E. A. (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLOS ONE*, *3*(10), e3376. https://doi.org/10.1371/journal.pone.0003376

Baker, P., Jackson, P., & Aitken, K. (2010). Bayesian estimation of marker dosage in sugarcane and other autopolyploids. *Theoretical and Applied Genetics*, *120*(8), 1653–1672. https://doi.org/10.1007/s00122-010-1283-z

Balsalobre, T. W. A., da Silva Pereira, G., Margarido, G. R. A., Gazaffi, R., Barreto, F. Z., Anoni, C. O., Cardoso-Silva, C. B., Costa, E. A., Mancini, M. C., Hoffmann, H. P., de

Souza, A. P., Garcia, A. A. F., & Carneiro, M. S. (2017). GBS-based single dosage markers for linkage and QTL mapping allow gene mining for yield-related traits in sugarcane. *BMC Genomics*, *18*. https://doi.org/10.1186/s12864-016-3383-x

Baral, K., Coulman, B., Biligetu, B., & Fu, Y.-B. (2018). Genotyping-by-Sequencing Enhances Genetic Diversity Analysis of Crested Wheatgrass [Agropyron cristatum (L.) Gaertn.]. *International Journal of Molecular Sciences*, *19*(9). https://doi.org/10.3390/ijms19092587

Barabaschi, D., Guerra, D., Lacrima, K., Laino, P., Michelotti, V., Urso, S., Valè, G., & Cattivelli, L. (2012). Emerging Knowledge from Genome Sequencing of Crop Species. *Molecular Biotechnology*, *50*(3), 250–266. https://doi.org/10.1007/s12033-011-9443-1

Barker, M. S., Arrigo, N., Baniaga, A. E., Li, Z., & Levin, D. A. (2016). On the relative abundance of autopolyploids and allopolyploids. *New Phytologist*, *210*(2), 391–398. https://doi.org/10.1111/nph.13698

Berger, E., Yorukoglu, D., Peng, J., & Berger, B. (2014). HapTree: A Novel Bayesian Framework for Single Individual Polyplotyping Using NGS Data. *PLOS Computational Biology*, *10*(3), e1003502. https://doi.org/10.1371/journal.pcbi.1003502

Bernardo, R. (2003). On the effectiveness of early generation selection in self-pollinated crops. *Crop Science*, *43*(4), 1558–1560. https://doi.org/10.2135/cropsci2003.1558

Berthouly-Salazar, C., Mariac, C., Couderc, M., Pouzadoux, J., Floc'h, J.-B., & Vigouroux, Y. (2016). Genotyping-by-Sequencing SNP Identification for Crops without a Reference Genome: Using Transcriptome Based Mapping as an Alternative Strategy. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.00777

Beissinger, T. M., Hirsch, C. N., Sekhon, R. S., Foerster, J. M., Johnson, J. M., Muttoni, G., Vaillancourt, B., Buell, C. R., Kaeppler, S. M., & Leon, N. de. (2013). Marker Density and Read Depth for Genotyping Populations Using Genotyping-by-Sequencing. *Genetics*, *193*(4), 1073–1081. https://doi.org/10.1534/genetics.112.147710

Blischak, P. D., Kubatko, L. S., & Wolfe, A. D. (2018). SNP genotyping and parameter estimation in polyploids using low-coverage sequencing data. *Bioinformatics (Oxford, England)*, *34*(3), 407–415. https://doi.org/10.1093/bioinformatics/btx587

Blumstein, D. M., Campbell, M. A., Hale, M. C., Sutherland, B. J. G., McKinney, G. J., Stott, W., & Larson, W. A. (2020). Comparative Genomic Analyses and a Novel Linkage Map for Cisco (Coregonus artedi) Provide Insights into Chromosomal Evolution and Rediploidization Across Salmonids. *G3: Genes, Genomes, Genetics*, *10*(8), 2863–2878. https://doi.org/10.1534/g3.120.401497

Bourke, P. M., Voorrips, R. E., Visser, R. G. F., & Maliepaard, C. (2018a). Tools for Genetic Studies in Experimental Populations of Polyploids. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00513

Bourke, P. M., van Geest, G., Voorrips, R. E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R. G. F., Arens, P., Smulders, M. J. M., & Maliepaard, C. (2018b). PolymapR-linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. *Bioinformatics (Oxford, England)*, *34*(20), 3496–3502. https://doi.org/10.1093/bioinformatics/bty371

Bouvet, J.-M., Makouanzi, G., Cros, D., & Vigneron, P. (2016). Modeling additive and non-additive effects in a hybrid population using genome-wide genotyping: Prediction accuracy implications. *Heredity*, *116*(2), 146–157. https://doi.org/10.1038/hdy.2015.78

Carnahan, H. L., & Hill, H. D. (1961). Cytology and Genetics of Forage Grasses. *Botanical Review*, *27*(1), 1–162. JSTOR.

Caruana, B. M., Pembleton, L. W., Constable, F., Rodoni, B., Slater, A. T., & Cogan, N. O. I. (2019). Validation of Genotyping by Sequencing Using Transcriptomics for Diversity and Application of Genomic Selection in Tetraploid Potato. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.00670

Chaikam, V., Molenaar, W., Melchinger, A. E., & Boddupalli, P. M. (2019). Doubled haploid technology for line development in maize: Technical advances and prospects. *Theoretical and Applied Genetics*, *132*(12), 3227–3243. https://doi.org/10.1007/s00122-019-03433-x

Chang, P. L., Dilkes, B. P., McMahon, M., Comai, L., & Nuzhdin, S. V. (2010). Homoeolog-specific retention and use in allotetraploid Arabidopsis suecica depends on parent of origin and network partners. *Genome Biology*, *11*(12), R125. https://doi.org/10.1186/gb-2010-11-12-r125

Chelaifa, H., Monnier, A., & Ainouche, M. (2010). Transcriptomic changes following recent natural hybridization and allopolyploidy in the salt marsh species Spartina x townsendii and Spartina anglica (Poaceae). *The New Phytologist*, *186*(1), 161–174. https://doi.org/10.1111/j.1469-8137.2010.03179.x

Chen, N., Hout, C. V. V., Gottipati, S., & Clark, A. G. (2014). Using Mendelian Inheritance To Improve High-Throughput SNP Discovery. *Genetics*, *198*(3), 847–857. https://doi.org/10.1534/genetics.114.169052

Chen, S., Ren, F., Zhang, L., Liu, Y., Chen, X., Li, Y., Zhang, L., Zhu, B., Zeng, P., Li, Z., Larkin, R. M., & Kuang, H. (2018). Unstable Allotetraploid Tobacco Genome due to Frequent Homeologous Recombination, Segmental Deletion, and Chromosome Loss. *Molecular Plant*, *11*(7), 914–927. https://doi.org/10.1016/j.molp.2018.04.009

Chen, Z. J. (2007). Genetic and Epigenetic Mechanisms for Gene Expression and Phenotypic Variation in Plant Polyploids. *Annual Review of Plant Biology*, *58*, 377–406. https://doi.org/10.1146/annurev.arplant.58.032806.103835

Chen, X.-X., Zhang, W., Liang, X.-Y., Liu, Y.-M., Xu, S.-J., Zhao, Q.-Y., Du, Y.-F., Zhang, L., Chen, X.-P., & Zou, C.-Q. (2019). Physiological and developmental traits associated with the grain yield of winter wheat as affected by phosphorus fertilizer management. *Scientific Reports*, *9*(1), 16580. https://doi.org/10.1038/s41598-019-53000-z

Clark, L. V., Dwiyanti, M. S., Anzoua, K. G., Brummer, J. E., Ghimire, B. K., Głowacka, K., Hall, M., Heo, K., Jin, X., Lipka, A. E., Peng, J., Yamada, T., Yoo, J. H., Yu, C. Y., Zhao, H., Long, S. P., & Sacks, E. J. (2019a). Genome-wide association and genomic prediction for biomass yield in a genetically diverse Miscanthus sinensis germplasm panel phenotyped at five locations in Asia and North America. *GCB Bioenergy*, *11*(8), 988–1007. https://doi.org/10.1111/gcbb.12620

Clark, L. V., Lipka, A. E., & Sacks, E. J. (2019b). polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. *G3 (Bethesda, Md.)*, *9*(3), 663–673. https://doi.org/10.1534/g3.118.200913

Clevenger, J., Chu, Y., Chavarro, C., Botton, S., Culbreath, A., Isleib, T. G., Holbrook, C. C., & Ozias-Akins, P. (2018). Mapping Late Leaf Spot Resistance in Peanut (Arachis hypogaea) Using QTL-seq Reveals Markers for Marker-Assisted Selection. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00083

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature Reviews Genetics*, *6*(11), 836–846. https://doi.org/10.1038/nrg1711

Combes, M.-C., Dereeper, A., Severac, D., Bertrand, B., & Lashermes, P. (2013). Contribution of subgenomes to the transcriptome and their intertwined regulation in the allopolyploid Coffea arabica grown at contrasted temperatures. *The New Phytologist*, *200*(1), 251–260. https://doi.org/10.1111/nph.12371

Cordeiro, G. M., Taylor, G. O., & Henry, R. J. (2000). Characterisation of microsatellite markers from sugarcane (Saccharum sp.), a highly polyploid species. *Plant Science*, *155*(2), 161–168. https://doi.org/10.1016/S0168-9452(00)00208-9

Crawford, J., Brown, P. J., Voigt, T., & Lee, D. K. (2016). Linkage mapping in prairie cordgrass (Spartina pectinata Link) using genotyping-by-sequencing. *Molecular Breeding*, *36*(5), 62. https://doi.org/10.1007/s11032-016-0484-9

Curwen-McAdams, C., & Jones, S. S. (2017). Breeding Perennial Grain Crops Based on Wheat. *Crop Science*, *57*(3), 1172–1188. https://doi.org/10.2135/cropsci2016.10.0869

Cui, L., Wall, P. K., Leebens-Mack, J. H., Lindsay, B. G., Soltis, D. E., Doyle, J. J., Soltis, P. S., Carlson, J. E., Arumuganathan, K., Barakat, A., Albert, V. A., Ma, H., & dePamphilis, C. W. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research*, *16*(6), 738–749. https://doi.org/10.1101/gr.4825606

Dar, T.-U.-H., & Rehman, R.-U. (2017). Introduction to Polyploidy. In T.-U.-H. Dar & R.-U. Rehman, *Polyploidy: Recent Trends and Future Perspectives* (pp. 1–13). Springer India. https://doi.org/10.1007/978-81-322-3772-3_1

da Silva Pereira, G., Gemenet, D. C., Mollinari, M., Olukolu, B. A., Wood, J. C., Diaz, F., Mosquera, V., Gruneberg, W. J., Khan, A., Buell, C. R., Yencho, G. C., & Zeng, Z.-B. (2020). Multiple QTL Mapping in Autopolyploids: A Random-Effect Model Approach with Application in a Hexaploid Sweetpotato Full-Sib Population. *Genetics*, *215*(3), 579–595. https://doi.org/10.1534/genetics.120.303080

Das, S., & Vikalo, H. (2015). SDhaP: Haplotype assembly for diploids and polyploids via semi-definite programming. *BMC Genomics*, *16*(1), 260. https://doi.org/10.1186/s12864-015-1408-5

Doyle, J. J., & Coate, J. E. (2019). Polyploidy, the Nucleotype, and Novelty: The Impact of Genome Doubling on the Biology of the Cell. *International Journal of Plant Sciences*, *180*(1), 1–52. https://doi.org/10.1086/700636

Eckardt, N. A. (2008). Epistasis and Genetic Regulation of Variation in the Arabidopsis Metabolome. *The Plant Cell*, *20*(5), 1185–1186. https://doi.org/10.1105/tpc.108.061051

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Estep, M. C., McKain, M. R., Diaz, D. V., Zhong, J., Hodge, J. G., Hodkinson, T. R., Layton, D. J., Malcomber, S. T., Pasquet, R., & Kellogg, E. A. (2014). Allopolyploidy, diversification, and the Miocene grassland expansion. *Proceedings of the National Academy of Sciences*, *111*(42), 15149–15154. https://doi.org/10.1073/pnas.1404177111

Flagel, L. E., & Wendel, J. F. (2010). Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation. *The New Phytologist*, *186*(1), 184–193. https://doi.org/10.1111/j.1469-8137.2009.03107.x

Garvin, M. R., Saitoh, K., & Gharrett, A. J. (2010). Application of single nucleotide polymorphisms to non-model species: A technical review. *Molecular Ecology Resources*, *10*(6), 915–934. https://doi.org/10.1111/j.1755-0998.2010.02891.x

Gao, L., Kielsmeier-Cook, J., Bajgain, P., Zhang, X., Chao, S., Rouse, M. N., & Anderson, J. A. (2015). Development of genotyping by sequencing (GBS)- and array-derived SNP markers for stem rust resistance gene Sr42. *Molecular Breeding*, *35*(11), 207. https://doi.org/10.1007/s11032-015-0404-4

Gao, S., Yan, Q., Chen, L., Song, Y., Li, J., Fu, C., & Dong, M. (2017). Effects of ploidy level and haplotype on variation of photosynthetic traits: Novel evidence from two Fragaria species. *PLoS ONE*, *12*(6). https://doi.org/10.1371/journal.pone.0179899

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., & Stephens, M. (2018). Genotyping Polyploids from Messy Sequencing Data. *Genetics*, *210*(3), 789–807. https://doi.org/10.1534/genetics.118.301468

Griffin, P. C., Robin, C., & Hoffmann, A. A. (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. *BMC Biology*, *9*(1), 19. https://doi.org/10.1186/1741-7007-9-19

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S., & Nusbaum, C. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nature Biotechnology*, *27*(2), 182–189. https://doi.org/10.1038/nbt.1523

Gompert, Z., & Mock, K. E. (2017). Detection of individual ploidy levels with genotyping-by-sequencing (GBS) analysis. *Molecular Ecology Resources*, *17*(6), 1156–1167. https://doi.org/10.1111/1755-0998.12657

Gottlieb, L. D. (1973). Genetic control of glutamate oxaloacetate transaminase isozymes in the diploid plant Stephanomeria exigua and its allotetraploid derivative. *Biochemical Genetics*, *9*(1), 97–107. https://doi.org/10.1007/BF00485595

Grandke, F., Ranganathan, S., van Bers, N., de Haan, J. R., & Metzler, D. (2017). PERGOLA: Fast and deterministic linkage mapping of polyploids. *BMC Bioinformatics*, *18*. https://doi.org/10.1186/s12859-016-1416-8

Grover, C. E., Salmon, A., & Wendel, J. F. (2012). Targeted sequence capture as a powerful tool for evolutionary analysis. *American Journal of Botany*, *99*(2), 312–319. https://doi.org/10.3732/ajb.1100323

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., & Brumfield, R. T. (2016). Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Systematic Biology*, *65*(5), 910–924. https://doi.org/10.1093/sysbio/syw036

Hackett, C. A., Milne, I., Bradshaw, J. E., & Luo, Z. (2007). TetraploidMap for Windows: Linkage map construction and QTL mapping in autotetraploid species. *The Journal of Heredity*, *98*(7), 727–729. https://doi.org/10.1093/jhered/esm086

Hackett, C. A., Boskamp, B., Vogogias, A., Preedy, K. F., & Milne, I. (2017). TetraploidSNPMap: Software for Linkage Analysis and QTL Mapping in Autotetraploid Populations Using SNP Dosage Data. *Journal of Heredity*, *108*(4), 438–442. https://doi.org/10.1093/jhered/esx022

He, J., Zhao, X., Laroche, A., Lu, Z.-X., 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*. https://doi.org/10.3389/fpls.2014.00484

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.-L., & Sorrells, M. E. (2013). Impact of Marker Ascertainment Bias on Genomic Selection Accuracy and Estimates of Genetic Diversity. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0074612

Hawkins, C., & Yu, L.-X. (2018). Recent progress in alfalfa (Medicago sativa L.) genomics and genomic selection. *The Crop Journal*, *6*(6), 565–575. https://doi.org/10.1016/j.cj.2018.01.006

Hinze, L. L., Hulse-Kemp, A. M., Wilson, I. W., Zhu, Q.-H., Llewellyn, D. J., Taylor, J. M., Spriggs, A., Fang, D. D., Ulloa, M., Burke, J. J., Giband, M., Lacape, J.-M., Van Deynze, A., Udall, J. A., Scheffler, J. A., Hague, S., Wendel, J. F., Pepper, A. E., Frelichowski, J., … Stelly, D. M. (2017). Diversity analysis of cotton (Gossypium hirsutum L.) germplasm using the CottonSNP63K Array. *BMC Plant Biology*, *17*(1), 37. https://doi.org/10.1186/s12870-017-0981-y

Jansky, S. H., & Spooner, D. M. (2018). The Evolution of Potato Breeding. In *Plant Breeding Reviews* (pp. 169–214). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119414735.ch4

Jeyasingh, P. D., Roy Chowdhury, P., Wojewodzic, M. W., Frisch, D., Hessen, D. O., & Weider, L. J. (2015). Phosphorus use and excretion varies with ploidy level in Daphnia. *Journal of Plankton Research*, *37*(6), 1210–1217. https://doi.org/10.1093/plankt/fbv095

Jiao, Y., & Paterson, A. H. (2014). Polyploidy-associated genome modifications during land plant evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1648), 20130355. https://doi.org/10.1098/rstb.2013.0355

Kadam, D. C., Potts, S. M., Bohn, M. O., Lipka, A. E., & Lorenz, A. J. (2016). Genomic Prediction of Single Crosses in the Early Stages of a Maize Hybrid Breeding Pipeline. *G3: Genes, Genomes, Genetics*, *6*(11), 3443–3453. https://doi.org/10.1534/g3.116.031286

Kandel, R., Yang, X., Song, J., & Wang, J. (2018). Potentials, Challenges, and Genetic and Genomic Resources for Sugarcane Biomass Improvement. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00151

Kashkush, K., Feldman, M., & Levy, A. A. (2003). Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genetics*, *33*(1), 102–106. https://doi.org/10.1038/ng1063

Kim, C., Guo, H., Kong, W., Chandnani, R., Shuang, L.-S., & Paterson, A. H. (2016). Application of genotyping by sequencing technology to a variety of crop breeding programs. *Plant Science*, *242*, 14–22. https://doi.org/10.1016/j.plantsci.2015.04.016

Koebner, R. M. D., & Summers, R. W. (2003). 21st century wheat breeding: Plot selection or plate detection? *Trends in Biotechnology*, *21*(2), 59–63. https://doi.org/10.1016/S0167-7799(02)00036-7

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., Wang, Z., Rasko, D. A., McCombie, W. R., Jarvis, E. D., & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature Biotechnology*, *30*(7), 693–700. https://doi.org/10.1038/nbt.2280

Kulkarni, M., Soolanayakanahally, R., Ogawa, S., Uga, Y., Selvaraj, M. G., & Kagale, S. (2017). Drought Response in Wheat: Key Genes and Regulatory Mechanisms Controlling Root System Architecture and Transpiration Efficiency. *Frontiers in Chemistry*, *5*. https://doi.org/10.3389/fchem.2017.00106

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., & Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01660

LaFramboise, T. (2009). Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Research*, *37*(13), 4181–4193. https://doi.org/10.1093/nar/gkp552

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Larsen, B., Gardner, K., Pedersen, C., Ørgaard, M., Migicovsky, Z., Myles, S., & Toldam-Andersen, T. B. (2018). Population structure, relatedness and ploidy levels in an apple gene bank revealed through genotyping-by-sequencing. *PLOS ONE*, *13*(8), e0201889. https://doi.org/10.1371/journal.pone.0201889

Leal-Bertioli, S. C. M., Godoy, I. J., Santos, J. F., Doyle, J. J., Guimarães, P. M., Abernathy, B. L., Jackson, S. A., Moretzsohn, M. C., & Bertioli, D. J. (2018). Segmental allopolyploidy in action: Increasing diversity through polyploid hybridization and homoeologous recombination. *American Journal of Botany*, *105*(6), 1053–1066. https://doi.org/10.1002/ajb2.1112

Levy, A. A., & Feldman, M. (2002). The Impact of Polyploidy on Grass Genome Evolution. *Plant Physiology*, *130*(4), 1587–1593. https://doi.org/10.1104/pp.015727

Li, H., Vikram, P., Singh, R. P., Kilian, A., Carling, J., Song, J., Burgueno-Ferreira, J. A., Bhavani, S., Huerta-Espino, J., Payne, T., Sehgal, D., Wenzl, P., & Singh, S. (2015). A high

density GBS map of bread wheat and its application for dissecting complex disease resistance traits. *BMC Genomics*, *16*(1). https://doi.org/10.1186/s12864-015-1424-5

Liu, H., Bayer, M., Druka, A., Russell, J. R., Hackett, C. A., Poland, J., Ramsay, L., Hedley, P. E., & Waugh, R. (2014). An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley. *BMC Genomics*, *15*(1), 104. https://doi.org/10.1186/1471-2164-15-104

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012, December 18). *SNP Markers and Their Impact on Plant Breeding* [Review Article]. International Journal of Plant Genomics; Hindawi. https://doi.org/10.1155/2012/728398

Manimekalai, R., Suresh, G., Kurup, H. G., Athiappan, S., & Kandalam, M. (2020). Role of NGS and SNP genotyping methods in sugarcane improvement programs. *Critical Reviews in Biotechnology*, *40*(6), 865–880. https://doi.org/10.1080/07388551.2020.1765730

Matsuoka, Y. (2011). Evolution of Polyploid Triticum Wheats under Cultivation: The Role of Domestication, Natural Hybridization and Allopolyploid Speciation in their Diversification. *Plant and Cell Physiology*, *52*(5), 750–764. https://doi.org/10.1093/pcp/pcr018

Matzke, M. A., Scheid, O. M., & Matzke, A. J. M. (1999). Rapid structural and epigenetic changes in polyploid and aneuploid genomes. *BioEssays*, *21*(9), 761–767. https://doi.org/10.1002/(SICI)1521-1878(199909)21:9<761::AID-BIES7>3.0.CO;2-C

McCarthy, E. W., Landis, J. B., Kurti, A., Lawhorn, A. J., Chase, M. W., Knapp, S., Le Comber, S. C., Leitch, A. R., & Litt, A. (2019). Early consequences of allopolyploidy alter floral evolution in Nicotiana (Solanaceae). *BMC Plant Biology*, *19*. https://doi.org/10.1186/s12870-019-1771-5

Ming, R., Moore, P. H., Wu, K.-K., D'hont, A., Glaszmann, J. C., Tew, T. L., Mirkov, T. E., Silva, J. da, Jifon, J., Rai, M., Schnell, R. J., Brumbley, S. M., Lakshmanan, P., Comstock, J. C., & Paterson, A. H. (2010). Sugarcane Improvement through Breeding and Biotechnology. In *Plant Breeding Reviews* (pp. 15–118). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470650349.ch2

Mirajkar, S. J., Devarumath, R. M., Nikam, A. A., Sushir, K. V., Babu, H., & Suprasanna, P. (2019). Sugarcane (Saccharum spp.): Breeding and Genomics. In J. M. Al-Khayri, S. M. Jain, & D. V. Johnson (Eds.), *Advances in Plant Breeding Strategies: Industrial and Food Crops: Volume 6* (pp. 363–406). Springer International Publishing. https://doi.org/10.1007/978-3-030-23265-8_11

Moghe, G. D., & Shiu, S.-H. (2014). The causes and molecular consequences of polyploidy in flowering plants. *Annals of the New York Academy of Sciences*, *1320*(1), 16–34. https://doi.org/10.1111/nyas.12466

Mollinari, M., & Garcia, A. A. F. (2018). Linkage analysis and haplotype phasing in experimental autopolyploid populations with high ploidy level using hidden Markov models. *BioRxiv*, 415232. https://doi.org/10.1101/415232

Moore, G., Devos, K. M., Wang, Z., & Gale, M. D. (1995). Cereal Genome Evolution: Grasses, line up and form a circle. *Current Biology*, *5*(7), 737–739. https://doi.org/10.1016/S0960-9822(95)00148-5

Motazedi, E., Finkers, R., Maliepaard, C., & de Ridder, D. (2018). Exploiting next-generation sequencing to solve the haplotyping puzzle in polyploids: A simulation study. *Briefings in Bioinformatics*, *19*(3), 387–403. https://doi.org/10.1093/bib/bbw126

Murat, F., Xu, J.-H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., Messing, J., & Salse, J. (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Research*, *20*(11), 1545–1557. https://doi.org/10.1101/gr.109744.110

Neigenfind, J., Gyetvai, G., Basekow, R., Diehl, S., Achenbach, U., Gebhardt, C., Selbig, J., & Kersten, B. (2008). Haplotype inference from unphased SNP data in heterozygous polyploids based on SAT. *BMC Genomics*, *9*, 356. https://doi.org/10.1186/1471-2164-9-356

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Olivera, P. D., Rouse, M. N., & Jin, Y. (2018). Identification of New Sources of Resistance to Wheat Stem Rust in Aegilops spp. In the Tertiary Genepool of Wheat. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01719

Osabe, K., Kawanabe, T., Sasaki, T., Ishikawa, R., Okazaki, K., Dennis, E. S., Kazama, T., & Fujimoto, R. (2012). Multiple Mechanisms and Challenges for the Application of

Allopolyploidy in Plants. *International Journal of Molecular Sciences*, *13*(7), 8696–8721. https://doi.org/10.3390/ijms13078696

Orsouw, N. J. van, Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., Poel, H. van der, Oeveren, J. van, Verstegen, H., & Eijk, M. J. T. van. (2007). Complexity Reduction of Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. *PLOS ONE*, *2*(11), e1172. https://doi.org/10.1371/journal.pone.0001172

Otto, S. P. (2007). The Evolutionary Consequences of Polyploidy. *Cell*, *131*(3), 452–462. https://doi.org/10.1016/j.cell.2007.10.022

Ouyang, Y., & Zhang, Q. (2013). Understanding Reproductive Isolation Based on the Rice Model. *Annual Review of Plant Biology*, *64*(1), 111–135. https://doi.org/10.1146/annurev-arplant-050312-120205

Payá-Milans, M., Olmstead, J. W., Nunez, G., Rinehart, T. A., & Staton, M. (2018). Comprehensive evaluation of RNA-seq analysis pipelines in diploid and polyploid species. *GigaScience*, *7*(12). https://doi.org/10.1093/gigascience/giy132

Peng, Z., Zhao, Z., Clevenger, J. P., Chu, Y., Paudel, D., Ozias-Akins, P., & Wang, J. (2020). Comparison of SNP Calling Pipelines and NGS Platforms to Predict the Genomic Regions Harboring Candidate Genes for Nodulation in Cultivated Peanut. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00222

Pereira, G. S., Garcia, A. A. F., & Margarido, G. R. A. (2018). A fully automated pipeline for quantitative genotype calling from next generation sequencing data in autopolyploids. *BMC Bioinformatics*, *19*(1), 398. https://doi.org/10.1186/s12859-018-2433-6

Pérez-de-Castro, A. M., Vilanova, S., Cañizares, J., Pascual, L., Blanca, J. M., Díez, M. J., Prohens, J., & Picó, B. (2012). Application of Genomic Tools in Plant Breeding. *Current Genomics*, *13*(3), 179. https://doi.org/10.2174/138920212800543084

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Poland, J. A., Brown, P. J., Sorrells, M. E., & Jannink, J.-L. (2012). Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach. *PLOS ONE*, *7*(2), e32253. https://doi.org/10.1371/journal.pone.0032253

Puritz, J. B., Matz, M. V., Toonen, R. J., Weber, J. N., Bolnick, D. I., & Bird, C. E. (2014). Demystifying the RAD fad. *Molecular Ecology*, *23*(24), 5937–5942. https://doi.org/10.1111/mec.12965

Ramsey, J., & Schemske, D. W. (1998). Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annual Review of Ecology and Systematics*, *29*(1), 467–501. https://doi.org/10.1146/annurev.ecolsys.29.1.467

Renny-Byfield, S., & Wendel, J. F. (2014). Doubling down on genomes: Polyploidy and crop plants. *American Journal of Botany*, *101*(10), 1711–1725. https://doi.org/10.3732/ajb.1400119

Sánchez Vilas, J., & Pannell, J. R. (2017). No difference in plasticity between different ploidy levels in the Mediterranean herb Mercurialis annua. *Scientific Reports*, *7*(1), 9484. https://doi.org/10.1038/s41598-017-07877-3

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, C., Shuang, Y., Xu, X., Pan, S., Cheng, S., Liu, X., … Universitat Pompeu Fabra. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, *485*(7400), 635–641. https://doi.org/10.1038/nature11119

Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, *243*(2), 281–296. https://doi.org/10.1007/s00425-015-2450-x

Scheben, A., Batley, J., & Edwards, D. (2017). Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnology Journal*, *15*(2), 149–161. https://doi.org/10.1111/pbi.12645

Schie, S., Chaudhary, R., & Debener, T. (2014). Analysis of a Complex Polyploid Plant Genome using Molecular Markers: Strong Evidence for Segmental Allooctoploidy in Garden Dahlias. *The Plant Genome*, *7*(3), plantgenome2014.01.0002. https://doi.org/10.3835/plantgenome2014.01.0002

Schmitz Carley, C. A., Coombs, J. J., Douches, D. S., Bethke, P. C., Palta, J. P., Novy, R. G., & Endelman, J. B. (2017). Automated tetraploid genotype calling by hierarchical clustering. *Theoretical and Applied Genetics*, *130*(4), 717–726. https://doi.org/10.1007/s00122-016-2845-5

Shen, J., Li, Z., Chen, J., Song, Z., Zhou, Z., & Shi, Y. (2016). SHEsisPlus, a toolset for genetic studies on polyploid species. *Scientific Reports*, *6*. https://doi.org/10.1038/srep24095

Shirasawa, K., Tanaka, M., Takahata, Y., Ma, D., Cao, Q., Liu, Q., Zhai, H., Kwak, S.-S., Cheol Jeong, J., Yoon, U.-H., Lee, H.-U., Hirakawa, H., & Isobe, S. (2017). A high-density SNP genetic map consisting of a complete set of homologous groups in autohexaploid sweetpotato (Ipomoea batatas). *Scientific Reports*, *7*. https://doi.org/10.1038/srep44207

Soltis, P. S., Marchant, D. B., Van de Peer, Y., & Soltis, D. E. (2015). Polyploidy and genome evolution in plants. *Current Opinion in Genetics & Development*, *35*, 119–125. https://doi.org/10.1016/j.gde.2015.11.003

Soltis, P. S., & Soltis, D. E. (2000). The role of genetic and genomic attributes in the success of polyploids. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(13), 7051–7057.

Soltis, D. E., Visger, C. J., Marchant, D. B., & Soltis, P. S. (2016). Polyploidy: Pitfalls and paths to a paradigm. *American Journal of Botany*, *103*(7), 1146–1166. https://doi.org/10.3732/ajb.1500501

Soltis, P. S., & Soltis, D. E. (2009). The Role of Hybridization in Plant Speciation. *Annual*

*Review of Plant Biology*, *60*(1), 561–588. https://doi.org/10.1146/annurev.arplant.043008.092039

Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., Normandeau, É., Laroche, J., Larose, S., Jean, M., & Belzile, F. (2013). An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping. *PLOS ONE*, *8*(1), e54603. https://doi.org/10.1371/journal.pone.0054603

Song, Q., & Chen, Z. J. (2015). Epigenetic and developmental regulation in plant polyploids. *Current Opinion in Plant Biology*, *24*, 101–109. https://doi.org/10.1016/j.pbi.2015.02.007

Spoelhof, J. P., Soltis, P. S., & Soltis, D. E. (2017). Pure polyploidy: Closing the gaps in autopolyploid research. *Journal of Systematics and Evolution*, *55*(4), 340–352. https://doi.org/10.1111/jse.12253

Stafne, E. T., Clark, J. R., Weber, C. A., Graham, J., & Lewers, K. S. (2005). Simple Sequence Repeat (SSR) Markers for Genetic Mapping of Raspberry and Blackberry. *Journal of the American Society for Horticultural Science*, *130*(5), 722–728. https://doi.org/10.21273/JASHS.130.5.722

Stift, M., Kolář, F., & Meirmans, P. G. (2019). Structure is more robust than other clustering methods in simulated mixed-ploidy populations. *Heredity*, *123*(4), 429–441. https://doi.org/10.1038/s41437-019-0247-6

Stolle, E., & Moritz, R. F. A. (2013). RESTseq – Efficient Benchtop Population Genomics with RESTriction Fragment SEQuencing. *PLOS ONE*, *8*(5), e63960. https://doi.org/10.1371/journal.pone.0063960

Sun, X., Liu, D., Zhang, X., Li, W., Liu, H., Hong, W., Jiang, C., Guan, N., Ma, C., Zeng, H., Xu, C., Song, J., Huang, L., Wang, C., Shi, J., Wang, R., Zheng, X., Lu, C., Wang, X., & Zheng, H. (2013). SLAF-seq: An Efficient Method of Large-Scale De Novo SNP Discovery and Genotyping Using High-Throughput Sequencing. *PLOS ONE*, *8*(3), e58700. https://doi.org/10.1371/journal.pone.0058700

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and Collinearity in Plant Genomes. *Science*, *320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Tinker, N. A., Bekele, W. A., & Hattori, J. (2016). Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. *G3: Genes|Genomes|Genetics*, *6*(4), 857–863. https://doi.org/10.1534/g3.115.024596

Todd, R. T., Forche, A., & Selmecki, A. (2017). Ploidy Variation in Fungi – Polyploidy, Aneuploidy, and Genome Evolution. *Microbiology Spectrum*, *5*(4). https://doi.org/10.1128/microbiolspec.FUNK-0051-2016

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: A simplified method for genomic genotyping in non-model organisms. *PeerJ*, *1*, e203. https://doi.org/10.7717/peerj.203

Truong, H. T., Ramos, A. M., Yalcin, F., Ruiter, M. de, Poel, H. J. A. van der, Huvenaars, K. H. J., Hogers, R. C. J., Enckevort, L. J. G. van, Janssen, A., Orsouw, N. J. van, & Eijk, M. J. T. van. (2012). Sequence-Based Genotyping for Marker Discovery and Co-Dominant Scoring in Germplasm and Populations. *PLOS ONE*, *7*(5), e37565. https://doi.org/10.1371/journal.pone.0037565

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Warren, W. C., & Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods*, *5*(3), 247–252. https://doi.org/10.1038/nmeth.1185

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. de F. (2016). Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*, *39*(3), 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Vining, K. J., Salinas, N., Tennessen, J. A., Zurn, J. D., Sargent, D. J., Hancock, J., & Bassil, N. V. (2017). Genotyping-by-sequencing enables linkage mapping in three octoploid cultivated strawberry families. *PeerJ*, *5*. https://doi.org/10.7717/peerj.3731

Voorrips, R. E., & Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC Bioinformatics*, *13*(1), 248. https://doi.org/10.1186/1471-2105-13-248

Wang, S., Meyer, E., McKay, J. K., & Matz, M. V. (2012). 2b-RAD: A simple and flexible method for genome-wide genotyping. *Nature Methods*, *9*(8), 808–810. https://doi.org/10.1038/nmeth.2023

Wang, S., Wong, D., Forrest, K., Allen, A., Chao, S., Huang, B. E., Maccaferri, M., Salvi, S., Milner, S. G., Cattivelli, L., Mastrangelo, A. M., Whan, A., Stephen, S., Barker, G., Wieseke, R., Plieske, J., Lillemo, M., Mather, D., Appels, R., … Akhunov, E. (2014). Characterization of polyploid wheat genomic diversity using a high-density 90 000 single nucleotide polymorphism array. *Plant Biotechnology Journal*, *12*(6), 787–796. https://doi.org/10.1111/pbi.12183

Wang, Y., Shahid, M. Q., Ghouri, F., Ercişli, S., Baloch, F. S., & Nie, F. (2019). Transcriptome analysis and annotation: SNPs identified from single copy annotated unigenes of three polyploid blueberry crops. *PLOS ONE*, *14*(4), e0216299. https://doi.org/10.1371/journal.pone.0216299

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, *18*(1), 586. https://doi.org/10.1186/s12859-017-2000-6

Yu, X., Zhang, M., Yu, Z., Yang, D., Li, J., Wu, G., & Li, J. (2020). An SNP-Based High-Density Genetic Linkage Map for Tetraploid Potato Using Specific Length Amplified Fragment Sequencing (SLAF-Seq) Technology. *Agronomy*, *10*(1), 114. https://doi.org/10.3390/agronomy10010114

Zhang, B., Xu, W., Liu, X., Mao, X., Li, A., Wang, J., Chang, X., Zhang, X., & Jing, R. (2017). Functional Conservation and Divergence among Homoeologs of TaSPL20 and TaSPL21, Two SBP-Box Genes Governing Yield-Related Traits in Hexaploid Wheat. *Plant Physiology*, *174*(2), 1177–1191. https://doi.org/10.1104/pp.17.00113

Zhang, J., Nagai, C., Yu, Q., Pan, Y.-B., Ayala-Silva, T., Schnell, R. J., Comstock, J. C., Arumuganathan, A. K., & Ming, R. (2012). Genome size variation in three Saccharum species. *Euphytica*, *185*(3), 511–519. https://doi.org/10.1007/s10681-012-0664-6

Zhang, S., Cai, Y., Guo, J., Li, K., Peng, R., Liu, F., Roberts, J. A., Miao, Y., & Zhang, X. (2019). Genotyping-by-Sequencing of Gossypium hirsutum Races and Cultivars Uncovers

Novel Patterns of Genetic Relationships and Domestication Footprints. *Evolutionary Bioinformatics Online*, *15*. https://doi.org/10.1177/1176934319889948

Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., & Bink, M. C. A. M. (2016). Probabilistic Multilocus Haplotype Reconstruction in Outcrossing Tetraploids. *Genetics*, *203*(1), 119–131. https://doi.org/10.1534/genetics.115.185579

Zhou, C., Olukolu, B., Gemenet, D. C., Wu, S., Gruneberg, W., Cao, M. D., Fei, Z., Zeng, Z.-B., George, A. W., Khan, A., Yencho, G. C., & Coin, L. J. M. (2017). *Assembly of whole-chromosome pseudomolecules for polyploid plant genomes using outcrossed mapping populations* [Preprint]. Bioinformatics. https://doi.org/10.1101/119271

Zörb, C., Ludewig, U., & Hawkesford, M. J. (2018). Perspective on Wheat Yield and Quality with Reduced Nitrogen Supply. *Trends in Plant Science*, *23*(11), 1029–1037. https://doi.org/10.1016/j.tplants.2018.08.012

Zych, K., Gort, G., Maliepaard, C. A., Jansen, R. C., & Voorrips, R. E. (2019). FitTetra 2.0 – improved genotype calling for tetraploids with multiple population and parental data support. *BMC Bioinformatics*, *20*. https://doi.org/10.1186/s12859-019-2703-y

# Chapter 2: Assessment of the novel polyRAD v1.2 variant calling pipeline's capability to correctly align sequence tags from paralogous loci and its impact on genome-wide association studies of polyploids

## 2.1 Abstract


### Background

The development of genotyping-by-sequencing (GBS) methods has facilitated genomics studies in non-model species, including polyploids. Recently software has been developed to call genotypes in polyploids, but limitations within the available software still present challenges. For example, variant and genotype calling methods have been established for autopolyploids but remain a challenge for both recent and ancient allopolyploids (e.g. wheat, maize, soybean, *Miscanthus*), particularly where the reference genome contains highly similar paralogous sequences that do not pair at meiosis. Alignment of sequence tags to the appropriate position within highly duplicated reference genomes remains a challenge inadequately addressed by existing alignment software. Although some variant calling pipelines can discriminate a paralogous locus from a Mendelian locus, the detection of these paralogous loci is typically for the exclusion of these loci from the downstream analysis of genomic studies, which hinders the opportunity to study these potentially important regions. We explored how to properly navigate through the uncertainty of GBS data and the significance of eliminating paralogous loci in downstream analysis using a newly developed pipeline that sorts sequence tags to their correct alignment locations based on the novel $H_{ind}/H_E$ statistic. In this study, we explored the challenges of variant calling methods in allopolyploids.

Results

Through simulated data we demonstrated that polyRAD's variant calling pipeline can align sequences to the correct position with high accuracy. The evaluations of empirical data further highlighted that the output from polyRAD provides markers concentrated in genomic regions to be included in downstream analysis when the reference genome is utilized. The concentration in genomic regions across the three studies led to a decrease in the number of loci included in the genome-wide analysis performed. Depsite the decrease in genomic coverage, polyRAD identified 78 significant associations observed for all 13 yield component traits assessed in the *Miscanthus* diversity panel compared to 61 and 83 associations identified by UNEAK and TASSEL respectively.

Conclusion

Our study directly addresses a knowledge gap noted by bioinformatic software users by assessing the impact of higher confidence in alignment position. We anticipate that this study and newly developed sorting pipeline of polyRAD v1.2 will result in improved genotyping quality, resulting in improved power for GWAS, GS, trait mapping, and population genetics.

## 2.2 Introduction

Many of humanity's most important crops are polyploids (e.g. wheat, sugarcane, canola, strawberry). Moreover, as polyploidization is a major theme of plant evolution, most diploid crops are either recent or ancient diploidized polyploids (e.g. maize and rice) (Stebbins, 1940; Moore et al., 1995; Tang et al., 2008). Thus, duplicate sequences are common within the genomes of plants and especially within recent polyploids. Next-generation sequencing (NGS)

methods have increased our ability to gain an understanding of polyploids. However, differentiating relatively similar, short sequences produced by NGS is difficult, and thus detecting paralogous loci has become a primary problem in genetic studies of polyploids (Dufresne, Stift, Vergilino, & Mable, 2014). The identification and differentiation of paralogous loci from one another is even more difficult without a reference genome (Gayral et al., 2013). Loci that are not filtered are prone to misaligning to the incorrect region of the genome in the variant calling process and therefore may lead to false conclusions in the downstream analysis (Kyriakidou, Tai, Anglin, Ellis, & Strömvik, 2018). For example, in a previously published study of the species *Robinia pseudoacacia L.*, approximately 20% of the variants detected with restriction site-associated DNA sequencing (RADseq) technology were labeled as paralogous loci (Verdu et al., 2016). Although some variant calling pipelines can discriminate a paralogous locus from a Mendelian locus, the detection of these paralogous loci is typically used to exclude these loci from the downstream analyses in genomics studies, which reduces their power. Ignoring duplicated loci limits our understanding of polyploid genomics. Overcoming this challenge could potentially increase our understanding of the variation in polyploid genomes that contributes to phenotypic diversity and increase the power to detect significant associations within genomic studies.

The development of genotyping by sequencing (GBS), sequence-based genotyping (SBG), and RADseq methods have enabled genomics studies in non-model species (Mastretta-Yanes et al., 2014; Elshire et al., 2011; Truong et al., 2012). These methods provide low-cost incomplete coverage of the entire genome but are prone to a high error rate, which can lead to bias (Gerard, Ferrão, Garcia, & Stephens, 2018; Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013). For most crops, including polyploids, GBS methods have become the preferred cost-effective

solution to obtaining quality sequencing data over other more costly sequencing approaches such as whole-genome sequencing (Chen et al., 2014). These cost-effective reduced-representation NGS methodologies, such as GBS, typically require the removal of most sequencing data in post-sequencing analysis resulting in an increased level of uncertainty that is not observed in more costly whole genome sequencing approaches (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012; Yu & Sun, 2013). Despite the need for filtering of many of the small read sequences from the GBS method, copious amounts of sequences pass through filtering and contribute to the characterization and identification of genomic regions of interest. Without the application of proper statistical methods for analyzing GBS sequence data from polyploids, inferences may be based on spurious interpretations originating from inflated estimates of inbreeding, heterozygote undercalling, and incorrect assumptions of the population structure (Davey et al., 2013). This phenomenon has been observed using simulated data; therefore, the principles that underlie polyploid bioinformatic software are specifically established to circumvent biases and build better-adapted software (Eaton, 2014).

GBS methods are based on the digestion and amplification of millions of sequence reads, but due to the short length of the sequence reads and the minimal variation within, it can be difficult for alignment software (e.g. Bowtie (Langmead & Salzberg, 2012), BWA (Li & Durbin, 2009)) to correctly and separately bin similar sequences that are in fact from different parts of the genome. Similar sequences that result from ancient or recent allopolyploidization, *i.e*. paralogous sequences, can represent homologous loci that possess different alleles or even functions; therefore, it is imperative to differentiate paralogous sequences from one another. In the post-sequencing analysis process, differentiating paralogous sequences from one another is especially difficult for organisms without a reference genome (Ohno, 1970). Though the power for

differentiating paralogous loci of duplicated genomes is higher with a reference genome than without, assigning and aligning the loci to the correct position in the reference genome remains difficult. Conventional variant calling software, used in the post-sequencing analysis, either filter out paralogous loci, arbitrarily assign the paralogous loci to a position in the reference genome, or merge the paralogous loci into one single locus (Nadukkalam Ravindran et al., 2018; Catchen et al., 2013; McKinney et al., 2017). Paralogous loci with low coverage, along with the inability to distinguish homology, are sources of error for GBS data, and the removal of paralogous loci contributes to the problem of missing data in downstream analysis of polyploids (Shafer et al., 2018). Additionally, the ability to correctly estimate allele dosage in polyploids is hampered by the low coverage of GBS data, contributing to errors in assigning tag sequences to the correct paralog, which can lead to erroneous conclusions in genomics studies (Clark, Lipka, & Sacks, 2019a; Gerard, Ferrão, Garcia, & Stephens, 2018).

At least four methods for the identification of paralogous loci have been employed in alignment and genotype calling software to alleviate biased downstream analysis, although each has limitations: 1) A method developed specifically for GBS datasets of self-pollinating species using clustering and a maximum likelihood-based approach allows the assignment of similar sequence tags to the appropriate group of tag sequences without the use of a reference genome (Tinker et al., 2016). 2) Another approach is based on the distribution of heterozygous individuals within a population and the allelic depth ratio within the heterozygous individuals (McKinney et al., 2017). The method includes the approximate proportion of heterozygous alleles in the population and the deviation of each locus from the expected Mendelian ratio (1:1) to detect likely paralogous loci. The loci with allelic depth ratios higher than the expected ratio are typically removed. 3) Another approach identifies loci with an excess of heterozygotes as

compared to expected heterozygotes based on the Hardy-Weinberg equation (HWE) to identify and remove paralogous loci (Lexer et al., 2014). However, because not all loci (i.e. diverged duplicates, isoloci) in polyploids behave in a Mendelian manner, the ratio of observed to expected heterozygotes may not be the most appropriate method to implement in genomic studies (Clark & Schreier, 2017). 4) In contrast to the previously discussed approaches, which are based on the assumptions of population statistics, an alternative method employed in allopolyploid studies identifies suspected paralogous loci based on sequence similarity, and utilization of a threshold based on the expected maximum of sequence similarity (Ravindran, Bentzen, Bradbury, & Beiko, 2018). Generally, each of the methods described can distinguish paralogous loci from non-paralogous loci based on the knowledge of the expected behavior of these populations. Different types of populations present different barriers, therefore, these common solutions typically only cater to one specific population type or a subset of population types.

Recently, a new pipeline in the bioinformatics software polyRAD was developed to sort GBS tag sequences to their correct alignment locations based on the novel $H_{ind}/H_E$ statistic, where $H_{ind}$ is the probability that two reads sampled from a single individual and locus correspond to different alleles, and $H_E$ is the expected heterozygosity at the same single locus (Clark et al., 2020). Employing pre-existing alignment software with the option of accommodating multiple alignments for each tag sequence, the sorting algorithm underlying the polyRAD variant calling pipeline assumes the possibility of multiple alignments with equal alignment scores or that the best alignment may not be correct. The $H_{ind}/H_E$ statistic is calculated for each assumed locus. For any assumed locus, if the $H_{ind}/H_E$ value exceeds the expected $H_{ind}/H_E$ value, these groups of tags are rearranged and the $H_{ind}/H_E$ is recalculated and comparison of the $H_{ind}/H_E$ and the expected

$H_{ind}/H_E$ is reassessed. The polyRAD algorithm performs a tabu search and rearranges tags multiple times assuming that groupings of tags at or below the expected value of $H_{ind}/H_E$ provide the most correct alignment location. The groups of tag sequences that cannot be adequately sorted are removed.

In this study, we investigated the potential value of including paralogous loci in genome-wide association studies using a newly developed pipeline based on read depth and population genetics statistics implemented in the R package, polyRAD (Clark et al., 2020). In particular, to test the potential of $H_{ind}/H_E$ for improving breeding outcomes in allopolyploid crops, we conducted three studies: 1) evaluation of the accuracy of the polyRAD algorithm for assigning sequence tags to paralogs, using a simulated population of bread wheat, *Triticum aestivum* (*T. aestivum*); 2) genome-wide association study (GWAS) of an actual panel of 273 *T. aestivum* breeding lines collected from the Midwest and the Eastern United States, comparing the impact on GWAS of filtering out paralogs using heterozygosity (standard strategy) with the strategy of correcting and using the paralogs with polyRAD; and 3) GWAS of a diversity panel of 568 *Miscanthus sinensis* collected from throughout its native range in East Asia, comparing the polyRAD variant calling pipeline to two previously-published pipelines (Table 3). We hypothesize that implementation of the $H_{ind}/H_E$ statistic to allow for the inclusion of paralogous loci will increase the effectiveness of downstream SNP trait association analyses and lead to the discovery of new significant trait-associated regions within the genome that would typically not be included in traditional analyses in which tags from paralogous loci are discarded.

## 2.3 Methods

### *The $H_{ind}/H_E$ statistic*

In allopolyploid species, diploid-like meiotic behavior is observed, thus many assumptions of diploid species can be applied to allopolyploids. The $H_{ind}/H_E$ statistic for assessing Mendelian behavior of loci based on sequence read depth has been described in a preprint by Clark et al. (2020). The $H_{ind}/H_E$ statistic underlies the sorting algorithm in polyRAD v1.2. The $H_{ind}/H_E$ equation is partially based on the assumptions underlying Hardy-Weinberg equilibrium (HWE) for diploid populations. In a diploid population, the expected heterozygosity for a single locus ($H_E$) can be determined by:

Equation 1.

$$H_E = 1 - \sum_{i=1}^{k} p_i^2$$

In Equation 1, one minus the summation of the squared allele frequency ($p_i$) of the $i^{th}$ allele across a total of $k$ alleles represents the probability of heterozygosity in a diploid. The assumptions of HWE are based on populations of diploid species in which there are two alleles but can be extended to multiple alleles as in Eqn. 1. The value of $H_E$ (Eqn. 1) ranges between zero to one with a value close to zero representing a small amount of heterozygosity whereas a value close to one represents a high level of heterozygosity. Moreover, $H_E$ is a meaningful measure of diversity in both diploid and polyploid populations, as it represents the probability that two alleles drawn at random from the population will be different.

$H_{ind}$ (Eqn. 2) is defined as the probability that if two sequencing reads were sampled without replacement from an NGS genotyping dataset at a given locus in an individual, they would represent different alleles (Clark et al. 2020). The 'ind' in $H_{ind}$ indicates that the statistic is calculated within each individual in the population. The expected value of $H_{ind}$ is:

Equation 2.

$$H_{ind} = \frac{ploidy-1}{ploidy} * (1 - F) * H_{E,}$$

where *F* is the coefficient of inbreeding. Given that inbreeding and population structure can affect inheritance in allopolyploids, *1- F* is included in the $H_{ind}$ equation to estimate the probability that two alleles drawn randomly from an individual will not be identical by descent, which accounts for the genetic similarity among relatives. The ploidy component of the $H_{ind}$, (*ploidy*-1)/*ploidy*, is defined as the probability that two sequencing reads originate from different chromosomes. The product of these two terms, multiplied by $H_E$, is, therefore, the probability that two sequencing reads from one individual will be different from each other. Thus, the $H_{ind}/H_E$ statistic provides a numeric value that is dependent on ploidy and inbreeding, which are assumed to be consistent within the population. The $H_{ind}/H_E$ statistic identifies alleles that may not behave in a Mendelian manner within allopolyploid species (e.g. paralogous sequence variants) in that they will have higher values than expected. Therefore, by using the $H_{ind}/H_E$ statistic we can filter these loci or adjust how alleles are assigned to loci, allowing the inclusion of paralogous loci in downstream analysis.

Empirical estimation of the $H_{ind}/H_E$ statistic in a GBS dataset using the averages across the population is estimated as:

Equation 3.

$$\widehat{H_{ind}}/\widehat{H_E} = \frac{\sum_{m=1}^{n} \hat{H}_{ind,m}/\hat{H}_E}{n} \cong \frac{ploidy-1}{ploidy}\,(1-F)$$

where *n* denotes the total number of individuals and *m* denotes each individual in Eqn. 3. The equation has been described in detail in the preprint by Clark et al. (2020). For example, in a diploid natural population without inbreeding the expected value of $H_{ind}/H_E$ would be 0.5, and Mendelian loci are expected to be around or below this expected value.

*polyRAD v1.2 sorting pipeline*

The goal of the sorting pipeline in polyRAD v1.2 (https://github.com/lvclark/polyRAD) is to retain loci that would have typically been removed in other variant calling pipelines, and correctly assign reads to these loci. In the sorting pipeline, the initial assignment of alleles to each locus is based on sequence similarity to the reference. Groups of tag sequences that are likely to belong to the same locus are then established based on negative read depth correlations. The sorting algorithm implemented in polyRAD v1.2 estimates the $H_{ind}/H_E$ based on read depth distribution, with $H_E$ estimated from allele frequencies averaged across the entire population from ratios of sequence read depth within each individual. The expected value of $H_{ind}/H_E$ acts as a threshold, and sets of alleles assigned to a single locus that exceeds the expected $H_{ind}/H_E$ are rearranged by the optimizing algorithm. When referring to this rearrangement of loci, the term 'sorting' is used. Sorting is performed on a set of tag sequences that aligned at multiple positions during the alignment process. One alignment position per tag is selected as being putatively correct by the sorting algorithm. Each tag sequence can contain multiple alleles found within the span of a sequencing read.

The optimizing algorithm in polyRAD v1.2 identifies tag sequences that represent loci that do not exceed the expected value of $H_{ind}/H_E$ and directly exports the loci into the output file because the rearrangement of these loci is not required. Groups of tag sequences that exceed the expected value of $H_{ind}/H_E$ statistic are rearranged and reevaluated by a flexible optimization method. The assumption underlying the algorithm is that the true alignment position yields loci with $H_{ind}/H_E$ values lower than or near the expected value. The flexible optimization method is iterated twenty-five times and the best solution is included in the output.

*Steps of the polyRAD v1.2 variant calling pipeline*

The steps required in polyRAD's variant calling are summarized below and in Fig. 1:

Step 1. Use TASSEL-GBS to identify all unique tags and read depths of all individuals

1.1. In the TASSEL-GBS pipeline, the GBSSeqToTagDBPlugin command allows tags from FASTQ files to be stored in an SQLite database using the input of the FASTQ files and a keyfile.

The TagExportToFastqPlugin command in the TASSEL-GBS pipeline will retrieve the tags from the database in the previous step and reformat the tags to allow the sequences to be readable by Bowtie2 software. This function will provide a FASTA file used in Step 2. The GetTagTaxaDistFromDBPlugin from the TASSEL-GBS pipeline provides a file of the depth of all tags for all samples in the population.

Step 2. Align tags to reference using alignment software

2.1. Using the bowtie2-build command, an index of the reference genome is created.

2.2. The bowtie2 command performs the alignment of the sequence reads of the FASTA file from the first previous steps to the reference genome. It is imperative to use the -k option along with this command, which will allow multiple tags to align to one position. From the suggestion of the developer, the -k option should represent a number higher than the number of subgenomes present in species. For example, one of the populations used in this study is an allohexaploid species *Triticum aestivum* with three subgenomes, so one would set the -k option to four.

Step 3. Group tags based on unordered sets of alignments

3.1. Using the Sequence/Alignment Map (SAM) file produced from alignment software, the process_sam_multi.py script in the polyRAD variant calling pipeline will generate two separate comma-separated value (.csv) files including 1) a file containing read depth by tag and individual 2) a file indicating the alignment position and number of mutations present in each tag sequence which will be used to reassign loci based on the estimated the $H_{ind}/H_E$ statistic.

3.2. From the suggestion of the software developer, filtering of individuals by average $H_{ind}/H_E$ is performed to remove individuals that are hybrids or not the expected ploidy. The inbreeding coefficient is also estimated based upon the frequency of the average $H_{ind}/H_E$ of each locus in a subset of randomly selected loci.

*Step 4. Sorting pipeline based on the $H_{ind}/H_E$ statistic*

4.1. The process_isoloci.py script uses $H_{ind}/H_E$ to sort tag sequences into putative loci and produce a file with the correct position for tag sequences that align to multiple positions

based on the loci with the value closest the estimated $H_{ind}/H_E$ statistic, as well as sequences that only aligned to one position. The estimated inbreeding and filtered individuals from step 3.2 are included on the command line code entered.

*Study 1: Simulation study to determine accuracy of assignment of tag sequences to alignment locations*

The ability of polyRAD to assign tag sequences to the correct position based on the reference genome was assessed, with the assumption that many of the simulated sequences would align to multiple positions and require rearranging. RAD sequence reads were generated from the *in silico* GBS variant simulation software, RADinitio (Rivera-Colón et al., 2019). The RADinitio simulation software tool emulated the RAD-seq library preparation process using a double digestion protocol with restriction enzymes *Pst*I and *Msp*I, producing sequence reads of 150 base pairs with 20X sequencing coverage (Peterson, Weber, Kay, Fisher, & Hoekstra, 2012). The *Triticum aestivum* (bread wheat) reference genome was used to construct reference alleles and alternative alleles of an experimental population of 400 individuals. The *in silico* phase of genotyping by NGS techniques modeled in the RADinitio simulation tool generated a population with a mutation rate of 7e-08, indel probability of 0.01, and insertion/deletion ratio 1.0, resulting in the extraction of 32,783 loci. From the 32,783 loci extracted only 14, 931 loci were retained for further processing. The other loci were discarded for reasons including lack of a second restriction site, the proximity of the loci to repetitive telomeric regions on the chromosome, or overlapping a cut site of another locus. The FASTA files provided from RADinito were converted to FASTQ files using a custom script that assigned a constant phred quality score to eliminate bias scoring among sequence reads. The average alignment accuracies for 14,931 loci among the novel sorting algorithm of polyRAD's variant calling pipeline, Bowtie2 (Langmead &

Salzberg, 2012), and Burrows-Wheel Aligner (BWA) alignment software (Li & Durbin, 2009) were compared. In particular, differences in the number of tag sequences from the SAM output files, along with the number of tag sequences that aligned correctly and the number of tag sequences that aligned incorrectly, from all three software were compared.

*Study 2: Comparison of the standard variant calling pipeline, TASSEL, with the polyRAD v1.2 sorting pipeline for a panel of 273 wheat breeding lines*

In Study 2, we reanalyzed a dataset for 273 winter wheat (*T. aestivum*) breeding lines that originated from the Midwestern and Eastern United States (Arruda et al., 2016). Of the 273 breeding lines, 185 belonged to the University of Illinois soft red winter wheat breeding program, and the remaining lines were selected from other land grant universities and private companies in the United States. The genome *of T. aestivum* is a hexaploid (2n = 6x = 42) consisting of three subgenomes denoted A, B, and D. The best linear unbiased estimators (BLUEs) of phenotypic measurements for Fusarium head blight (FHB) disease severity, incidence, and incidence-severity kernel index were calculated using a mixed model approach to evaluate resistance to the fungal plant pathogen *Fusarium graminearum*. FHB is an important disease of wheat (Mehta, 2014). GBS libraries were prepared using the protocol described in Poland et al., 2012. Sequence data was obtained by Illumina HiSeq2000 at the University of Illinois W.M. Keck Center for Comparative and Functional Genomics in Urbana, IL. Descriptions of the techniques used to collect phenotypic measurements, isolate DNA, and sequence DNA were described previously (Arruda et al., 2016*).*

Variant calling was performed using two workflows, polyRAD and TASSEL-GBSv2, which were compared to each other in terms of power and sensitivity in GWAS. To ensure the best

comparison, parameters for TASSEL mirrored the previously published study (Arruda et al. 2016), both when TASSEL was used for variant calling and when it was used in Step 1 of the polyRAD pipeline (described above). Using Bowtie2 in Step 2 of the polyRAD pipeline, 2,672,226 sequence reads were aligned to the wheat reference genome using the multiple alignment options in the software. Of these sequences, 30.2% aligned to more than one location in the reference genome. From these sequences, a total of 87,385 markers were identified from the novel sorting algorithm implemented in polyRAD, which filtered to remove loci that were not present in at least 100 individuals. Using the method discussed in 3.2 of the 'Steps of the polyRAD v1.2 variant calling pipeline', we observed two peaks while estimating the $H_{ind}/H_E$ from 1,000 loci randomly selected from the dataset (Fig.2). Based on the peak at 0.1, the inbreeding coefficient was estimated to be 0.8, consistent with wheat being self-fertilizing and highly inbred. The second $H_{ind}/H_E$ peak at 0.9 represented paralogous loci and would have resulted in a negative inbreeding coefficient if those loci were treated as Mendelian. After running the sorting algorithm, genotype calling and imputation were performed using the polyRAD genotype calling function, which takes into account population structure, and no further filtering was performed. Using the TASSEL-GBSv2 pipeline, 32,483 markers were identified, then subjected to filtering identical to that in the previous study (Arruda et al. 2016). The previous study assumed high levels of misaligned sequences before imputation and performed three consecutive filtering criteria for the removal of markers by excluding those with i) missing data greater than 50% ii) the minor allele frequency less than 5% or iii) the percentage of heterozygotes greater than 10%. After filtering, a total of 19,992 markers from TASSEL-GBSv2 were included in the GWAS and used for comparison with those output by polyRAD.

Tag-based haplotypes were output by the polyRAD pipeline, which was used for genotype calling for both variant calling methods (polyRAD and TASSEL). A single tag may span multiple SNPs, and more than two tags may correspond to one locus. Throughout this manuscript, the term 'marker' refers to these tag-based haplotypes in this study. Estimated haplotype dosages were used as numeric genotypes in both Studies 2 and 3.

Genome-wide association analyses were performed with GAPIT3 software (Wang et al. 2018) using the multiple loci mixed linear model. The p-values calculated for each marker were adjusted to reduce false positives based upon the false discovery rate (FDR) method proposed by Benjamini and Hochberg (1995). Markers with adjusted p-values below 0.10 were considered significant in this study, for consistency with the analysis performed by Arruda et al. (2016). Significant associations identified in both studies were reported with respect to the *T. aestivum* reference genome v2.2 published by the International Wheat Genome Sequencing Consortium (Consortium (IWGSC), 2014). In particular, we compared the number of significant associations and their locations with respect to the reference genome between the two variant calling pipelines. Analysis of the significant associations was performed to determine if any of the potential causative variants were shared across both pipelines based on the average linkage disequilibrium (LD) of 1.06 Mb observed in a recent wheat study (Bhatta et al., 2019).

*Study 3: Comparison of the standard variant calling pipelines, UNEAK and TASSEL, with the polyRAD v1.2 sorting pipeline on a diversity panel of 568 M. sinensis*

*Miscanthus* is a relatively recent allopolyploid; thus, paralogous loci in *M. sinensis* are frequent (Swaminathan et al., 2012). With this knowledge, we selected a previously studied diversity

panel of 568 *M. sinensis* accessions to assess the utility of polyRAD's variant calling pipeline (Clark et al., 2019b). The diversity panel was phenotyped at six locations: 1) Sapporo, Japan by Hokkaido University (HU), 2) Leamington, ON by New Energy Farms (NEF), 3) Fort Collins, CO by Colorado State University (CSU), 4) Urbana, IL by the University of Illinois (UI), 5) Chuncheon, Korea by Kangwon National University (KNU), and 6) Zhuji, China by Zhejiang University (ZJU).

In the previous study (Clark et al., 2019b), genetic markers were called with the UNEAK pipeline (Lu et al., 2013), which identifies variants without the use of a reference genome, because a reference genome was not available for *Miscanthus* at the time that the original study was conducted. The output from the novel referenced-based sorting algorithm in polyRAD was compared to the output from the standard reference-based TASSEL pipeline, and with output from the UNEAK pipeline used in the previously published study by Clark et al. (2019b). Tag-based haplotypes were used as markers in the polyRAD and standard TASSEL pipeline. Tag-based haplotypes were used as markers in GWAS regardless of genotyping method. The default setting in polyRAD output tag-based haplotypes and the output of TASSEL was also processed through polyRAD to group SNPs into tags. UNEAK is tag-based but only allows one SNP per tag. In all cases the most common tag for a given locus was omitted, and the remaining tags were used as markers in GWAS, with values ranging from 0 to 2 indicating their estimated copy number. the term 'marker' for this study refers to these tag-based haplotypes for all pipelines. The unified mixed linear model approach implemented in GAPIT3 software (Wang et al., preprint) was performed using false discovery rate (FDR) method proposed by Benjamini and Hochberg (1995), and markers with adjusted p-values below 0.05 were considered significant in the previous study. We performed the polyRAD variant calling method as described for the

wheat panel in Study 2; based on the peak at 0.3, the inbreeding coefficient was estimated to be 0.4. TagDigger software (Clark & Sacks, 2016) was used to compare tag sequences from all three variant calling pipelines. A database of tag sequences was created for each pipeline. The sequences were considered a match when the entire tag matched, a subset of their tags matched, or when the tag sequences were presented as a shorter version of each other.

The number of significant markers from the genome-wide association analysis was compared using the polyRAD variant calling pipeline, the standard TASSEL pipeline (Bradbury et al., 2007), and the UNEAK non-reference SNP discovery pipeline (Lu et al., 2013). The previous GWAS analyses identified 27 significant markers associated with biomass yield and 298 unique markers associated with twelve yield component traits. To maximize the power across all 13 traits studied, we applied the multi-locus mixed-model approach described by Segura et al. (2012) implemented in GAPIT3 software (Wang et al., 2020). The p-values calculated for each marker were adjusted to reduce false positives based upon the false discovery rate (FDR) method proposed by Benjamini and Hochberg (1995), and markers with adjusted p-values below 0.05 were considered significant in this study. A custom script (https://github.com/wittney/polyRAD_eval_scripts.git) was used to evaluate the distance of markers included in the GWAS to the nearest gene, based on the position in the *M. sinensis* v7.1 reference genome (Nordberg et al., 2014). The three pipelines were further compared on the basis of 1) significant associations, based upon the position with respect to the *M. sinensis* reference genome assuming that markers within 1,000 base pairs represent the same associated region based upon the linkage disequilibrium of *M. sinensis* observed in Slavov et al. (2014) , 2) number of markers included in the genome-wide association studies, 3) distance in base pairs of all of markers included in each variant calling to the nearest gene, and 4) distance in base pairs of

significant markers identified in each variant calling pipeline to the nearest gene. Each marker identified from the variant calling pipelines to the nearest gene was categorized as either being within a gene, outside of a gene but within 5 kb to the nearest gene, between the distance of 5 kb and 30 kb to the nearest gene or of a distance greater than 30 kb to the nearest gene. Variant calling pipelines were compared in terms of proportions of markers in each of these four categories.

## 2.4 Results

### *Study 1: Simulation study to determine accuracy of assignment of tag sequences to alignment locations*

Compared to Bowtie2 and BWA, the polyRAD sorting algorithm based on the novel $H_{ind}/H_E$ statistic provided fewer tag sequences to be included in downstream analysis (Table 3). After sorting, a total of 6,320 tag sequences were obtained by polyRAD, whereas the standard alignment option in Bowtie2 produced 7,672 tag sequences, and BWA produced 7,498 (Table 4). Of the total sorted sequences output from polyRAD that were aligned correctly, 46% aligned to multiple positions in the Bowtie2 output. From the sequences that were aligned to multiple locations, 95% of the sequences were sorted correctly by polyRAD. Overall, the alignment error rate of the novel sorting algorithm underlying the polyRAD variant calling pipeline was lower than that of Bowtie2 and BWA alignment softwares (Table 4). Following the steps described in Clark et al. 2020, the multiple alignment option in Bowtie2 alignment software was used to prepare input for the polyRAD pipeline, and a total of 14,931 tag sequences were aligned. The intermediate sorting output file from polyRAD's pipeline revealed that over two-thirds of the total tag sequences from the Bowtie2 alignment software aligned more than one time, with 31%

of tag sequences aligning twice and 46% of tag sequences aligning three times, while only 23% of tag sequences aligned one time. It was not possible to estimate $H_{ind}/H_E$ for 41% of the tag sequences from the Bowtie2 alignment output, likely due to monomorphic loci and/or low read depth, causing polyRAD to exclude these loci from its final output.

Of the 6,320 tag sequences from polyRAD, 133 were aligned incorrectly, which were further assessed in Bowtie2 and BWA. Of the 133 incorrectly aligned tag sequences from polyRAD, 45% were also incorrect in the standard option in Bowtie2. Approximately 18% of the incorrectly aligned tag sequences aligned correctly in Bowtie2, though the many of the tag sequences that were aligned incorrectly in the sorting pipeline were not in the final output of the Bowtie2 standard pipeline. In comparison with BWA, all tag sequences that were incorrectly aligned in polyRAD were also not aligned correctly in BWA.

*Study 2: Comparison of the standard variant calling pipeline, TASSEL, with the polyRAD v1.2 sorting pipeline for a panel of 273 wheat breeding lines*

Though more markers were included in the GWAS using the novel sorting pipeline in polyRAD (n = 87,385) as compared to TASSEL (n = 19,992), fewer significant associations were identified by polyRAD (n = 6) than TASSEL (n = 8) (Table 5). Moreover, two of the significantly associated genomic locations were shared between the pipelines. Approximately 90% of the randomly sampled loci used for the estimation of the $H_{ind}/H_E$ statistic was above the expected value of $H_{ind}/H_E$, indicating the possibility of misalignment and were thus sorted using the novel algorithm in polyRAD (Fig 3.). Two loci in Table 5 that were significant in both of the genotyping approaches were within approximately 1 Mb of one another, on chromosome 4A (0.08 Mb apart) and 6A (1.36 Mb apart). Pairs of significant hits from polyRAD and TASSEL on

these chromosomes could potentially represent the same causative variant given the high linkage disequilibrium in wheat (Wang et al., 2010).

*Study 3: Comparison of the standard variant calling pipelines, UNEAK and TASSEL, with the polyRAD v1.2 sorting pipeline on a diversity panel of 568 M. sinensis*

The previously published GWAS that used the non-reference-based UNEAK pipeline identified 46,177 markers (Clark et al., 2019b). In contrast, with the standard reference-based TASSEL pipeline, 1,024,980 markers were obtained and used in a new GWAS. With the novel sorting algorithm implemented in polyRAD, after filtering based upon removing loci that were not present in at least 100 individuals or were above the maximum allowed $H_{ind}/H_E$ of 0.71, 86,580 markers were identified and used in the subsequent GWAS. Approximately 54% of the loci from polyRAD were estimated to be above the expected $H_{ind}/H_E$ of 0.3 before sorting, indicating the possibility of misalignment and were thus sorted using the novel algorithm in polyRAD (Figure 4). The peak $H_{ind}/H_E$ was slightly above the expected value $H_{ind}/H_E$ of 0.30 (Figure 4a) compared to the $H_{ind}/H_E$ distribution after the rearrangement of loci by polyRAD (Figure 4b). The frequency distribution peak was slightly below the estimated $H_{ind}/H_E$ (Figure 4b), reflecting optimization of $H_{ind}/H_E$ from the correction of tag alignment locations.

The total number of significant associations identified from GWAS based on UNEAK, TASSEL, and polyRAD were 60, 83, and 78, respectively (Table 6). No significant markers were shared among the pipelines. A larger number of markers were included in the GWAS using the TASSEL (n = 1,024,980) standard pipeline in comparison to the UNEAK (n = 46,177) and polyRAD pipelines (n = 86,580). Large differences among the pipelines were observed for which markers were included in GWAS analysis. TASSEL output provided ten times more

markers to be included in the GWAS than polyRAD. polyRAD shared all tag sequences with TASSEL, but shared only 2,931 tag sequences with UNEAK, which were also included in the TASSEL output (Figure 5).

Although the number of markers included in the TASSEL pipeline was higher than in the other variant calling pipelines, many of the markers included in TASSEL were not from gene-rich regions. Although the number of markers shared between polyRAD and UNEAK were few, they were similar in terms of outputting markers concentrated in gene-rich regions (Fig. 6a). With this discovery of only a small number of tag sequences shared between variant calling pipelines, further evaluation revealed that UNEAK displayed the highest percentage of markers within or in close proximity of genes based upon annotated genomic regions. The genome-wide study performed with polyRAD provided an output with 78% of the total markers included either within a gene or within five thousand base pairs of the closest gene. The previously published study performed with UNEAK pipeline provided an output with 79% of the total number of markers included either within a gene or within five thousand base pairs of the closest gene (Figure 6a). Compared to UNEAK and polyRAD, the standard TASSEL output included less markers in genomic regions (Figure 6a). Comparing only the 222 significant associations from all three variant calling pipelines, polyRAD GWAS provided 79% of markers with a distance either within a gene or less than five thousand base pairs away from the closest gene. Despite these differences among the pipelines in terms of distribution of loci across the genome, significant associations from all three pipelines showed similar patterns of close proximity to genes (Fig. 6b).

## 2.5 Discussion

In this study we sought to compare the output and downstream analysis of the UNEAK non-reference pipeline and the reference-based TASSEL variant calling pipeline, as well as the alignment software Bowtie2 and BWA, to the new polyRAD variant calling pipeline. This study addresses the advantages, disadvantage, and potential biases of performing genomic studies with software developed to accommodate allopolyploid species to the standard variant calling pipelines developed for diploids. We expected that the sorting and filtering of polyRAD based upon the $H_{ind}/H_E$ statistic would provide a higher confidence in the position chosen by alignment software, more stringent filtering, and the inclusion of sequences that would typically be omitted in allopolyploid genomic studies. Although in recent years more software catering to polyploids have been developed, softwares that may or may not account for the complex polyploid genomic structure are still commonly used among the polyploid research community (Nguyen et al., 2020; Tong et al., 2020; Jordan et al., 2018; Qu et al., 2017). Across all three of our studies the results demonstrated that i) compared to other alignment software the novel sorting algorithm provides higher accuracy in the alignment position chosen, ii) differing significant associations are identified among GWAS derived from different pipelines depending on how variants are called, and iii) with the use of the reference genome the loci included in the output of polyRAD are more concentrated towards genic regions compared to the TASSEL standard pipeline.

### *Study 1: Simulation study reveals a higher accuracy of assignment of tag sequences to alignment locations*

Overall the error rate of polyRAD was lower than the error rates of Bowtie2 and BWA aligner software, but polyRAD output provided approximately 17% fewer aligned tag sequences than

Bowtie2 and BWA. BWA produced an output that resulted in the greatest number of tag sequences but the inclusion of these tag sequences increased the error rate. Specifically, the inclusion of these misaligned sequences decreased the accuracy rates of BWA and Bowtie2 aligner softwares.

Forty-one percent of the tag sequences removed by polyRAD were filtered because $H_{ind}/H_E$ could not be calculated, presumably due to low read depth. A potential limitation of the $H_{ind}/H_E$ statistic is the need for an accurate estimate of inbreeding for the population (Clark et al. 2020). If inbreeding were overestimated, the $H_{ind}/H_E$ threshold would be too low and result in some Mendelian loci being filtered from the dataset. Our population was simulated without any inbreeding, and because we were following the polyRAD variant calling workflow as if it were an empirical dataset, the inbreeding coefficient was estimated to be zero. From this simulation study we are able remove the potential limitation of the $H_{ind}/H_E$ to confirm the novel sorting algorithm's ability to assign tags and filter tag sequences at a higher accuracy compared to standard pipelines. polyRAD outperformed BWA and Bowtie2 but the stringent filtering of polyRAD leads to an output with fewer sequences.

*Study 2: Comparison of the standard variant calling pipeline, TASSEL, with the polyRAD v1.2 sorting pipeline identifies differing significant associations in a panel of 273 wheat breeding lines*

Although the polyRAD sorting pipeline discarded many markers in Study 1 because their $H_{ind}/H_E$ could not be calculated, we hypothesized that it would reduce the need to filter markers based on heterozygosity, resulting in a net increase in the number of markers available for

GWAS. Compared to the previously published analysis using TASSEL (Arruda et al., 2016), approximately four times more markers were included from polyRAD's novel sorting pipeline. For SEV and INC traits analyzed in Study 2, fewer significant associations were identified using polyRAD's novel sorting pipeline as compared to TASSEL but overall the difference of significant associations was small (6 vs. 8). This may indicate that there is no significant difference between either pipeline's ability to detect potential causative variants in the wheat breeding population, although many other factors contribute to the power to detect significant associations in complex traits such as FHB. For example, a population size greater than the 273 available in this study would be expected to allow for greater potential to detect causative variants (Long & Langley, 1999).

When comparing the significant associations, both TASSEL and polyRAD identified markers on chromosome 6A to be associated with incidence (Table 5). The other commonalities among both pipelines were identified on chromosome 6A and 4A. Other mapping studies have also detected QTL for FHB resistance in wheat on 4A and 6A, suggesting that these regions are truly associated with this trait (Buerstmayr et al., 2009). No significant associations within 1.06 Mb were shared for severity and incidence-severity kernel index traits. The previous study validated the trait associations based upon known regions associated with FHB resistance. SNP-trait associations near the major-effect QTL (Fhb1) for FHB resistance in *T. aestivum* on chromosome 3B, which had been introgressed in 97 of the winter wheat breeding lines included in this study, were found with TASSEL only but for two different traits (Bernardo et al., 2012; Liu et al., 2008; Zhou et al., 2002). We identified approximately 2,080 markers in the polyRAD dataset that were in the genomic region of Fhb1, whereas in the previous study it was suggested that only a few markers included in the analysis were near Fhb1. Only one significant marker was

identified to be associated with the severity trait near Fhb1 in the previous study, but the other three markers in the genomic region near Fhb1 did not meet the significance threshold. We expected the increase of markers within this genomic region by polyRAD relative to TASSEL pipeline to lead to an increased propensity to detect and characterize the Fhb1 region associated with the severity trait. Unfortunately, we did not find many markers associated with the Fhb1 genomic region. We assume that the difference in filtering methods resulted in the removal of markers closer to Fhb1 by TASSEL. This may suggest that there is no significant difference in the ability polyRAD to identify more significant associations. In our study, polyRAD uniquely identified a significant marker association located on chromosome 1B. Notably, previous studies also detected QTL for FHB resistance on 1B that accounted for 12% to 16% of variation (Fuentes et al., 2005; Gilsinger et al., 2005; Shen et al., 2003), indicating high confidence in the ability of polyRAD to facilitate detection of marker-trait associations. The unique marker identified from the TASSEL pipeline on chromosome 7B also has literature supporting the presence of a QTL, but this literature does not support the same level of variation as chromosome 1B (Buerstmayr et al., 2009).

Within this study, two peaks in the histogram to estimate the level of inbreeding were observed (Fig. 2), which suggest that there is a high number of misaligned loci in the dataset. Although the dataset used in this study was based on the estimated inbreeding level of 0.8, when comparing the output of markers to be included in the GWAS there was not a difference in the output of markers based on differing levels of $H_{ind}/H_E$ estimated 0.1 (inbreeding of 0.8) and 0.9 (inbreeding of -0.8).

The difference in output and significant associations observed between the standard TASSEL pipeline and polyRAD could be attributed to the difference in how the markers were called,

filtered and genotyped. polyRAD standard variant calling pipeline filters based upon the $H_{ind}/H_E$ statistic, but filtering based upon observed heterozygosity is common in GBS pipelines and is used to filter markers in the standard TASSEL pipeline (McKinney et al., 2017). This filtering method incorporated in TASSEL is useful in diploid species but the similarity of sequences, repetitive sequences, and paralogous regions of allopolyploid species can weaken the reliability of genotype calls (Perea et al., 2016; Li et al., 2015). Specifically, in outcrossing species such as *Miscanthus*, the overcalling of heterozygotes occurs often (Perea et al.,2016). Because the genotype calling method in this study of genome-wide study was the same across all studies using the reference genome we can infer that the difference in the methods of filtering led to the differences observed in each output.

*Study 3: Comparison of the standard variant calling pipelines, UNEAK and TASSEL, with the polyRAD v1.2 sorting pipeline on a diversity panel of 568 M. sinensis reveals greater coverage in genomic regions*

Overall, all three softwares performed well, but within the genome-wide studies, the polyRAD pipeline resulted in fewer significant associations for all 13 traits. A similar result was also observed in Study 2. We focused our attention to three traits known to be associated with biomass yield including dry biomass yield, compressed circumference and culm length. For these three traits, the polyRAD pipeline (25) identified fewer significant associations than UNEAK (29) and TASSEL (31). Moreover, none of the significant marker-trait associations identified from the output of the three pipelines studied were within 1,000 base pairs of each other, indicating that each method identified different associated regions. These unexpected results led to an evaluation of the tags shared among the variant calling pipelines overall and the proximity of the included tags to the nearest gene, as an indirect method of ascertaining whether a marker

was likely to be linked to a causative variant. Although TASSEL provided ten times more markers than polyRAD, the probability of the markers from the TASSEL pipeline being located less than 5,000 bp to a gene was low (Fig. 6). When we used the reference genome, all the tag sequences from polyRAD were also output by TASSEL (Fig. 5). Thus, the polyRAD output had a higher concentration of gene-rich regions than TASSEL. UNEAK's non-reference output was comparable to polyRAD (79% within or five thousand base pairs within a gene vs 78% within or five thousand base pairs within a gene). Thus, 13,273 of the tags from UNEAK were from gene-rich regions but many of the tags identified were not the same tags that were identified by polyRAD, indicating different genomic coverage. Ultimately, greater power is achieved by increasing the sample size of the population than by increasing the number of markers, thus the greater number of significant associations observed from TASSEL relative to polyRAD and UNEAK may be spurious results (Long & Langley, 1999).

Within this study, one peak in the histogram to estimate the $H_{ind}/H_E$ was observed (Fig. 4), which suggest that there are fewer misaligned loci than Study 2 in the dataset. The difference of the frequency distribution of Figure 4a and Figure 4b displays the effect of the sorting algorithm. This supports our hypothesis that polyRAD may be better suited for natural populations than populations with high levels of inbreeding such as breeding populations.

## 2.6 Conclusions

NGS methods have contributed to greater understanding of polyploids in recent years, but many recommendations from recent NGS studies have not been implemented into software tools specific to polyploids, leading to a disconnect between population studies and translational research. It is largely known that the most popular variant calling softwares available were

developed for diploid species, but few software have been developed to accommodate the complex genomic structure of polyploid species, resulting in a lowered confidence in the alignment position chosen in most variant calling pipelines. Using both simulated and empirical data, we found that the polyRAD variant calling pipeline hones in on high-quality markers, improving downstream analysis by reducing computational time and multiple testing correction in comparison to pipelines that generate large volumes of low-quality markers. We discovered that all pipelines studied, polyRAD, TASSEL and UNEAK, generated markers that were significantly associated with our traits of interest. However, in contrast to TASSEL, polyRAD and UNEAK markers concentrated in gene-rich regions, reducing computational time by generating a smaller dataset without losing many markers in LD with causative loci. For this reason, when the reference genome is available, we recommend using polyRAD variant calling pipeline to minimize the amount of time computationally while concentrating the genome-wide analysis on the detection of genes or genomic regions associated with a trait of interest. This ability to mine high quality markers from GBS data in allopolyploid organisms may also make GBS a more appealing choice in comparison to costly SNP array technologies. We also recognized from this study that more stringent filtering may result in fewer markers being included in the downstream analysis. Further evaluation of the tag sequence output, specifically of Study 1 and Study 3, revealed that the different variant calling pipelines and alignment software provide non-redundant makers. Therefore, as an alternative to the previous recommendation, for allopolyploid species that lack a reference genome, we recommend performing SNP calling with multiple softwares to maximize genomic coverage. These suggestions would be better suited for natural populations over breeding populations because Study 2 revealed that there was minimal filtering in the inbred population compared to the

outcrossing population used in Study 3. This may suggest that polyRAD is more appropriate for natural populations or modifications to the software will be required to accommodate populations with high levels of inbreeding such as breeding populations.

2.7 Tables & Figures

Table 3. Description of three populations used to evaluate the efficacy of a novel DNA sequence tag sorting algorithm in the polyRAD variant calling pipeline.

| | Study 1 | Study 2 | Study 3 |
|---|---|---|---|
| Species | *Triticum aestivum* | *Triticum aestivum* | *Miscanthus sinensis* |
| Number of individuals | 400 | 273 | 568 |
| Population type | Natural | Breeding | Natural |
| Data type | Simulated | Empirical | Empirical |
| Chromosome number | 2n = 42 | 2n = 42 | 2n = 38 |
| Ploidy | Allohexaploid | Allohexaploid | Allotetraploid |
| Inbreeding coefficient | 0.0 | 0.8 | 0.4 |

Table 4. Number of correctly aligned tags compared to the number of incorrectly aligned tags from the 14,931 tag sequences included the output in the *Triticum aestivum* simulation study. The output of the novel sorting pipeline, polyRAD, was compared with BWA (Li & Durbin, 2009) and Bowtie2 (Langmead & Salzberg, 2012) alignment software. The percentage in parentheses indicates the error rate of each software.

| | BWA | Bowtie2 | polyRAD |
|---|---|---|---|
| Number of correctly aligned tags | 7,498 | 7,672 | 6,320 |
| Number of incorrectly aligned tags | 596 (7.9%) | 460 (6.0%) | 133 (2.1%) |

Table 5. Significant associations and their locations with respect to the reference genome from genome-wide association analyses in Study 2 of Fusarium head blight resistance in a panel of bread wheat (*Tricticum aestivum*) breeding lines from the Midwest and Eastern United States using TASSEL standard pipeline from a previously published study (Arruda et al., 2016) and a new variant calling pipeline using polyRAD.

| Trait | Chromosome | Position |
|---|---|---|
| Severity (SEV) | | |
| *TASSEL (Arruda et al., 2016)* | | |
| IWGSC_CSS_3B_scaff_10676713_7175 | 3B | 10676713 |
| Incidence (INC) | | |
| *TASSEL (Arruda et al., 2016)* | | |
| IWGSC_CSS_7DS_scaff_3876750_2023 | 7D | 3876750 |
| IWGSC_CSS_6AL_scaff_5780077_12152 | 6A | 5780077 |
| IWGSC_CSS_4DS_scaff_2300354_4482 | 4D | 2300354 |
| IWGSC_CSS_4AL_scaff_7146617_11335 | 4A | 7146617 |
| IWGSC_CSS_7AS_scaff_4132011_1400 | 7A | 4132011 |
| *polyRAD* | | |
| IWGSC_6AS_V1_4392152-8296 | 6A | 4392152 |
| IWGSC_6AS_V1_4422943-5747 | 6A | 4422943 |
| IWGSC_4AL_V2_7062964-5695 | 4A | 7062964 |
| IWGSC_1BL_V1_3815477-9966 | 1B | 3815477 |
| Incidence–severity–kernel Index (ISK) | | |
| *TASSEL (Arruda et al., 2016)* | | |
| IWGSC_CSS_3B_scaff_10676713_7175 | 3B | 10676713 |
| IWGSC_CSS_7DS_scaff_3876750_2023 | 7D | 3876750 |
| *polyRAD* | | |
| IWGSC_3DS_V1_1085926-0213 | 3D | 1085926 |
| IWGSC_2AL_V1_4296034-7158 | 2A | 4296034 |

Table 6. Number of significant associations from genome-wide association analyses of biomass yield and 12 yield component traits in a panel of *Miscanthus sinensis* collected in six locations (Study 3). Markers were called using the UNEAK pipeline from a previously published study (Clark et al. 2019), the TASSEL-GBSv2 pipeline, and a new variant calling pipeline using polyRAD.

| Trait | ZJU | HU + NEF + CSU + UI + KNU | HU + NEF + CSU + UI + KNU + ZJU | Total |
|---|---|---|---|---|
| *Basal circumference (cm)* | | | | |
| UNEAK | SK | SK | 1 | 1 |
| TASSEL | SK | SK | 4 | 4 |
| polyRAD | SK | SK | 10 | 10 |
| *Compressed circumference (cm)* | | | | |
| UNEAK | 1 | 0 | 0 | 1 |
| TASSEL | 7 | 1 | 1 | 9 |
| polyRAD | 5 | 1 | 2 | 8 |
| *Compressed circumference/basal circumference* | | | | |
| UNEAK | 15 | 0 | 0 | 15 |
| TASSEL | 5 | 2 | 0 | 7 |
| polyRAD | 3 | 1 | 0 | 4 |
| *Culm length (cm)* | | | | |
| UNEAK | 1 | 0 | 1 | 2 |
| TASSEL | 3 | 1 | 1 | 5 |
| polyRAD | 0 | 0 | 4 | 4 |
| *Culm node number* | | | | |
| UNEAK | 1 | 0 | 1 | 2 |
| TASSEL | 2 | 2 | 3 | 7 |

*Table 6 (cont.)*

| | | | | |
|---|---|---|---|---|
| polyRAD | 0 | 3 | 2 | 5 |
| *Culms per footprint (#/cm$^2$ )* | | | | |
| UNEAK | 0 | 2 | 1 | 3 |
| TASSEL | 1 | 0 | 5 | 6 |
| polyRAD | 5 | 1 | 3 | 10 |
| *Culm volume (cm$^3$)* | | | | |
| UNEAK | 0 | 0 | 0 | 0 |
| TASSEL | 0 | 1 | 1 | 2 |
| polyRAD | 0 | 2 | 1 | 3 |
| *Diameter of basal internode (mm)* | | | | |
| UNEAK | 0 | 5 | 5 | 10 |
| TASSEL | 0 | 2 | 3 | 5 |
| polyRAD variant calling pipeline | 0 | 1 | 1 | 2 |
| *Diameter of topmost internode (mm)* | | | | |
| UNEAK | 0 | 0 | 0 | 0 |
| TASSEL | 1 | 2 | 4 | 7 |
| polyRAD | 0 | 3 | 0 | 3 |
| *Dry biomass yield (g/plant)* | | | | |
| UNEAK | 0 | 0 | 26 | 26 |
| TASSEL | 7 | 3 | 7 | 17 |
| polyRAD | 9 | 1 | 3 | 13 |
| *Internode length (cm)* | | | | |
| UNEAK | 0 | 0 | 0 | 0 |
| TASSEL | 0 | 3 | 1 | 4 |

*Table 6 (cont.)*

| | | | | |
|---|---|---|---|---|
| polyRAD | 0 | 1 | 0 | 1 |
| *Proportion of reproductive culms* | | | | |
| UNEAK | 0 | NC | NC | 0 |
| TASSEL | 0 | NC | NC | 0 |
| polyRAD | 0 | NC | NC | 0 |
| *Total number of culms* | | | | |
| UNEAK | 0 | 1 | 0 | 1 |
| TASSEL | 1 | 6 | 3 | 10 |
| polyRAD | 2 | 10 | 4 | 16 |

HU, Hokkaido University in Sapporo, Japan; NEF, New Energy Farms in Leamington, ON; CSU, Colorado State University in Fort Collins, CO; UI, the University of Illinois in Urbana, IL; KNU, Kangwon National University in Chuncheon, Korea; ZJU, Zhejiang University in Zhuji, China.


Figure 1. Overview of the variant calling pipeline implemented in polyRAD v1.2. The highlighted region represents where the sorting pipeline is integrated with TASSEL's variant calling process.

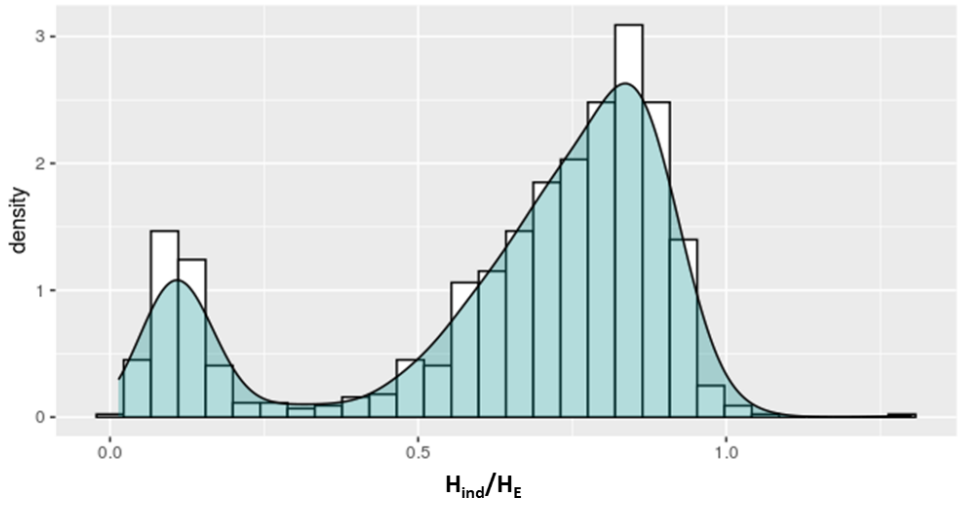



Figure 2. Frequency distribution of $H_{ind}/H_E$ across loci in Study 2 of 273 breeding lines *Triticum aestivum* collected from the Midwest and Eastern United States. Peaks estimated at 0.1 and 0.9 were observed from the 1,000 loci randomly selected from the dataset previously studied (Arruda et al., 2016). The inbreeding was estimated to be 0.8 from peak observed at 0.1 and employed in polyRAD to sort and filter loci included in the dataset.

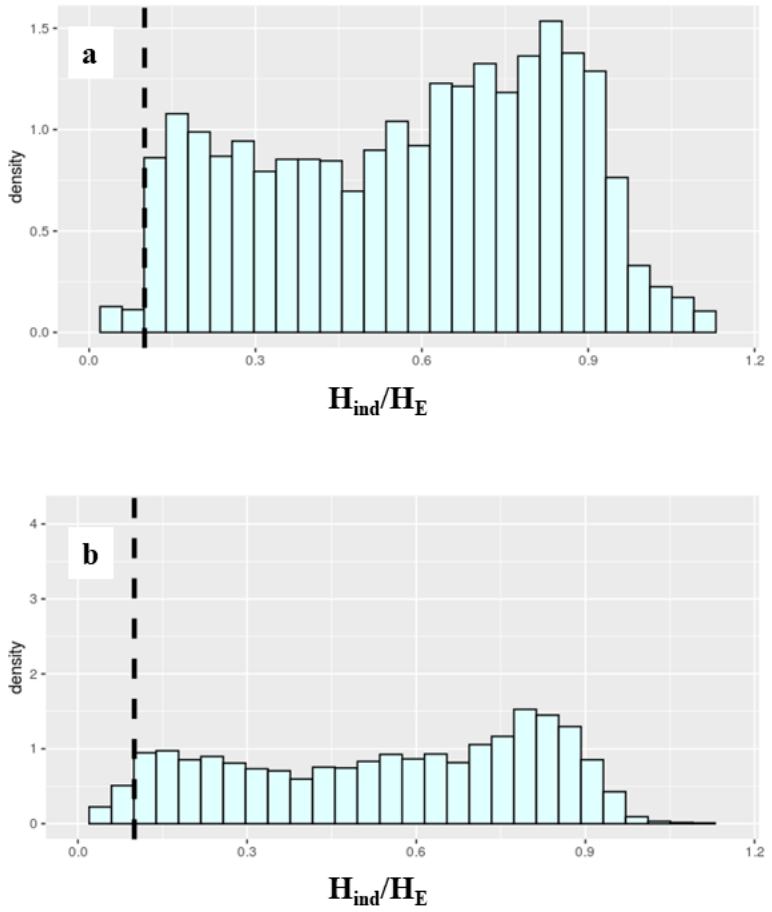



Figure 3. Frequency distribution of $H_{ind}/H_E$ across loci in a previously studied in Study 2 of breeding lines of *Triticum aestivum* collected from the Midwest and Eastern United States (Arruda et al., 2016). The dashed white line represents the expected value of the $H_{ind}/H_E$ assuming a Mendelian locus. Loci above this threshold are expected to be non-Mendelian. (a) The frequency distribution of the $H_{ind}/H_E$ across loci prior to sorting. The loci above the dashed line represent the loci expected to undergo additional sorting by the novel sorting algorithm in polyRAD. (b) The frequency distribution of the $H_{ind}/H_E$ across loci after undergoing additional sorting by the novel sorting algorithm in polyRAD.

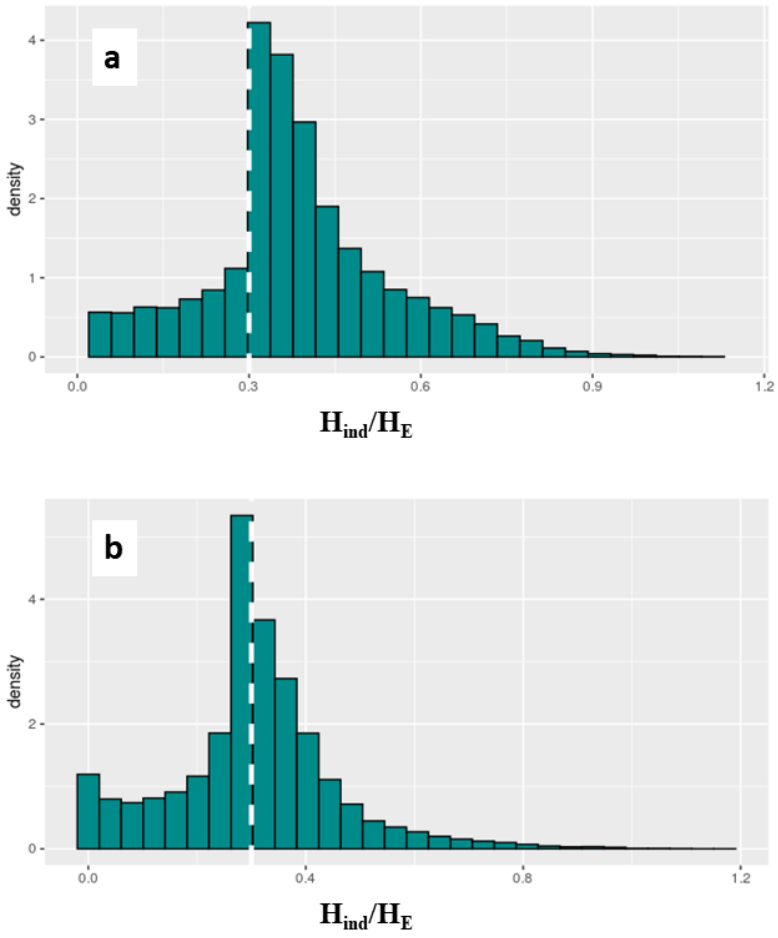



Figure 4. Frequency distribution of $H_{ind}/H_E$ across loci in a previously studied in Study 3 of a *Miscanthus sinensis* diversity panel (Clark et al., 2019). The dashed white line represents the expected value of the $H_{ind}/H_E$ assuming a Mendelian locus. Loci above this threshold are expected to be non-Mendelian. (a) The frequency distribution of the $H_{ind}/H_E$ across loci prior to sorting. The loci above the dashed line represent the loci expected to undergo additional sorting by the novel sorting algorithm in polyRAD. (b) The frequency distribution of the $H_{ind}/H_E$ across loci after undergoing additional sorting by the novel sorting algorithm in polyRAD.

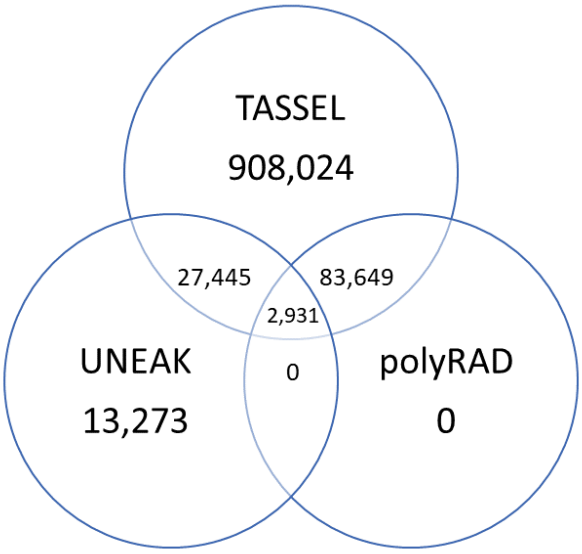



Figure 5**.** Venn diagram of the number of single nucleotide polymorphisms output by three variant calling pipelines, TASSEL, UNEAK and polyRAD, for a *Miscanthus sinensis* diversity panel (Clark et al., 2019b) from Study 3. The total number of shared tag sequences between two variant calling pipelines and the novel sorting pipeline, polyRAD, is highlighted in the figure.

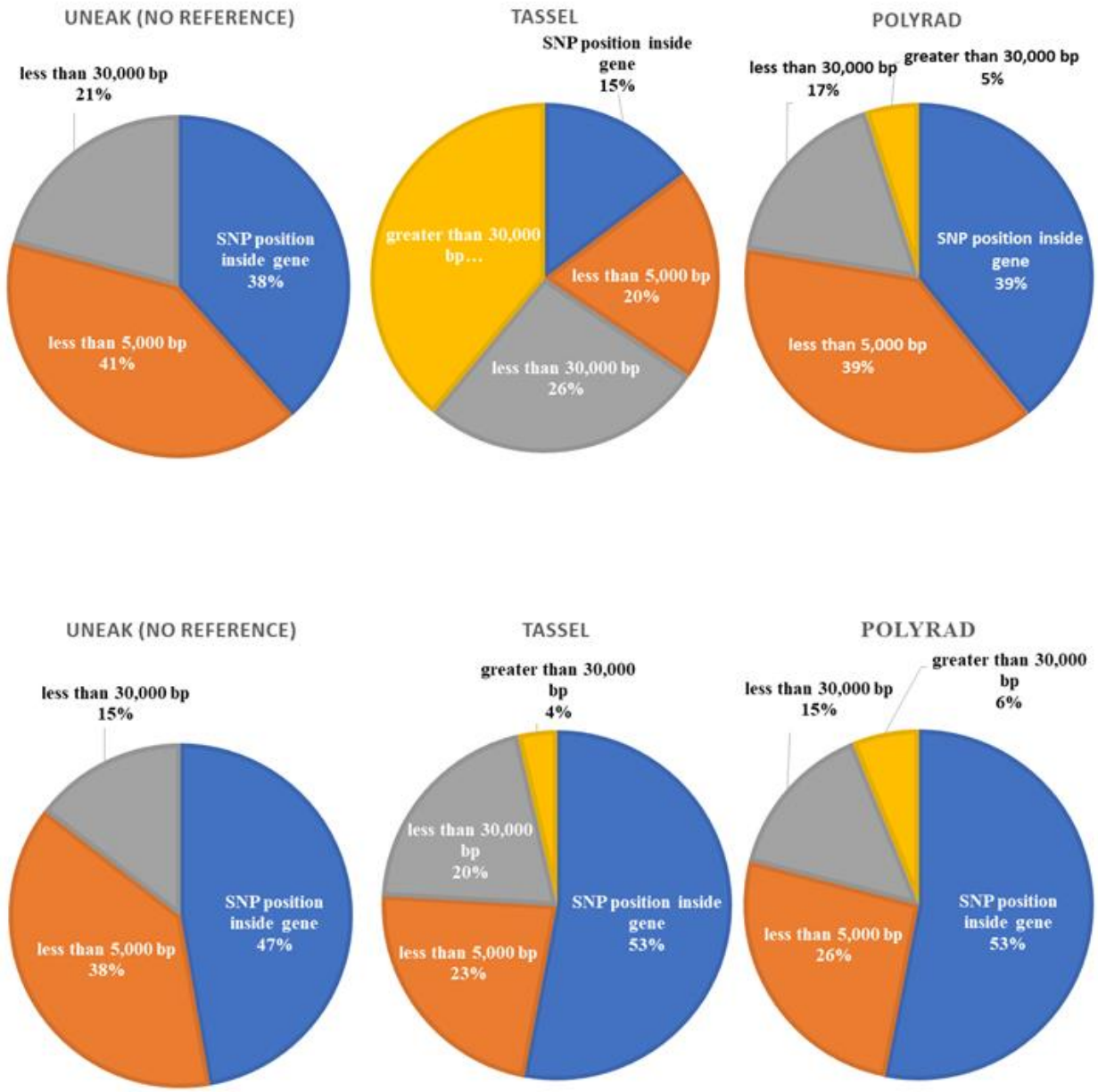



Figure 6. Distance in base pairs of all of single nucleotide polymorphisms (SNPs) to the nearest gene from three variant calling pipelines (UNEAK, TASSEL, polyRAD) from genome-wide association analysis of a *Miscanthus sinensis* diversity panel for biomass and 12 yield traits performed in Study 3 (Clark et al., 2019b). The following distances are indicated by four colors: blue) SNP position located inside a gene orange) SNP position located less than 5,000 bps to the nearest gene gray) SNP position located less than 30,000 bps to the nearest gene yellow) SNP position located greater than 30,000 bps to the nearest gene. (a) All SNPs output by each pipeline. (b) Significant SNPs only.

## 2.8 Literature Cited


Arruda, M. P., Brown, P., Brown-Guedira, G., Krill, A. M., Thurber, C., Merrill, K. R., … Kolb, F. L. (2016). Genome-Wide Association Mapping of Fusarium Head Blight Resistance in Wheat using Genotyping-by-Sequencing. *The Plant Genome*, *9*(1). https://doi.org/10.3835/plantgenome2015.04.0028

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. JSTOR.

Bhatta, M., Shamanin, V., Shepelev, S., Baenziger, P. S., Pozherukova, V., Pototskaya, I., & Morgounov, A. (2019). Marker-Trait Associations for Enhancing Agronomic Performance, Disease Resistance, and Grain Quality in Synthetic and Bread Wheat Accessions in Western Siberia. *G3: Genes, Genomes, Genetics*, *9*(12), 4209–4222. https://doi.org/10.1534/g3.119.400811

Bernardo, A. N., Ma, H., Zhang, D., & Bai, G. (2012). Single nucleotide polymorphism in wheat chromosome region harboring Fhb1 for Fusarium head blight resistance. *Molecular Breeding*, *29*(2), 477–488. https://doi.org/10.1007/s11032-011-9565-y

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *The American Journal of Human Genetics*, *103*(3), 338–348. https://doi.org/10.1016/j.ajhg.2018.07.015

Buerstmayr, H., Ban, T., & Anderson, J. A. (2009). QTL mapping and marker-assisted selection for Fusarium head blight resistance in wheat: A review. *Plant Breeding*, *128*(1), 1–26. https://doi.org/10.1111/j.1439-0523.2008.01550.x

Buggs, R. J. A., Elliott, N. M., Zhang, L., Koh, J., Viccini, L. F., Soltis, D. E., & Soltis, P. S. (2010). Tissue-specific silencing of homoeologs in natural populations of the recent allopolyploid Tragopogon mirus. *The New Phytologist*, *186*(1), 175–183. https://doi.org/10.1111/j.1469-8137.2010.03205.x

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Chen, N., Hout, C. V. V., Gottipati, S., & Clark, A. G. (2014). Using Mendelian Inheritance To Improve High-Throughput SNP Discovery. *Genetics*, *198*(3), 847–857. https://doi.org/10.1534/genetics.114.169052

Clark, L. V., Mays, W., Lipka, A. E., & Sacks, E. J. (2020). A population-level statistic for assessing Mendelian behavior of genotyping-by-sequencing data from highly duplicated genomes. *BioRxiv*, 2020.01.11.902890. https://doi.org/10.1101/2020.01.11.902890

Clark, L. V., Lipka, A. E., & Sacks, E. J. (2019a). polyRAD: Genotype Calling with Uncertainty from Sequencing Data in Polyploids and Diploids. *G3 (Bethesda, Md.)*, *9*(3), 663–673. https://doi.org/10.1534/g3.118.200913

Clark, L. V., Dwiyanti, M. S., Anzoua, K. G., Brummer, J. E., Ghimire, B. K., Głowacka, K., Hall, M., Heo, K., Jin, X., Lipka, A. E., Peng, J., Yamada, T., Yoo, J. H., Yu, C. Y., Zhao, H., Long, S. P., & Sacks, E. J. (2019b). Genome-wide association and genomic prediction for biomass yield in a genetically diverse Miscanthus sinensis germplasm

panel phenotyped at five locations in Asia and North America. *GCB Bioenergy*, *11*(8), 988–1007. https://doi.org/10.1111/gcbb.12620

Clark, L. V., & Sacks, E. J. (2016). TagDigger: User-friendly extraction of read counts from GBS and RAD-seq data. *Source Code for Biology and Medicine*, *11*(1), 11. https://doi.org/10.1186/s13029-016-0057-7

Clark, L. V., & Schreier, A. D. (2017). Resolving microsatellite genotype ambiguity in populations of allopolyploid and diploidized autopolyploid organisms using negative correlations between allelic variables. *Molecular Ecology Resources*, *17*(5), 1090–1103. https://doi.org/10.1111/1755-0998.12639

Consortium (IWGSC), T. I. W. G. S., Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., … Wang, L. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*(6403), eaar7191. https://doi.org/10.1126/science.aar7191

Corwin, J. A., & Kliebenstein, D. J. (2017). Quantitative Resistance: More Than Just Perception of a Pathogen. *The Plant Cell*, *29*(4), 655–665. https://doi.org/10.1105/tpc.16.00915

Cuthbert, P. A., Somers, D. J., & Brulé-Babel, A. (2007). Mapping of Fhb2 on chromosome 6BS: A gene controlling Fusarium head blight field resistance in bread wheat (Triticum aestivum L.). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *114*(3), 429–437. https://doi.org/10.1007/s00122-006-0439-3

DaCosta, J. M., & Sorenson, M. D. (2014). Amplification Biases and Consistent Recovery of Loci in a Double-Digest RAD-seq Protocol. *PLOS ONE*, *9*(9), e106713. https://doi.org/10.1371/journal.pone.0106713

Davey, J. W., Cezard, T., Fuentes-Utrilla, P., Eland, C., Gharbi, K., & Blaxter, M. L. (2013).

Special features of RAD Sequencing data: Implications for genotyping. *Molecular Ecology*, *22*(11), 3151–3164. https://doi.org/10.1111/mec.12084

Dixon, J., Braun, H.-J., Kosina, P., & Crouch, J. (n.d.). *Wheat Facts and Futures 2009*. 105.

Dufresne, F., Stift, M., Vergilino, R., & Mable, B. K. (2014). Recent progress and challenges in population genetics of polyploid organisms: An overview of current state-of-the-art molecular and statistical tools. *Molecular Ecology*, *23*(1), 40–69. https://doi.org/10.1111/mec.12581

Eaton, D. A. R. (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, *30*(13), 1844–1849. https://doi.org/10.1093/bioinformatics/btu121

https://doi.org/10.3835/plantgenome2011.08.0024

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLOS ONE*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Fuentes, R. G., Mickelson, H. R., Busch, R. H., Dill-Macky, R., Evans, C. K., Thompson, W. G., Wiersma, J. V., Xie, W., Dong, Y., & Anderson, J. A. (2005). Resource Allocation and Cultivar Stability in Breeding for Fusarium Head Blight Resistance in Spring Wheat. *Crop Science*, *45*(5), 1965–1972. https://doi.org/10.2135/cropsci2004.0589

Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., … Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, *22*(11), 3165–3178. https://doi.org/10.1111/mec.12089

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., … Galtier, N.

(2013). Reference-Free Population Genomics from Next-Generation Transcriptome Data and the Vertebrate–Invertebrate Gap. *PLOS Genetics*, *9*(4), e1003457. https://doi.org/10.1371/journal.pgen.1003457

Gerard, D., Ferrão, L. F. V., Garcia, A. A. F., & Stephens, M. (2018). Genotyping Polyploids from Messy Sequencing Data. *Genetics*, *210*(3), 789–807. https://doi.org/10.1534/genetics.118.301468

Gilsinger, J., Kong, L., Shen, X., & Ohm, H. (2005). DNA markers associated with low Fusarium head blight incidence and narrow flower opening in wheat. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *110*(7), 1218–1225. https://doi.org/10.1007/s00122-005-1953-4

Green, A. J., Berger, G., Griffey, C. A., Pitman, R., Thomason, W., Balota, M., & Ahmed, A. (2012). Genetic Yield Improvement in Soft Red Winter Wheat in the Eastern United States from 1919 to 2009. *Crop Science*, *52*(5), 2097–2108. https://doi.org/10.2135/cropsci2012.01.0026

Jia, H., Zhou, J., Xue, S., Li, G., Yan, H., Ran, C., … Ma, Z. (2018). A journey to understand wheat Fusarium head blight resistance in the Chinese wheat landrace Wangshuibai. *The Crop Journal*, *6*(1), 48–59. https://doi.org/10.1016/j.cj.2017.09.006

Jordan, K. W., Wang, S., He, F., Chao, S., Lun, Y., Paux, E., Sourdille, P., Sherman, J., Akhunova, A., Blake, N. K., Pumphrey, M. O., Glover, K., Dubcovsky, J., Talbert, L., & Akhunov, E. D. (2018). The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. *The Plant Journal*, *95*(6), 1039–1054. https://doi.org/10.1111/tpj.14009

Katju, V., & Bergthorsson, U. (2013). Copy-number changes in evolution: Rates, fitness effects

and adaptive significance. *Frontiers in Genetics*, *4*. https://doi.org/10.3389/fgene.2013.00273

Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, *12*(5), e1004842. https://doi.org/10.1371/journal.pcbi.1004842

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., & Strömvik, M. V. (2018). Current Strategies of Polyploid Plant Genome Sequence Assembly. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01660

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Lexer, C., Wüest, R. O., Mangili, S., Heuertz, M., Stölting, K. N., Pearman, P. B., Forest, F., Salamin, N., Zimmermann, N. E., & Bossolini, E. (2014). Genomics of the divergence continuum in an African plant biodiversity hotspot, I: Drivers of population divergence in Restio capensis (Restionaceae). *Molecular Ecology*, *23*(17), 4373–4386. https://doi.org/10.1111/mec.12870

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Z., Baniaga, A. E., Sessa, E. B., Scascitelli, M., Graham, S. W., Rieseberg, L. H., & Barker, M. S. (2015). Early genome duplications in conifers and other seed plants. *Science Advances*, *1*(10), e1501084. https://doi.org/10.1126/sciadv.1501084

Liu, Y., Salsman, E., Fiedler, J. D., Hegstad, J. B., Green, A., Mergoum, M., … Li, X. (2019).

Genetic Mapping and Prediction Analysis of FHB Resistance in a Hard Red Spring Wheat Breeding Population. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01007

Liu, S., Pumphrey, M. O., Gill, B. S., Trick, H. N., Zhang, J. X., Dolezel, J., Chalhoub, B., & Anderson, J. A. (2008). Toward positional cloning of Fhb1, a major QTL for Fusarium head blight resistance in wheat. *Cereal Research Communications*, *36*(6), 195–201. https://doi.org/10.1556/CRC.36.2008.Suppl.B.15

Long, A. D., & Langley, C. H. (1999). The Power of Association Studies to Detect the Contribution of Candidate Genetic Loci to Variation in Complex Traits. *Genome Research*, *9*(8), 720–731.

Lu, F., Lipka, A. E., Glaubitz, J., Elshire, R., Cherney, J. H., Casler, M. D., Buckler, E. S., & Costich, D. E. (2013). Switchgrass Genomic Diversity, Ploidy, and Evolution: Novel Insights from a Network-Based SNP Discovery Protocol. *PLoS Genetics*, *9*(1). https://doi.org/10.1371/journal.pgen.1003215

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science (New York, N.Y.)*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Mastretta-Yanes, A., Zamudio, S., Jorgensen, T. H., Arrigo, N., Alvarez, N., Piñero, D., & Emerson, B. C. (2014). Gene Duplication, Population Genomics, and Species-Level Differentiation within a Tropical Mountain Shrub. *Genome Biology and Evolution*, *6*(10), 2611–2624. https://doi.org/10.1093/gbe/evu205

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. https://doi.org/10.1111/1755-0998.12613

Moore, G., Devos, K. M., Wang, Z., & Gale, M. D. (1995). Cereal Genome Evolution: Grasses, line up and form a circle. *Current Biology*, *5*(7), 737–739. https://doi.org/10.1016/S0960-9822(95)00148-5

Nadukkalam Ravindran, P., Bentzen, P., Bradbury, I. R., & Beiko, R. G. (2018). PMERGE: Computational filtering of paralogous sequences from RAD-seq data. *Ecology and Evolution*, *8*(14), 7002–7013. https://doi.org/10.1002/ece3.4219

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*(6), 443–451. https://doi.org/10.1038/nrg2986

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I. V., & Dubchak, I. (2014). The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research*, *42*(Database issue), D26-31. https://doi.org/10.1093/nar/gkt1069

Nguyen, T. K., Ha, S. T. T., & Lim, J. H. (2020). Analysis of chrysanthemum genetic diversity by genotyping-by-sequencing. *Horticulture, Environment, and Biotechnology*. https://doi.org/10.1007/s13580-020-00274-2

Ohno, S. (1970). Polyploidy: Duplication of the Entire Genome. In S. Ohno (Ed.), *Evolution by Gene Duplication* (pp. 98–106). Springer. https://doi.org/10.1007/978-3-642-86659-3_17

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B.,

Speicher, M. R., Zschocke, J., & Trajanoski, Z. (2014). A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics*, *15*(2), 256–278. https://doi.org/10.1093/bib/bbs086

Parry, D. W., Jenkinson, P., & McLEOD, L. (1995). Fusarium ear blight (scab) in small grain cereals—A review. *Plant Pathology*, *44*(2), 207–238. https://doi.org/10.1111/j.1365-3059.1995.tb02773.x

Perea, C., De La Hoz, J. F., Cruz, D. F., Lobaton, J. D., Izquierdo, P., Quintero, J. C., Raatz, B., & Duitama, J. (2016). Bioinformatic analysis of genotype by sequencing (GBS) data with NGSEP. *BMC Genomics*, *17*(Suppl 5). https://doi.org/10.1186/s12864-016-2827-7

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS ONE*, *7*(5), e37135. https://doi.org/10.1371/journal.pone.0037135

Poland, J., Endelman, J., Dawson, J., Rutkoski, J., Wu, S., Manes, Y., Dreisigacker, S., Crossa, J., Sánchez-Villeda, H., Sorrells, M., & Jannink, J.-L. (2012). Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *The Plant Genome*, *5*(3), 103–113. https://doi.org/10.3835/plantgenome2012.06.0006

Qu, C., Jia, L., Fu, F., Zhao, H., Lu, K., Wei, L., Xu, X., Liang, Y., Li, S., Wang, R., & Li, J. (2017). Genome-wide association mapping and Identification of candidate genes for fatty acid composition in Brassica napus L. using SNP markers. *BMC Genomics*, *18*(1), 232. https://doi.org/10.1186/s12864-017-3607-8

Ravindran, P. N., Bentzen, P., Bradbury, I. R., & Beiko, R. G. (2018). PMERGE: Computational

filtering of paralogous sequences from RAD-seq data. *Ecology and Evolution*, *8*(14), 7002–7013. https://doi.org/10.1002/ece3.4219

Rivera-Colón, A. G., Rochette, N. C., & Catchen, J. M. (2019). *Simulation with RADinitio Improves RADseq Experimental Design and Sheds Light on Sources of Missing Data* [Preprint]. https://doi.org/10.1101/775239

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2018). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, 907–917. https://doi.org/10.1111/2041-210X.12700@10.1111/(ISSN)2041-210x.PracticalToolsFieldMethodsMEE32018

Shen, X., Ittu, M., & Ohm, H. W. (2003). Quantitative Trait Loci Conditioning Resistance to Fusarium Head Blight in Wheat Line F201R. *Crop Science*, *43*(3), 850–857. https://doi.org/10.2135/cropsci2003.8500

Slavov, G. T., Nipper, R., Robson, P., Farrar, K., Allison, G. G., Bosch, M., Clifton-Brown, J. C., Donnison, I. S., & Jensen, E. (2014). Genome-wide association studies and prediction of 17 traits related to phenology, biomass and cell wall composition in the energy grass Miscanthus sinensis. *New Phytologist*, *201*(4), 1227–1239. https://doi.org/10.1111/nph.12621

Stebbins, G. L. (1940). The Significance of Polyploidy in Plant Evolution. *The American Naturalist*, *74*(750), 54–66. JSTOR.

Swaminathan, K., Chae, W. B., Mitros, T., Varala, K., Xie, L., Barling, A., Glowacka, K., Hall, M., Jezowski, S., Ming, R., Hudson, M., Juvik, J. A., Rokhsar, D. S., & Moose, S. P. (2012). A framework genetic map for Miscanthus sinensis from RNAseq-based markers

shows recent tetraploidy. *BMC Genomics*, *13*, 142. https://doi.org/10.1186/1471-2164-13-142

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and Collinearity in Plant Genomes. *Science*, *320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., Buckler, E. S., & Zhang, Z. (2016). GAPIT Version 2: An Enhanced Integrated Tool for Genomic Association and Prediction. *The Plant Genome*, *9*(2). https://doi.org/10.3835/plantgenome2015.11.0120

Tong, Z., Zhou, J., Xiu, Z., Jiao, F., Hu, Y., Zheng, F., Chen, X., Li, Y., Fang, D., Li, S., Wu, X., Zeng, J., Zhao, S., Jian, J., & Xiao, B. (2020). Construction of a high-density genetic map with whole genome sequencing in Nicotiana tabacum L. *Genomics*, *112*(2), 2028–2033. https://doi.org/10.1016/j.ygeno.2019.11.015

Tinker, N. A., Bekele, W. A., & Hattori, J. (2016). Haplotag: Software for Haplotype-Based Genotyping-by-Sequencing Analysis. *G3: Genes, Genomes, Genetics*, *6*(4), 857–863. https://doi.org/10.1534/g3.115.024596

Truong, H. T., Ramos, A. M., Yalcin, F., Ruiter, M. de, Poel, H. J. A. van der, Huvenaars, K. H. J., Hogers, R. C. J., Enckevort, L. J. G. van, Janssen, A., Orsouw, N. J. van, & Eijk, M. J. T. van. (2012). Sequence-Based Genotyping for Marker Discovery and Co-Dominant Scoring in Germplasm and Populations. *PLOS ONE*, *7*(5), e37565. https://doi.org/10.1371/journal.pone.0037565

Verdu, C. F., Guichoux, E., Quevauvillers, S., De Thier, O., Laizet, Y., Delcamp, A., …

Mariette, S. (2016). Dealing with paralogy in RADseq data: In silico detection and single nucleotide polymorphism validation in Robinia pseudoacacia L. *Ecology and Evolution*, *6*(20), 7323–7333. https://doi.org/10.1002/ece3.2466

Waples, R. K., Seeb, L. W., & Seeb, J. E. (2016). Linkage mapping with paralogs exposes regions of residual tetrasomic inheritance in chum salmon (Oncorhynchus keta). *Molecular Ecology Resources*, *16*(1), 17–28. https://doi.org/10.1111/1755-0998.12394

Waples, R. K., Seeb, J. E., & Seeb, L. W. (2017). Congruent population structure across paralogous and nonparalogous loci in Salish Sea chum salmon (Oncorhynchus keta). *Molecular Ecology*, *26*(16), 4131–4144. https://doi.org/10.1111/mec.14163

Wang J.,Zhang Z., 2018 GAPIT Version 3:An Interactive Analytical Tool for Genomic Association and Prediction.

Wang, K., Dickson, S. P., Stolle, C. A., Krantz, I. D., Goldstein, D. B., & Hakonarson, H. (2010). Interpretation of Association Signals and Identification of Causal Variants from Genome-wide Association Studies. *American Journal of Human Genetics*, *86*(5), 730–742. https://doi.org/10.1016/j.ajhg.2010.04.003

Xue, S., Li, G., Jia, H., Xu, F., Lin, F., Tang, M., … Ma, Z. (2010). Fine mapping Fhb4, a major QTL conditioning resistance to Fusarium infection in bread wheat (Triticum aestivum L.). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *121*(1), 147–156. https://doi.org/10.1007/s00122-010-1298-5

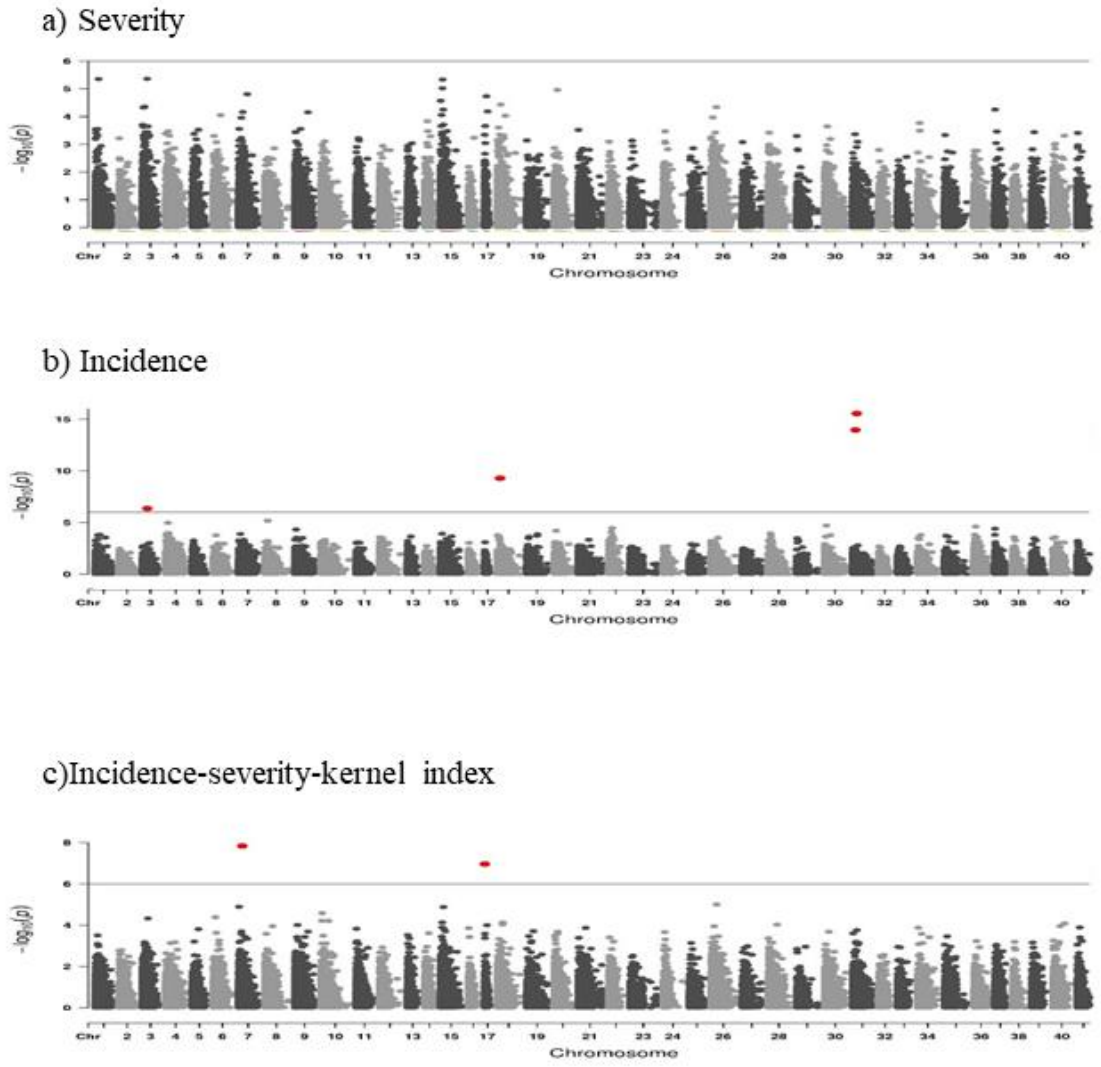
Xue, S., Xu, F., Tang, M., Zhou, Y., Li, G., An, X., … Ma, Z. (2011). Precise mapping Fhb5, a major QTL conditioning resistance to Fusarium infection in bread wheat (Triticum aestivum L.). *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *123*(6), 1055–1063. https://doi.org/10.1007/s00122-011-1647-z

Yu, X., & Sun, S. (2013). Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, *14*(1), 274. https://doi.org/10.1186/1471-2105-14-274

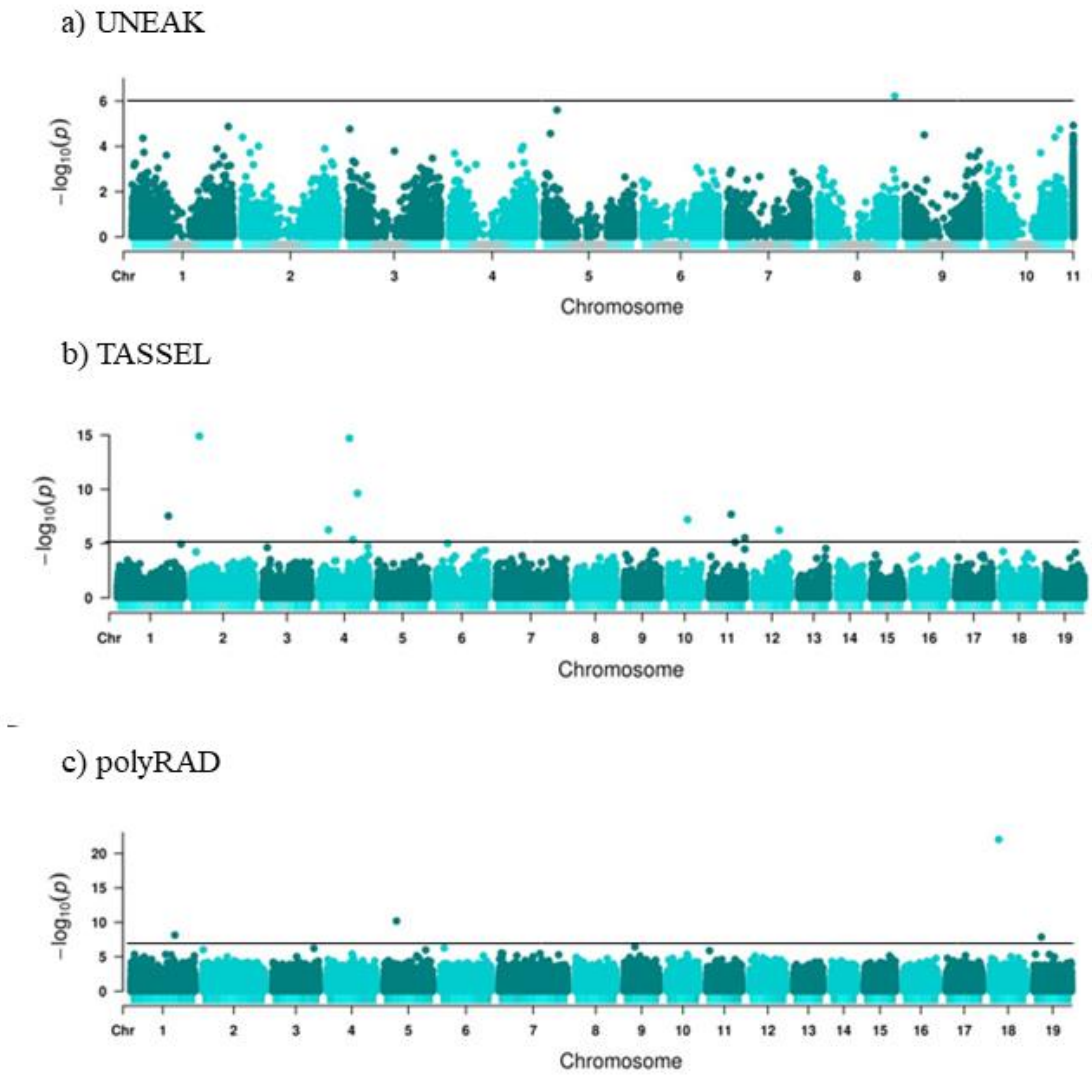
Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355–360. https://doi.org/10.1038/ng.546

Zhou, W., Kolb, F. L., Bai, G., Shaner, G., & Domier, L. L. (2002). Genetic analysis of scab resistance QTL in wheat with microsatellite and AFLP markers. *Genome*, *45*(4), 719–727. https://doi.org/10.1139/g02-034

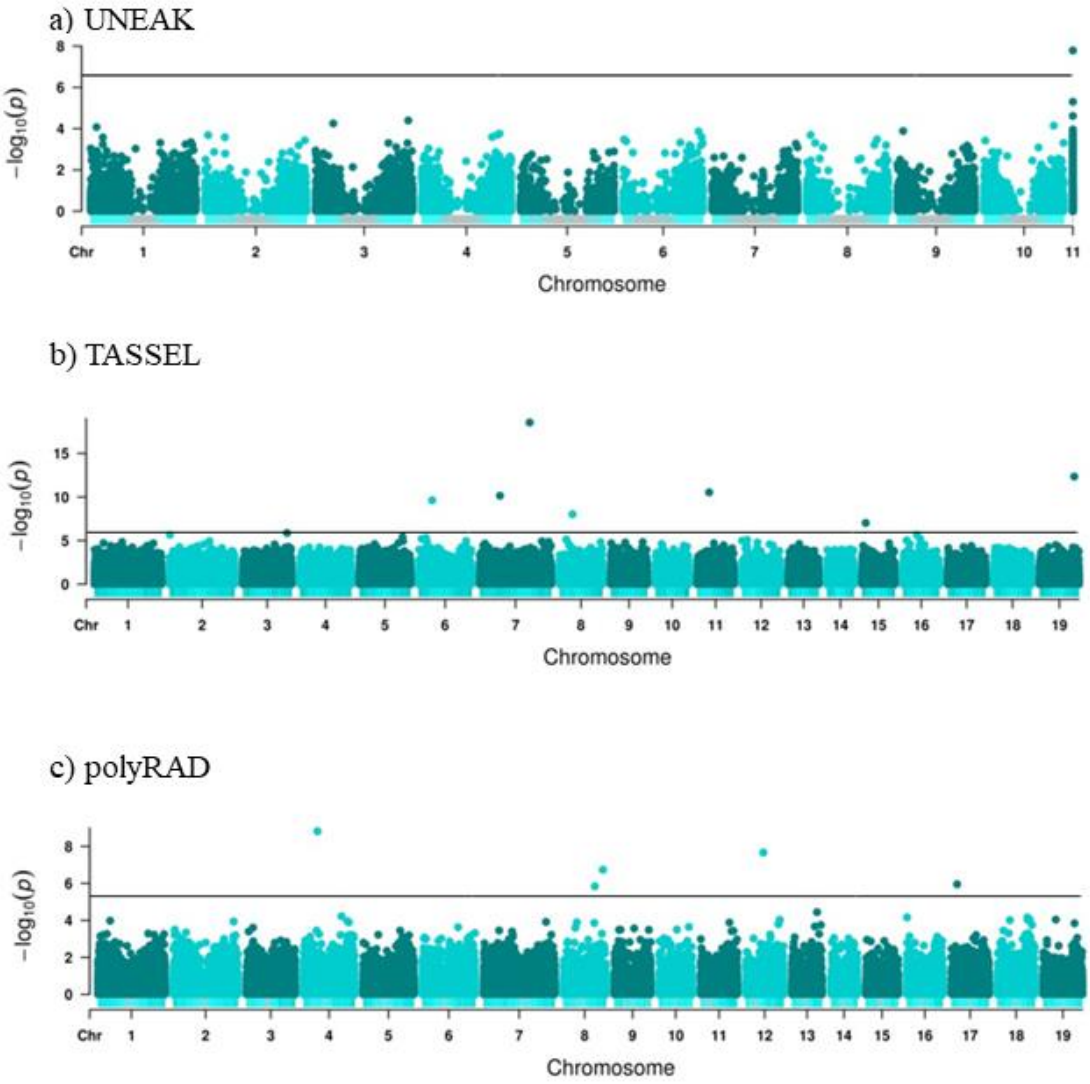
## 2.9 Supplementary Figures



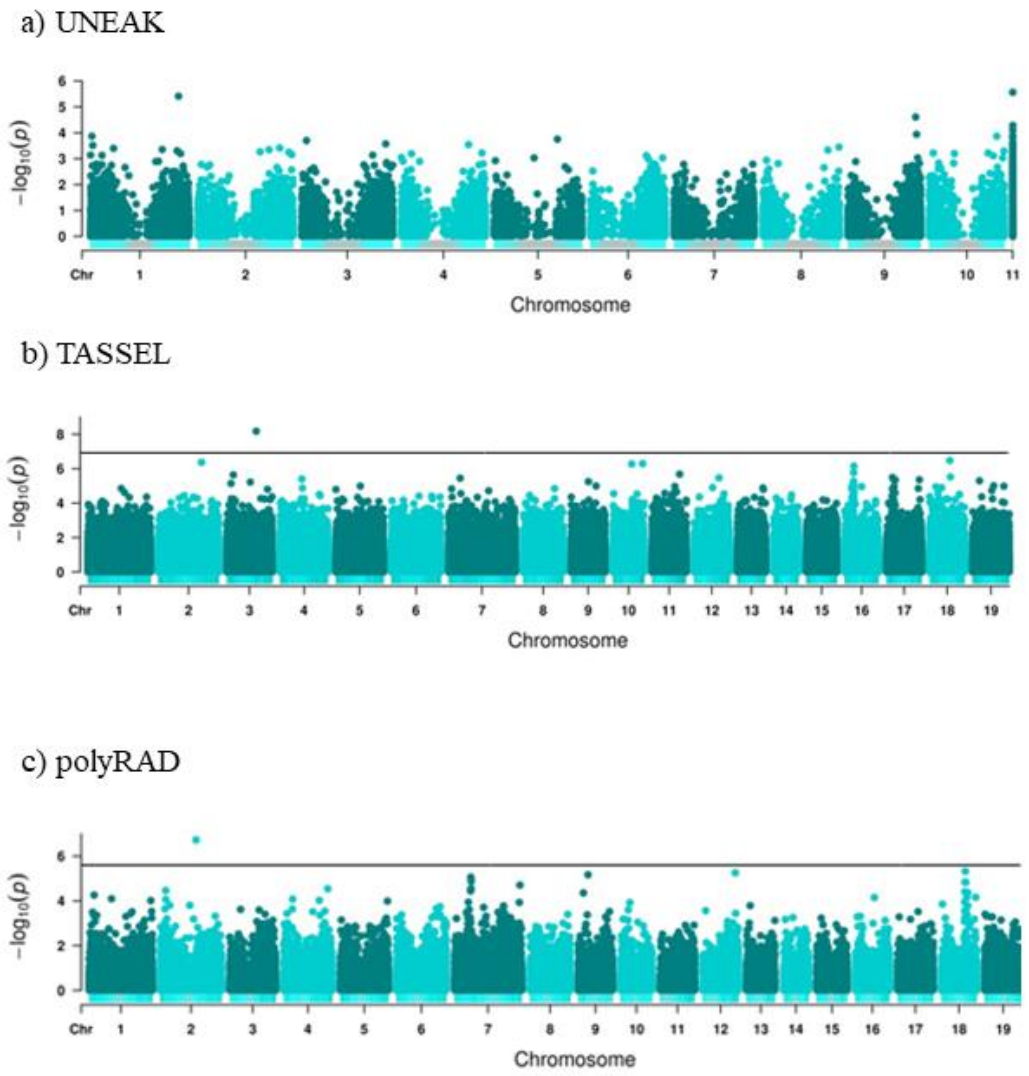

Supplementary Figure 1. A genome wide association study assessing three traits associated with fusarium head blight within 273 *Triticum aestivum* breeding lines collected from the Midwest and Eastern United States was conducted to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. A total of 87,385 SNP markers were included in the genome-wide association analysis and positions of SNPs were aligned with respect to the *Triticum aestivum* v. 2.2 reference genome. The analysis has resulted in the identification of a) 0 significant SNPs associated with severity, b) 4 significant SNPs associated with incidence, and c) 2 significant SNPs associated with incidence-severity kernel index.



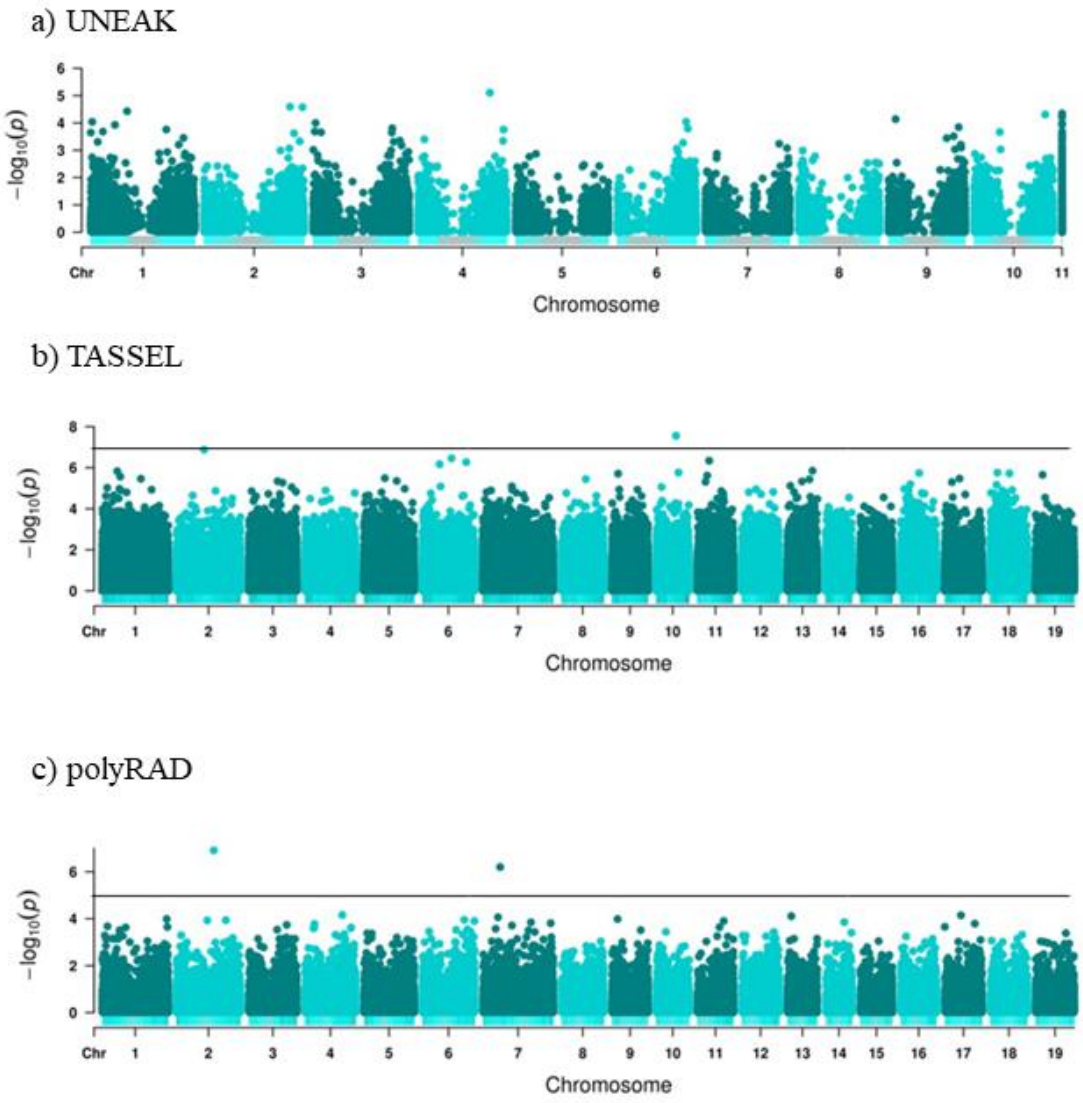

Supplementary Figure 2. A genome wide association study assessing the trait basal circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 4 significant SNPs associated with TASSEL pipeline, and c) 10 significant SNPs associated with polyRAD variant calling pipeline.
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).



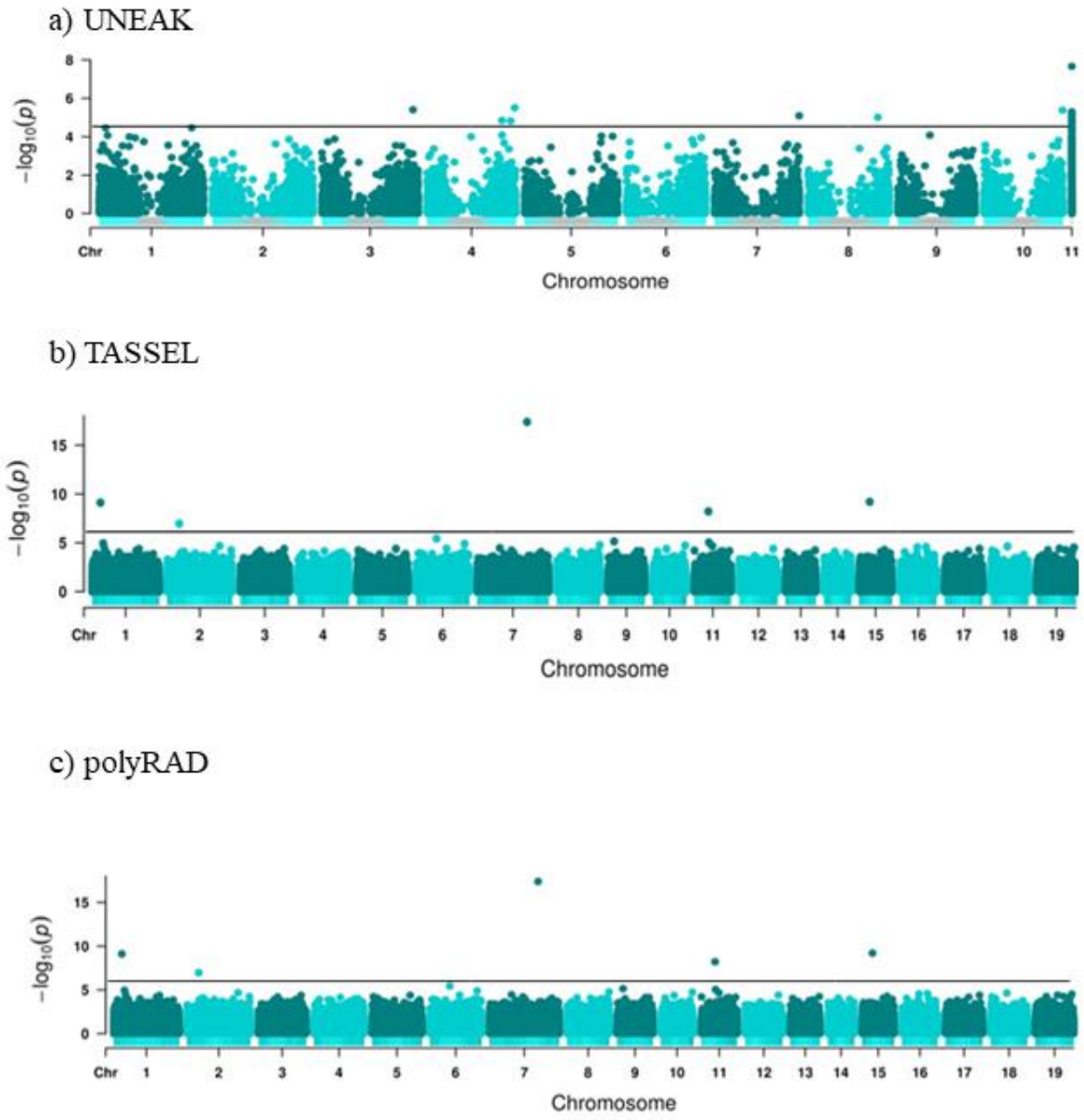

Supplementary Figure 3. A genome wide association study assessing the trait compressed circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 7 significant SNPs associated with TASSEL pipeline, and c) 5 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Zhuji, China by Zhejiang University (ZJU).



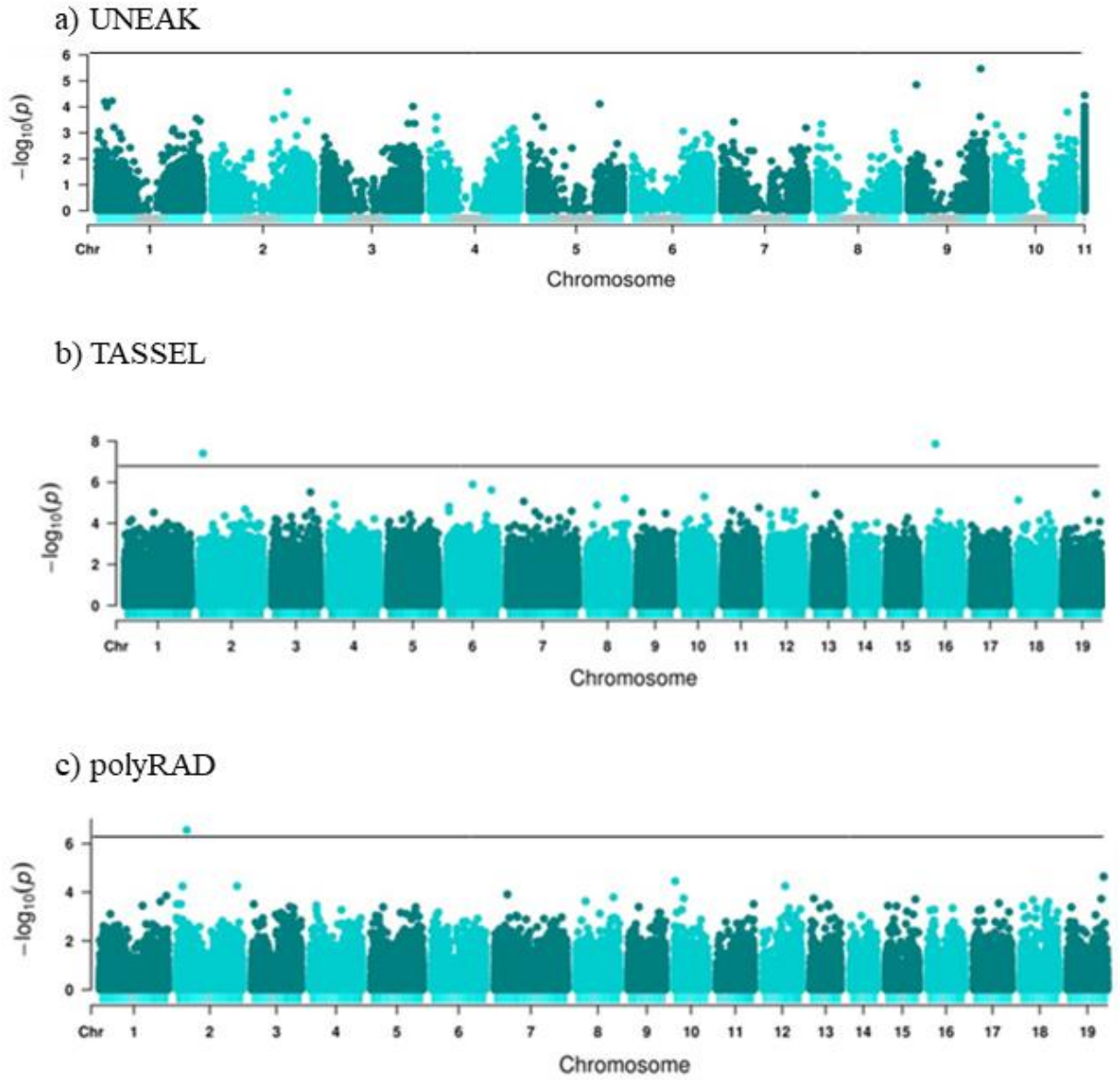

Supplementary Figure 4. A genome wide association study assessing the trait compressed circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).



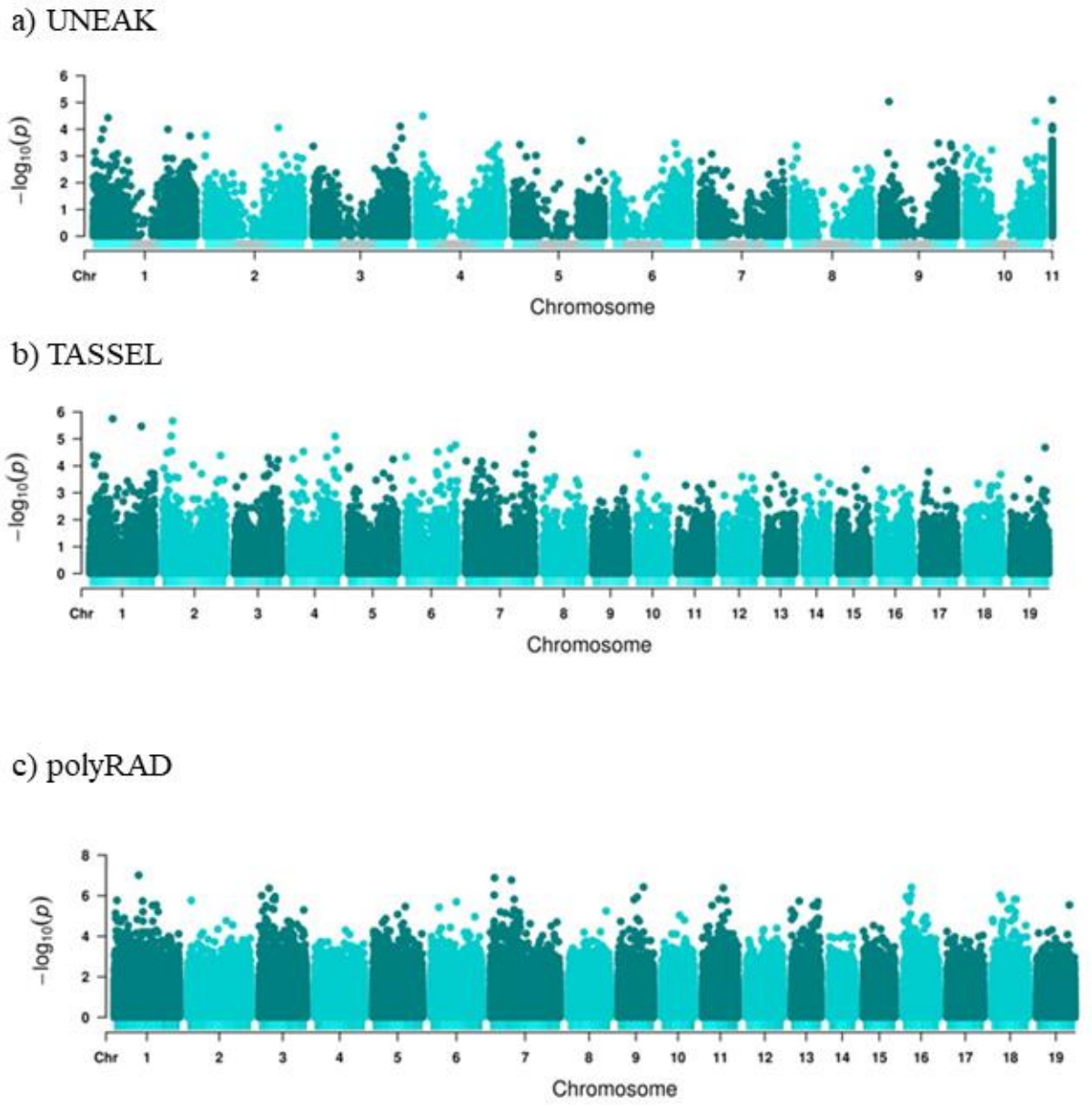

Supplementary Figure 5. A genome wide association study assessing the trait compressed circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 2 significant SNPs associated with polyRAD variant calling pipeline.
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).




Supplementary Figure 6. A genome wide association study assessing the trait compressed circumference divided by the basal circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 15 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 5 significant SNPs associated with TASSEL pipeline, and c) 3 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Zhuji, China by Zhejiang University (ZJU).



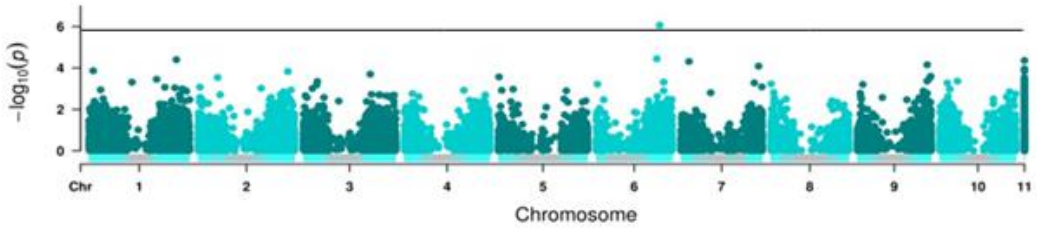

Supplementary Figure 7. A genome wide association study assessing the trait compressed circumference divided by compressed circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 2 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).



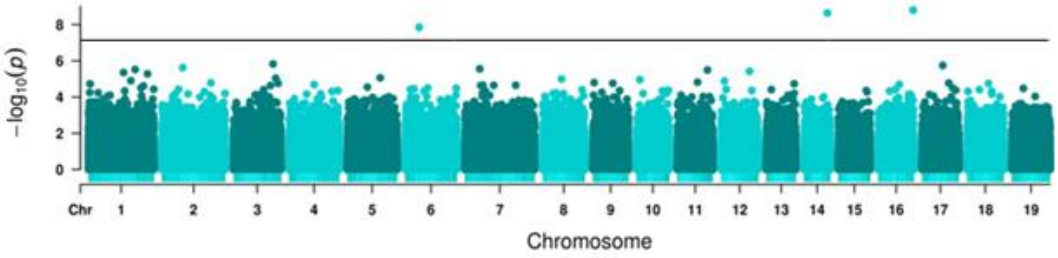

Supplementary Figure 8. A genome wide association study assessing the trait compressed circumference divided by basal circumference, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 0 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.

Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).

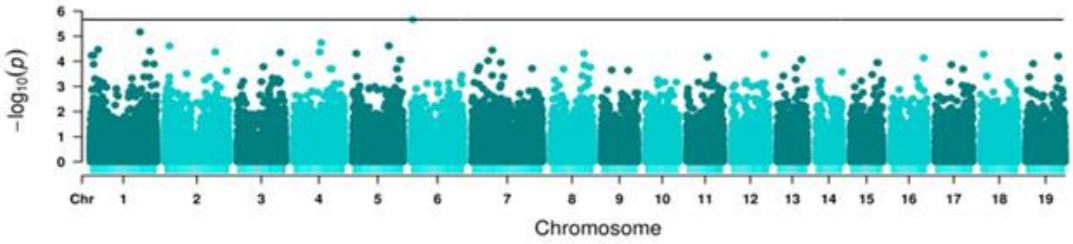
a) UNEAK




b) TASSEL

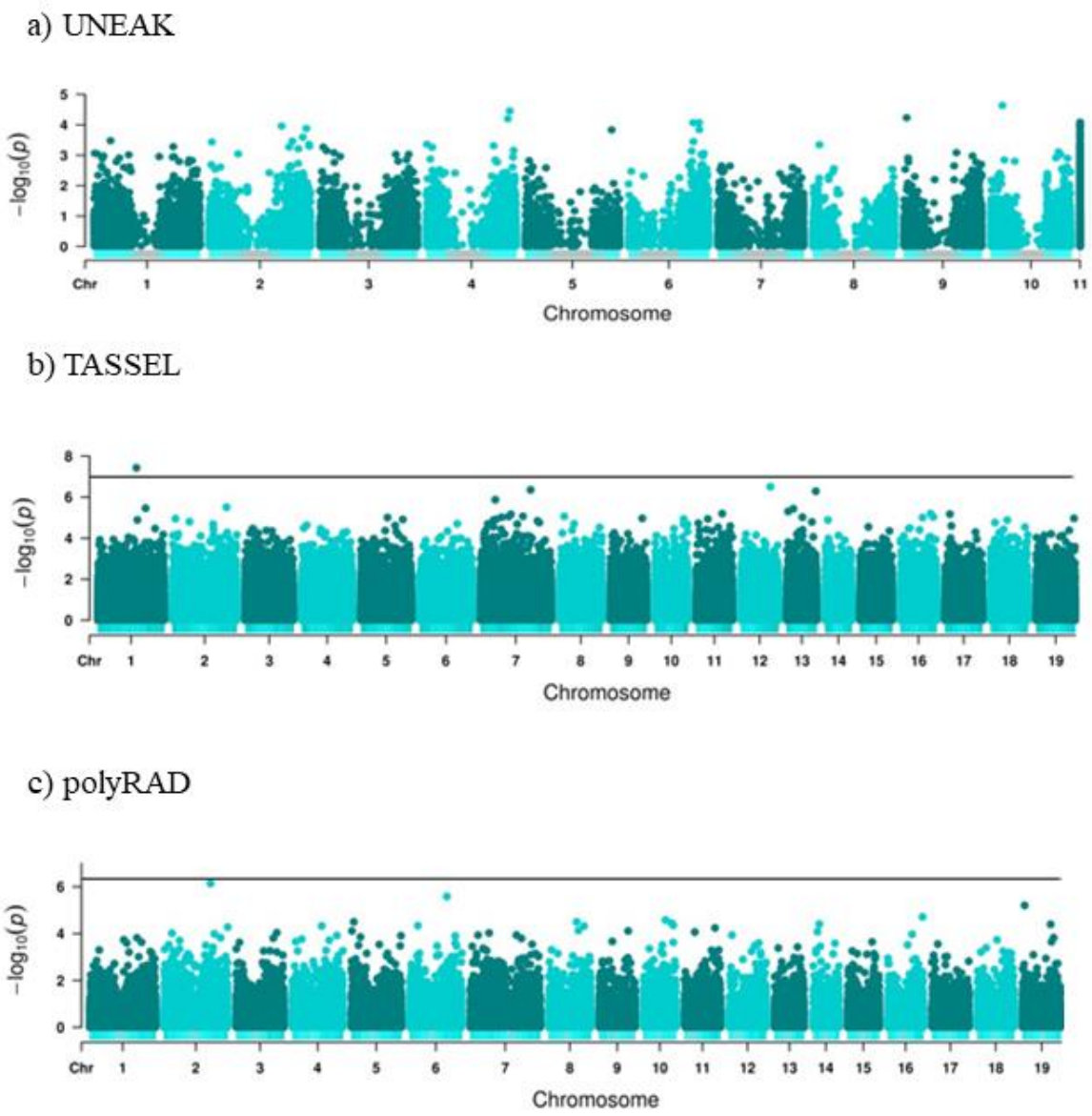



c) polyRAD

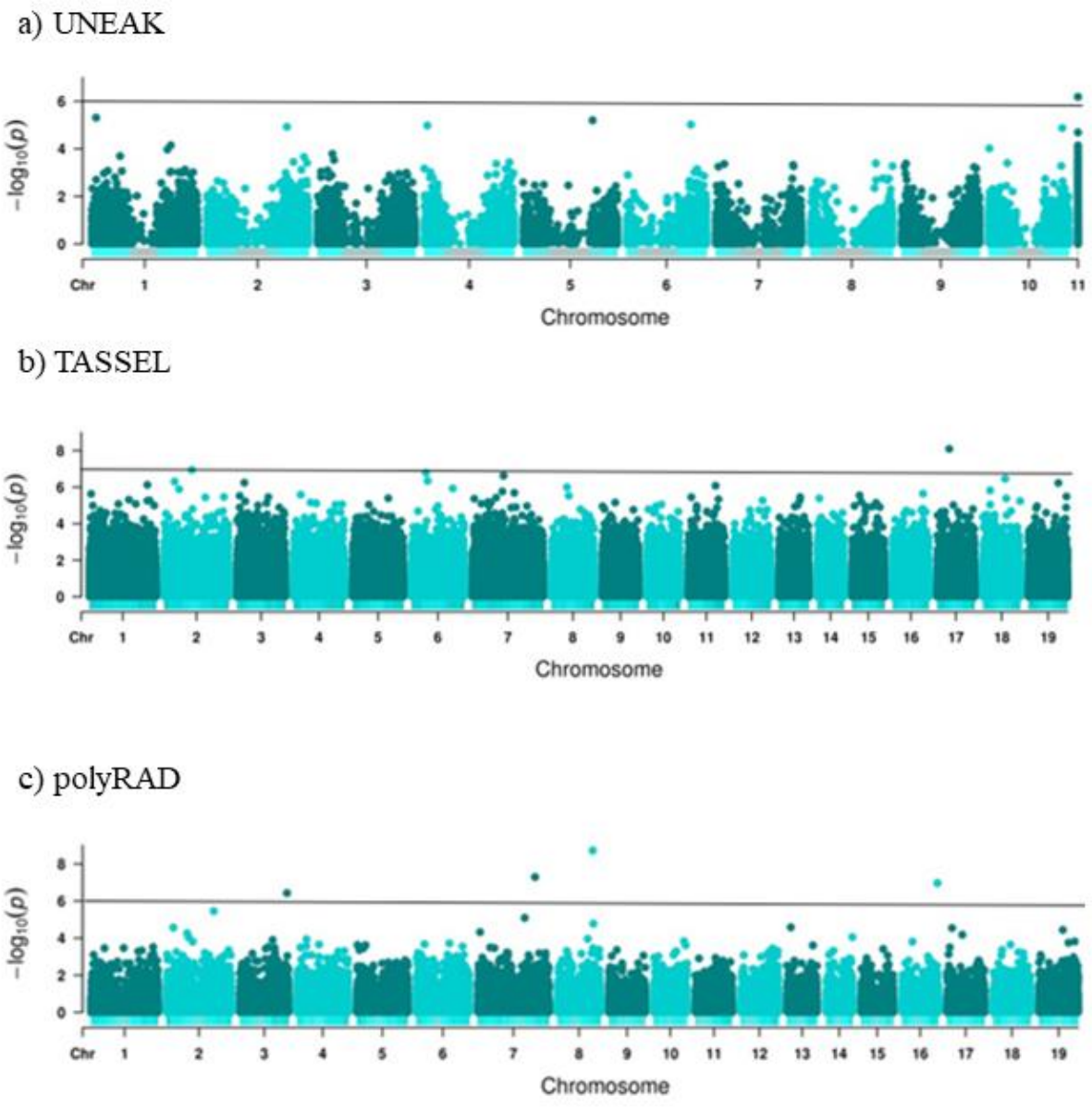



Supplementary Figure 9. A genome wide association study assessing the trait compressed circumference divided by the culm length, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 3 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.

Location(s): Zhuji, China by Zhejiang University (ZJU).

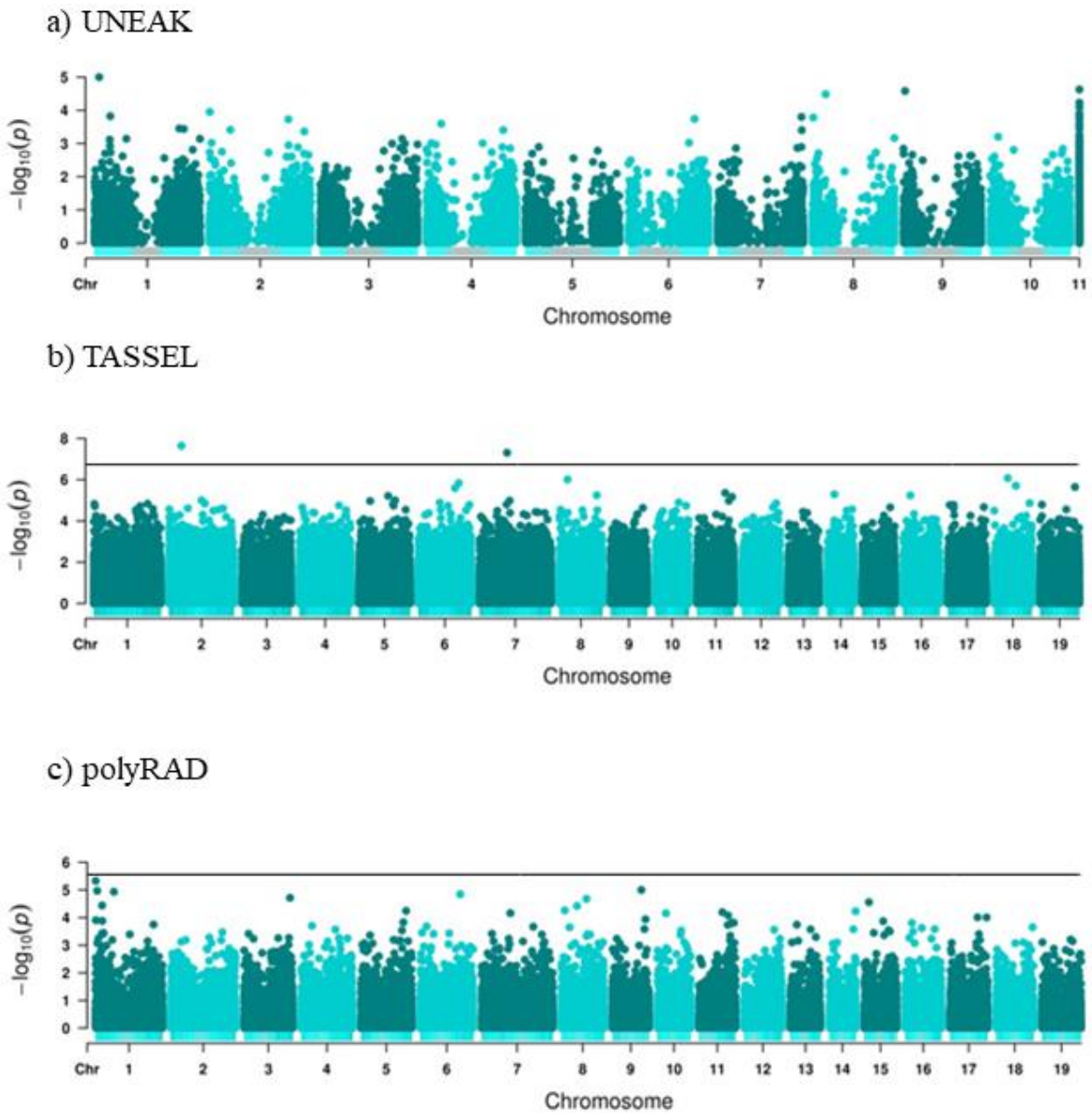



Supplementary Figure 10. A genome wide association study assessing the trait culm length, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).

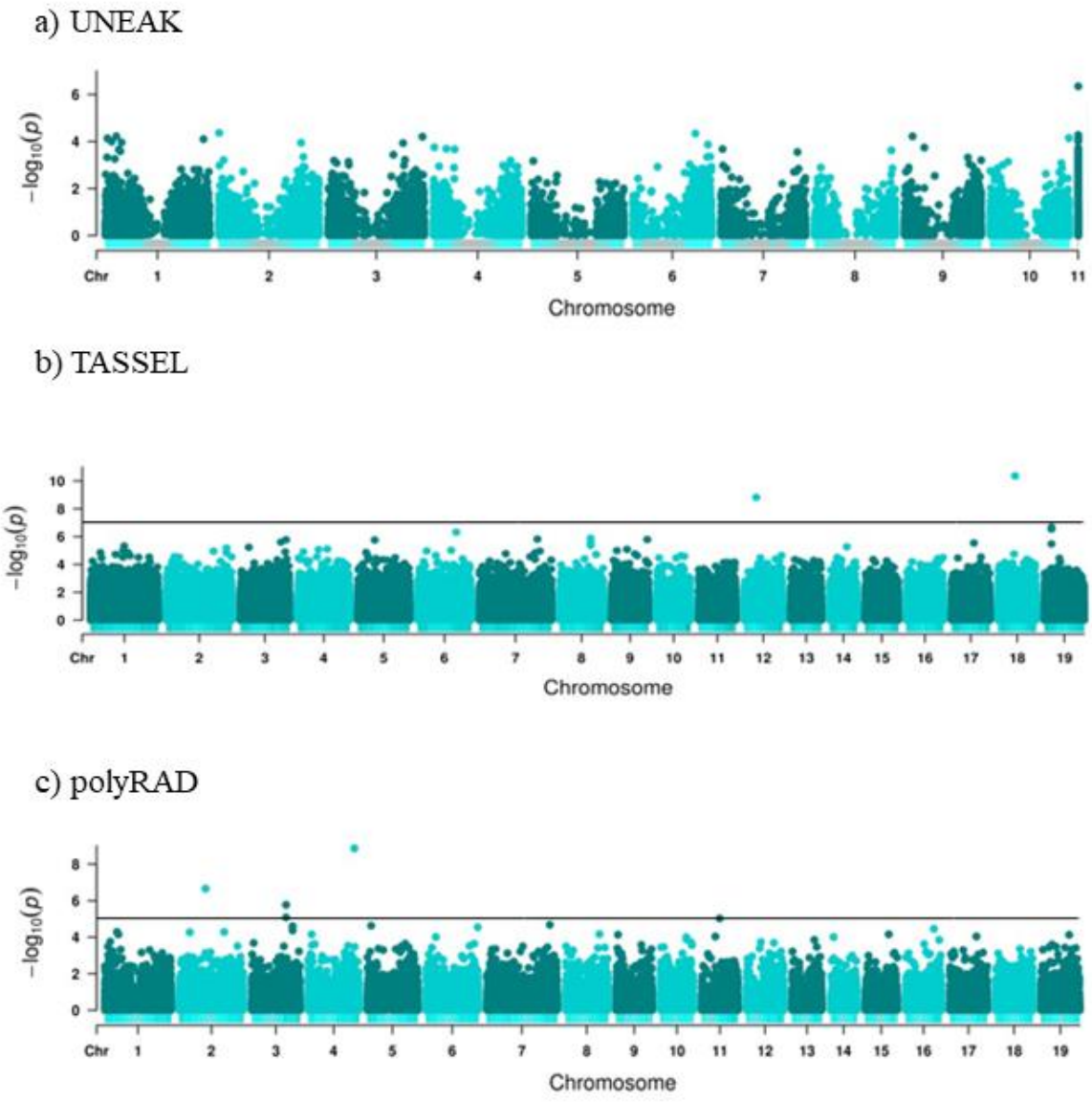



Supplementary Figure 11. A genome wide association study assessing the trait culm length, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 4 significant SNPs associated with polyRAD variant calling pipeline.

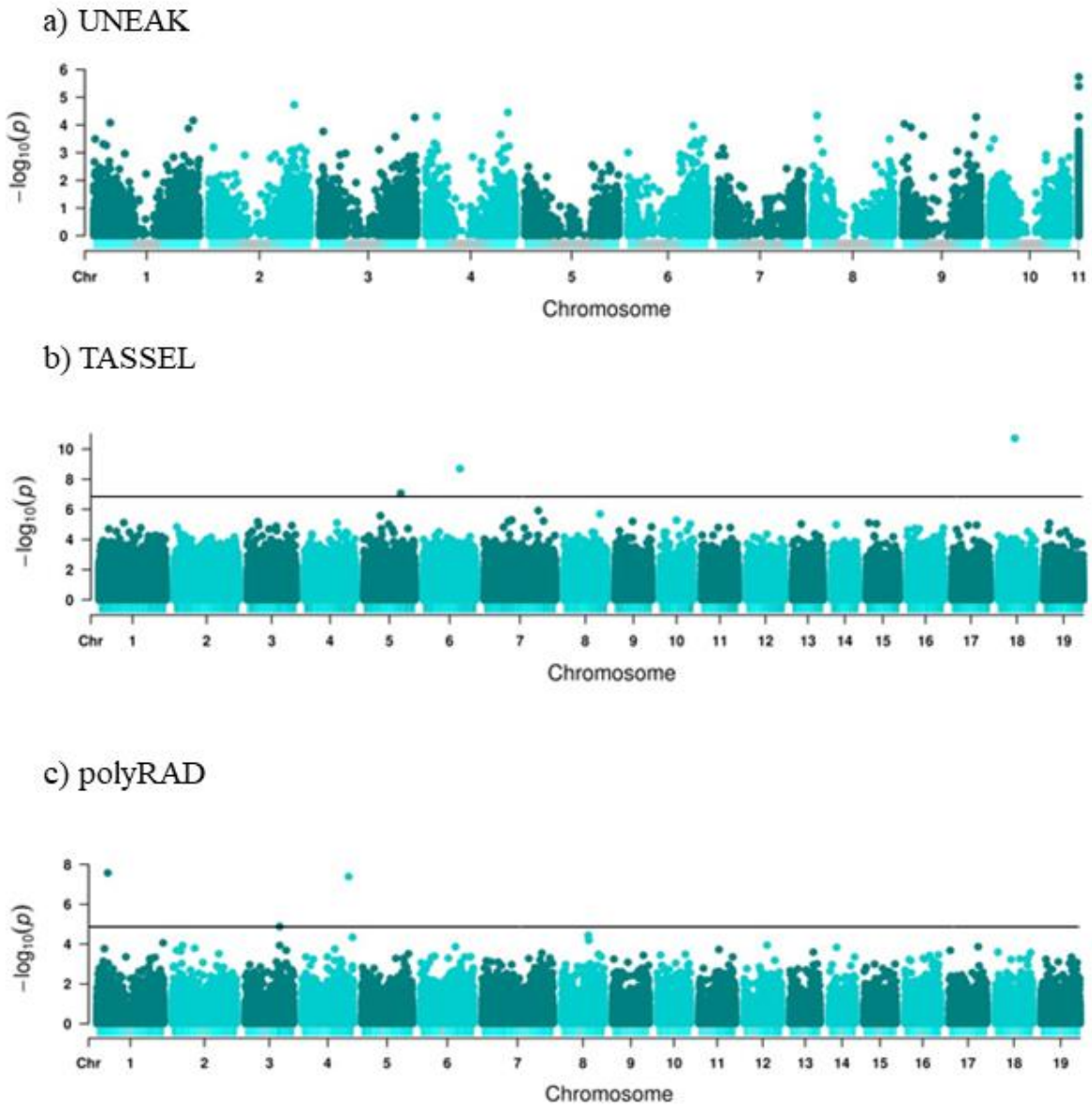
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).



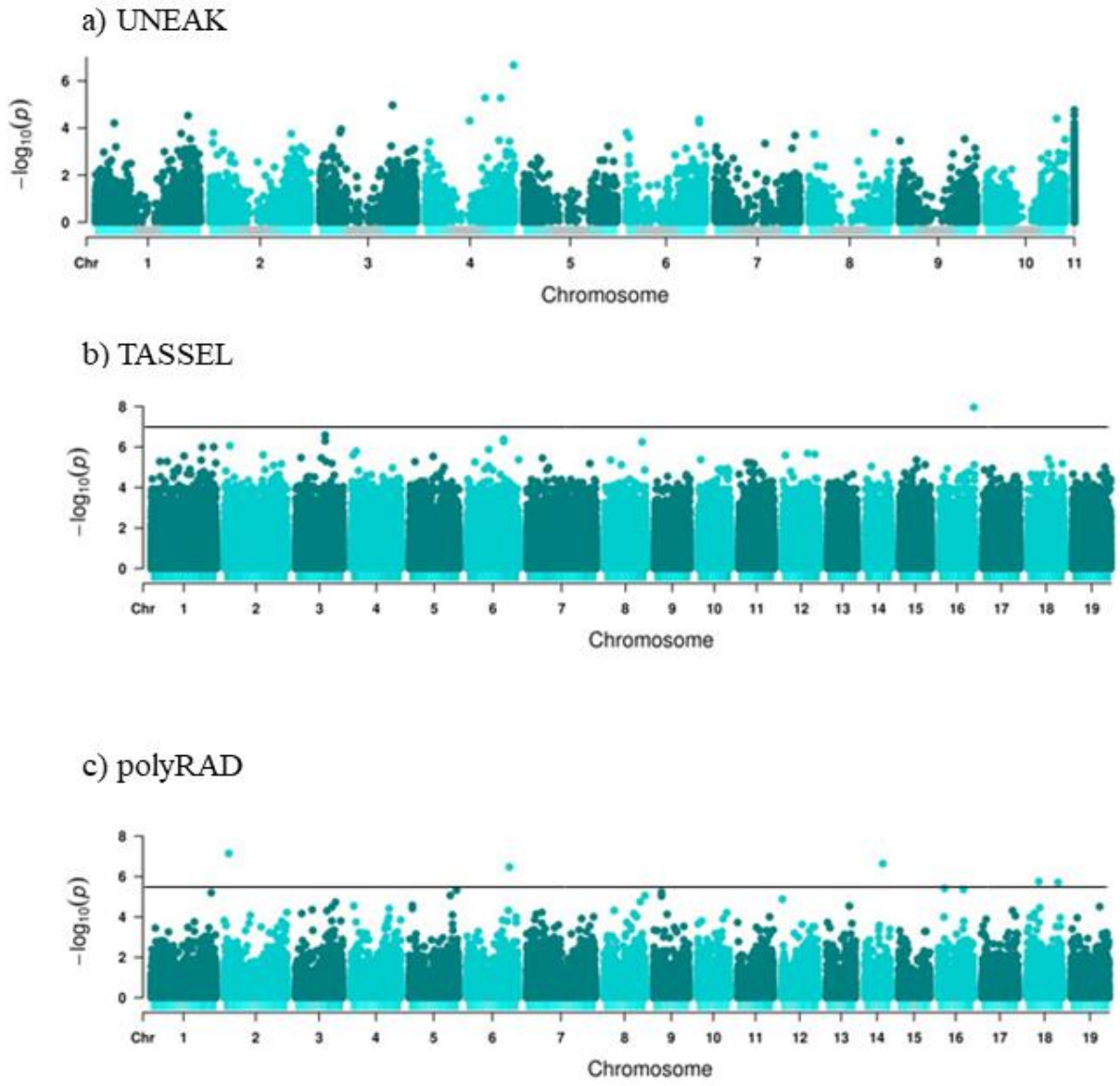

Supplementary Figure 12. A genome wide association study assessing the trait compressed circumference divided by the culm node number, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 2 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Zhuji, China by Zhejiang University (ZJU).



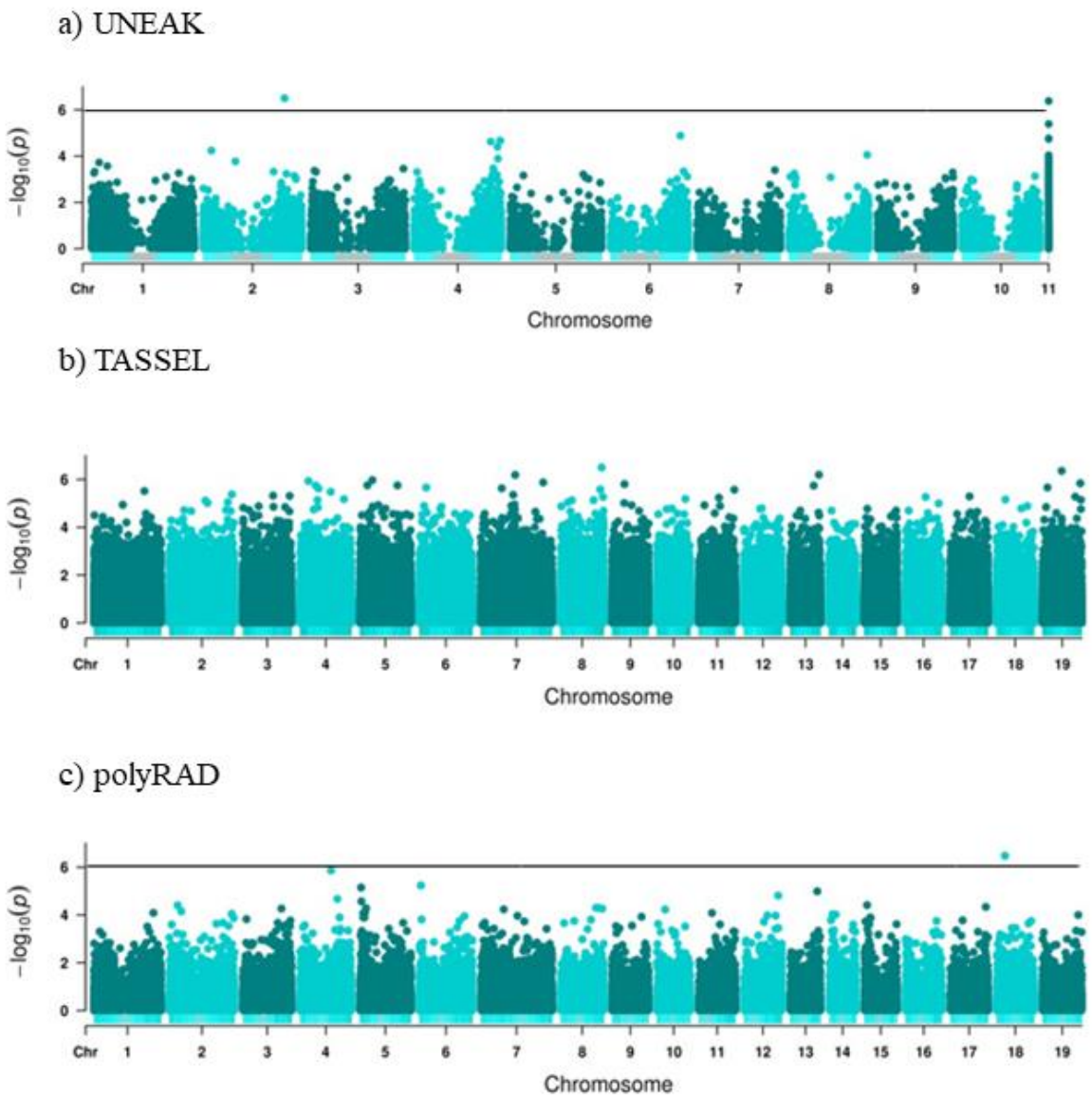

Supplementary Figure 13. A genome wide association study assessing the trait culm node number, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 2 significant SNPs associated with TASSEL pipeline, and c) 3 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).



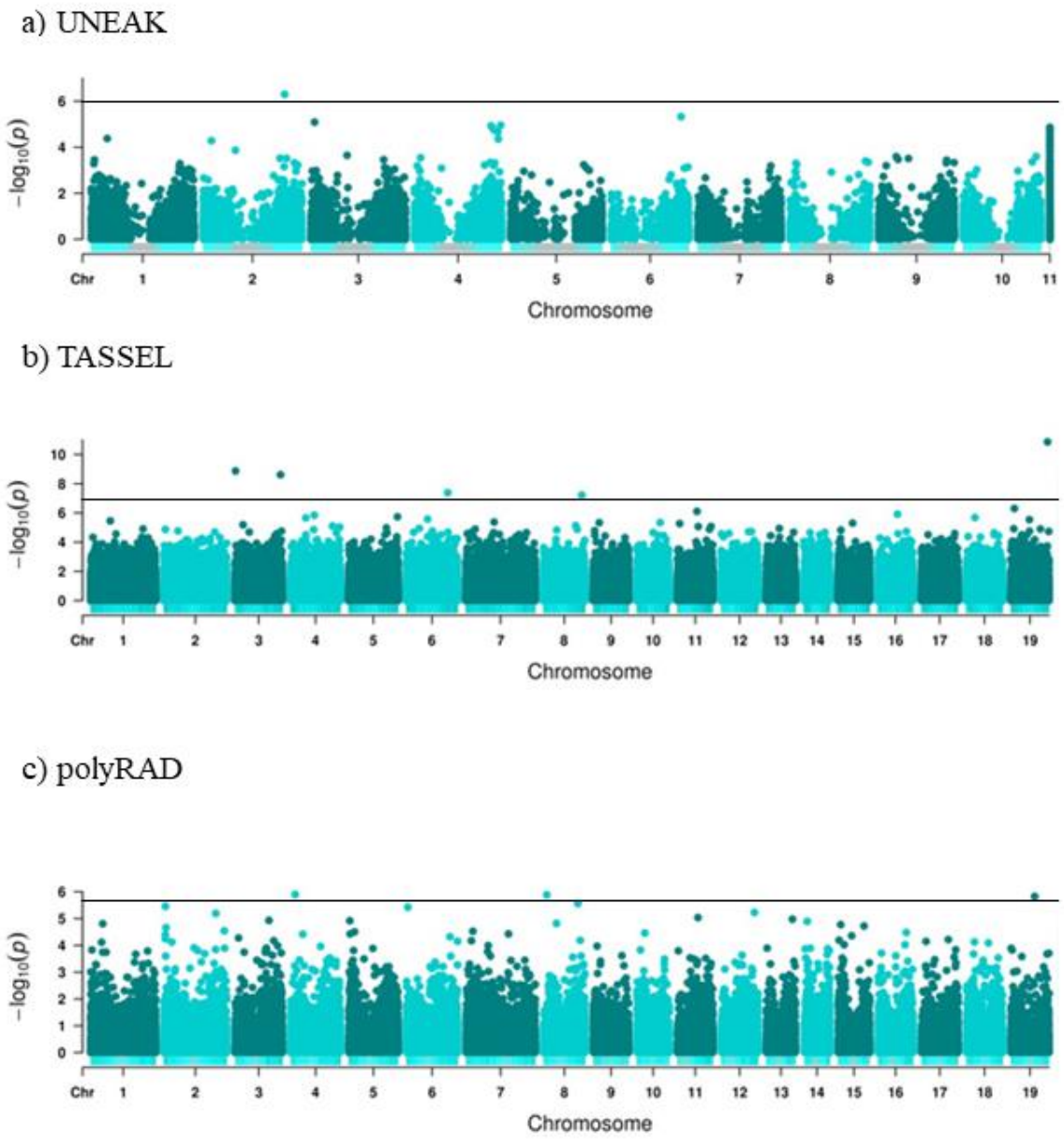

Supplementary Figure 14. A genome wide association study assessing the trait culm node number, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 3 significant SNPs associated with TASSEL pipeline, and c) 2 significant SNPs associated with polyRAD variant calling pipeline. Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).



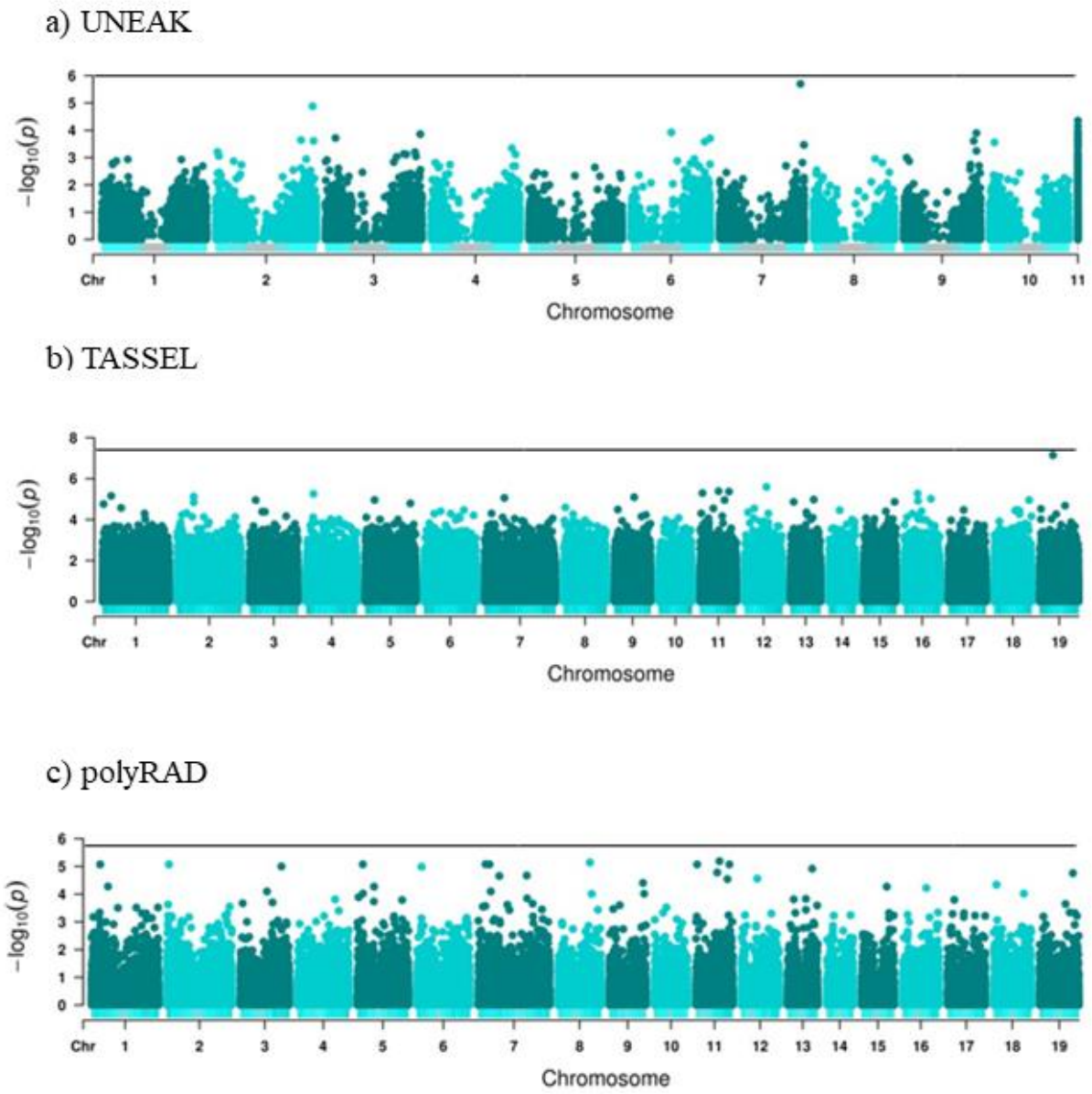

Supplementary Figure 15. A genome wide association study assessing the trait compressed circumference divided by the culms per footprint, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 5 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Zhuji, China by Zhejiang University (ZJU).




Supplementary Figure 16. A genome wide association study assessing the trait culms per footprint, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 2 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 0 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).

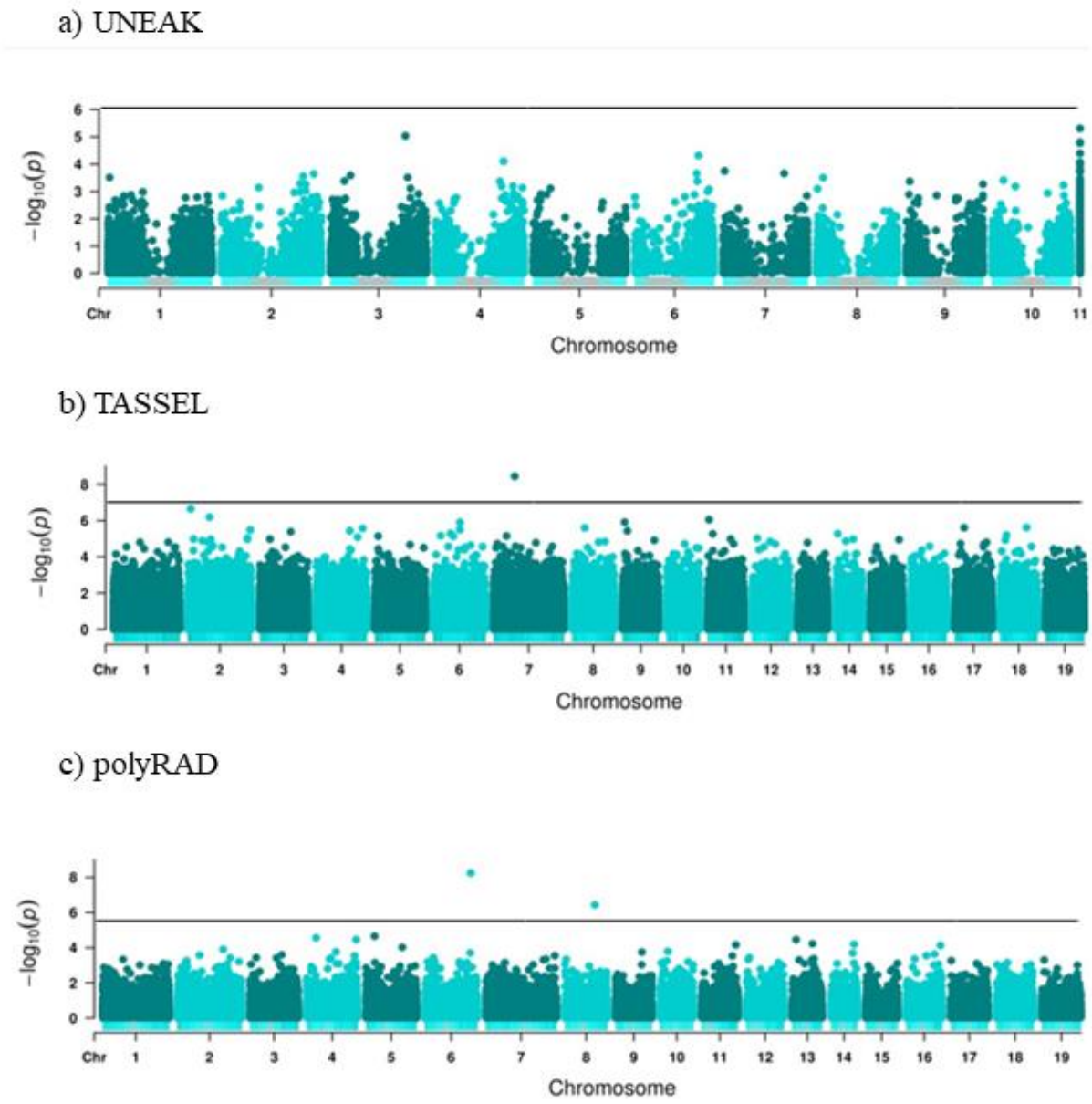



Supplementary Figure 17. A genome wide association study assessing the trait culms per footprint, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 5 significant SNPs associated with TASSEL pipeline, and c) 3 significant SNPs associated with polyRAD variant calling pipeline. Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).

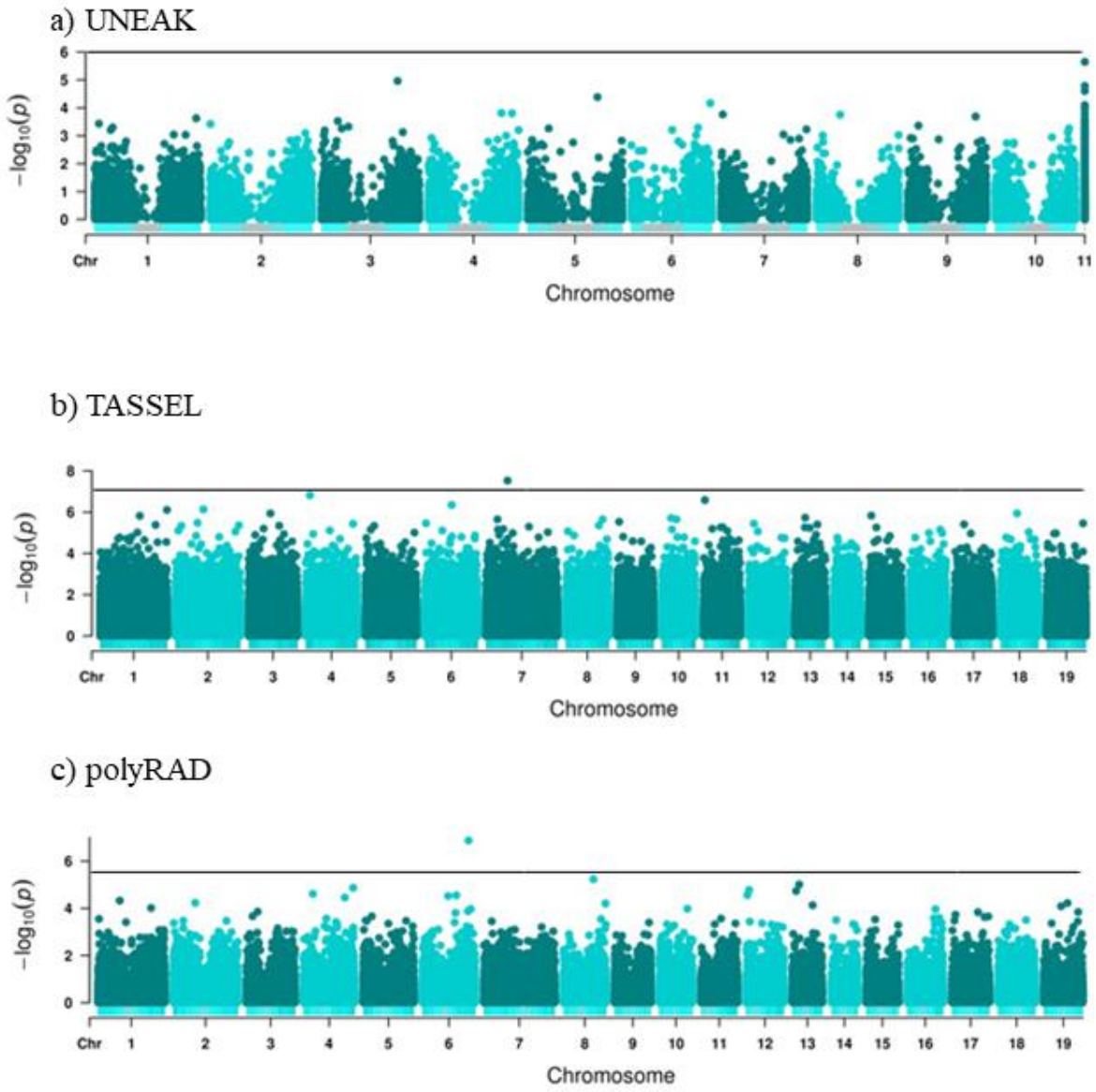



Supplementary Figure 18. A genome wide association study assessing culm volume, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 0 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.

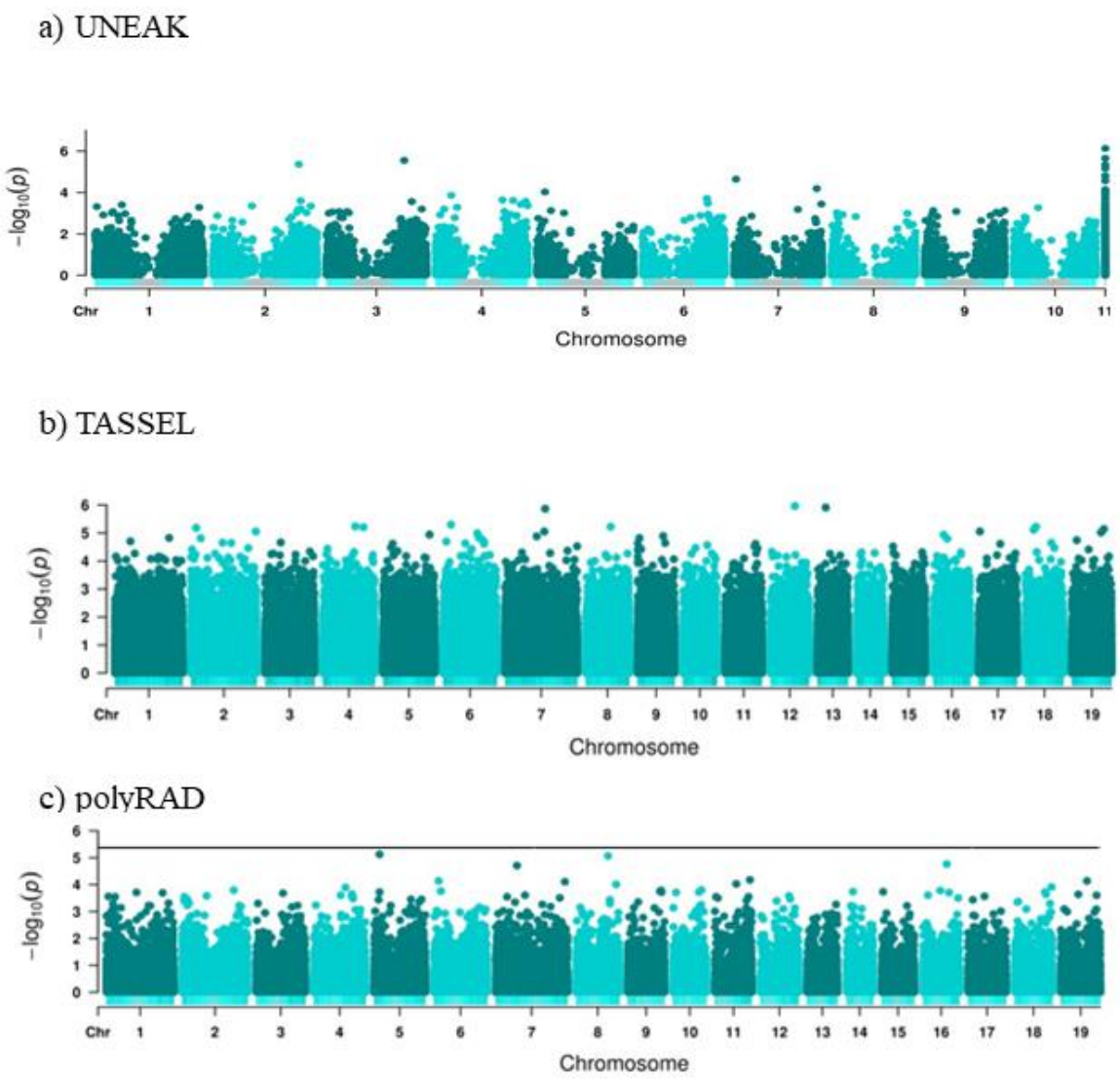
Location(s): Zhuji, China by Zhejiang University (ZJU).



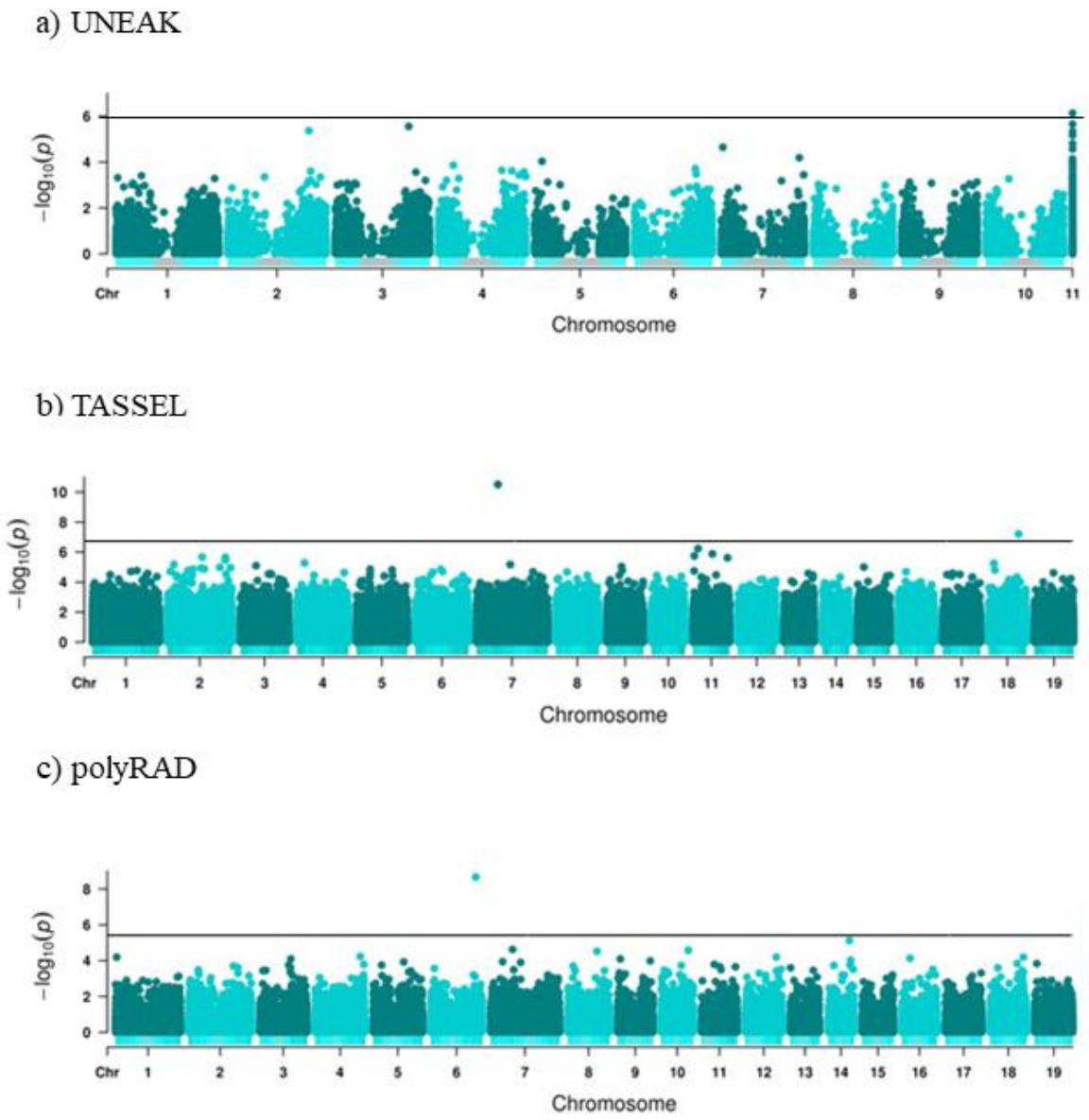

Supplementary Figure 19. A genome wide association study assessing culm volume, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 2 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).



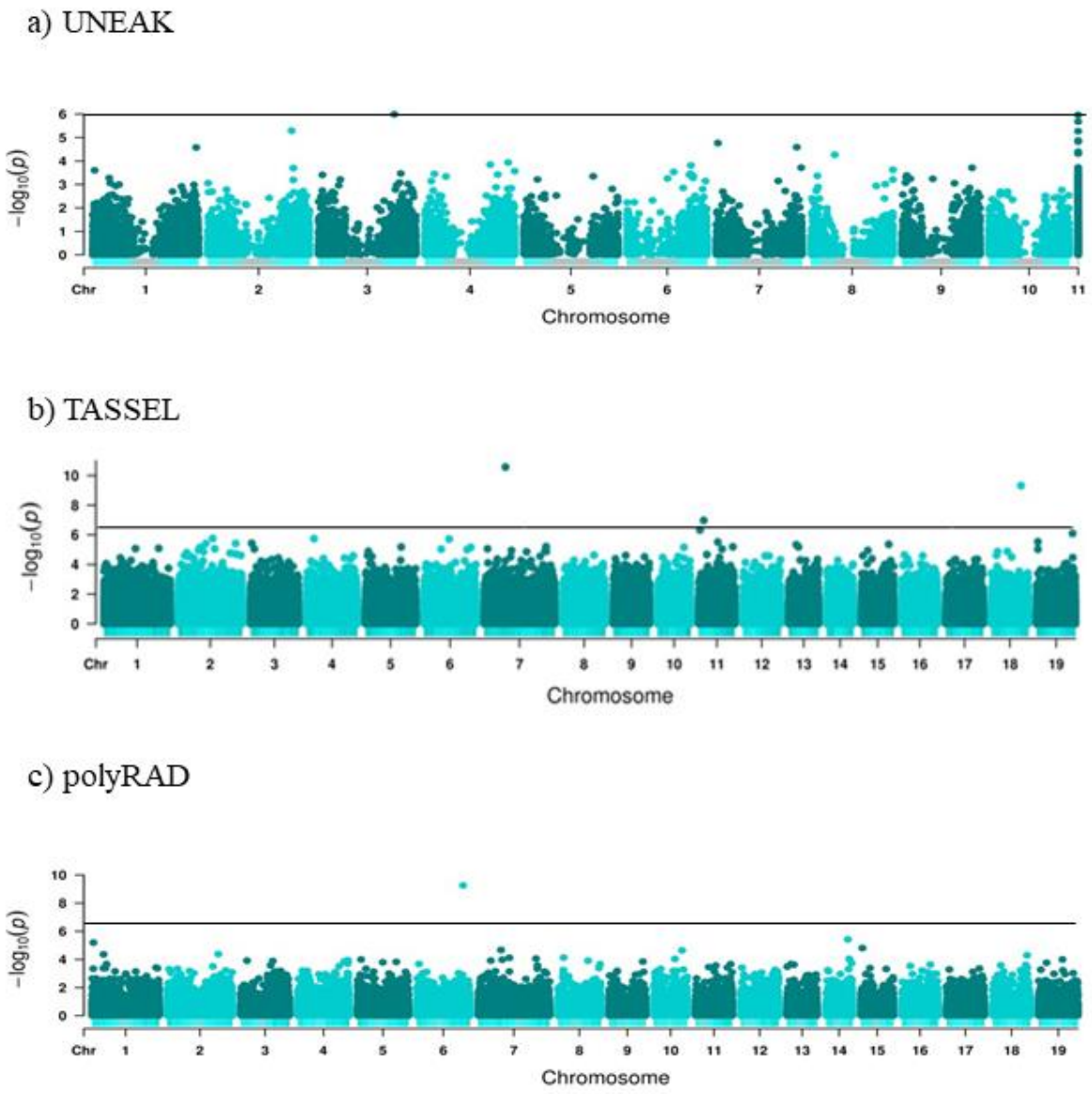

Supplementary Figure 20. A genome wide association study assessing culm volume, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline.
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).



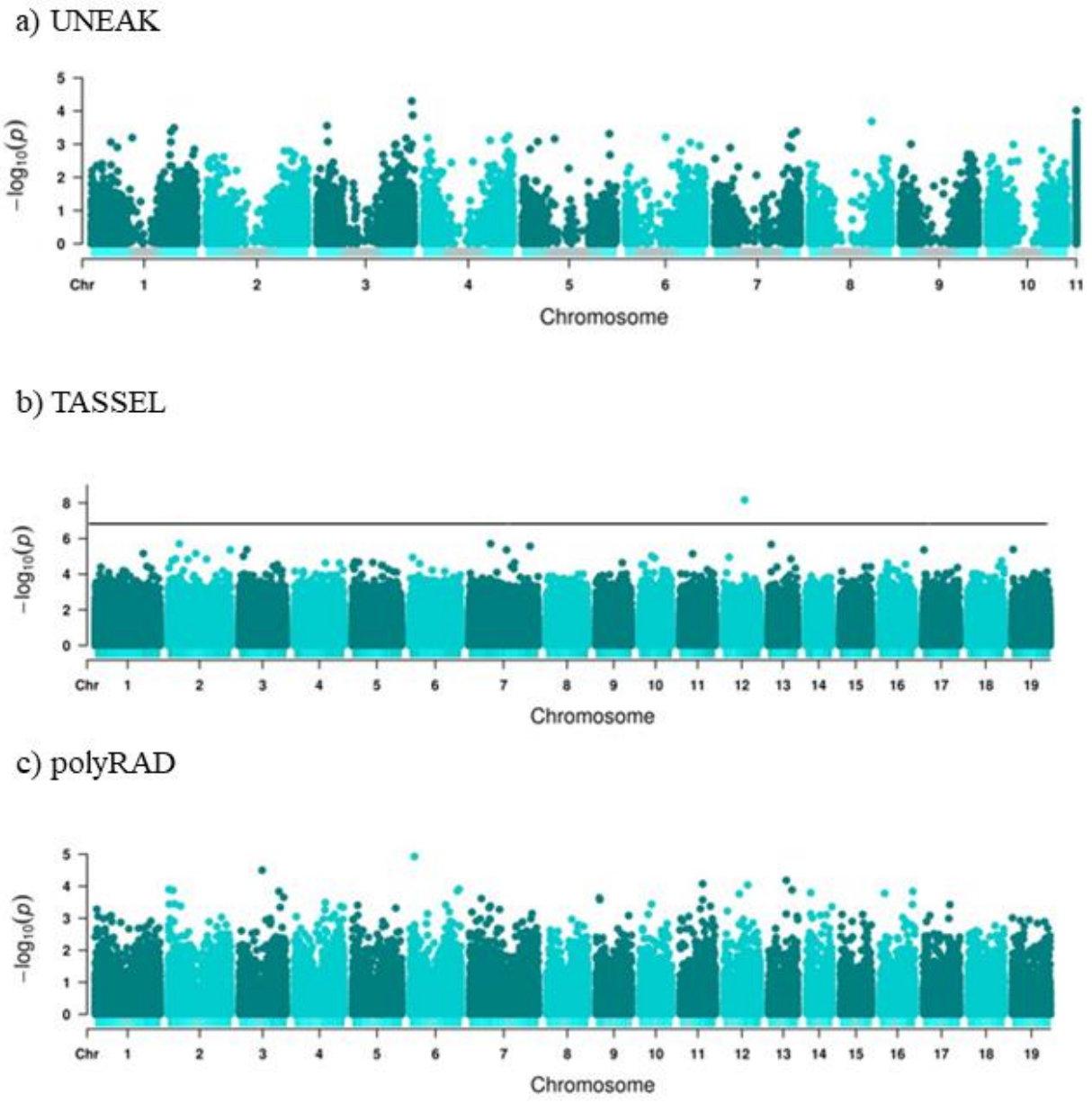

Supplementary Figure 21. A genome wide association study assessing the diameter of the basal internode, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 0 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Zhuji, China by Zhejiang University (ZJU).



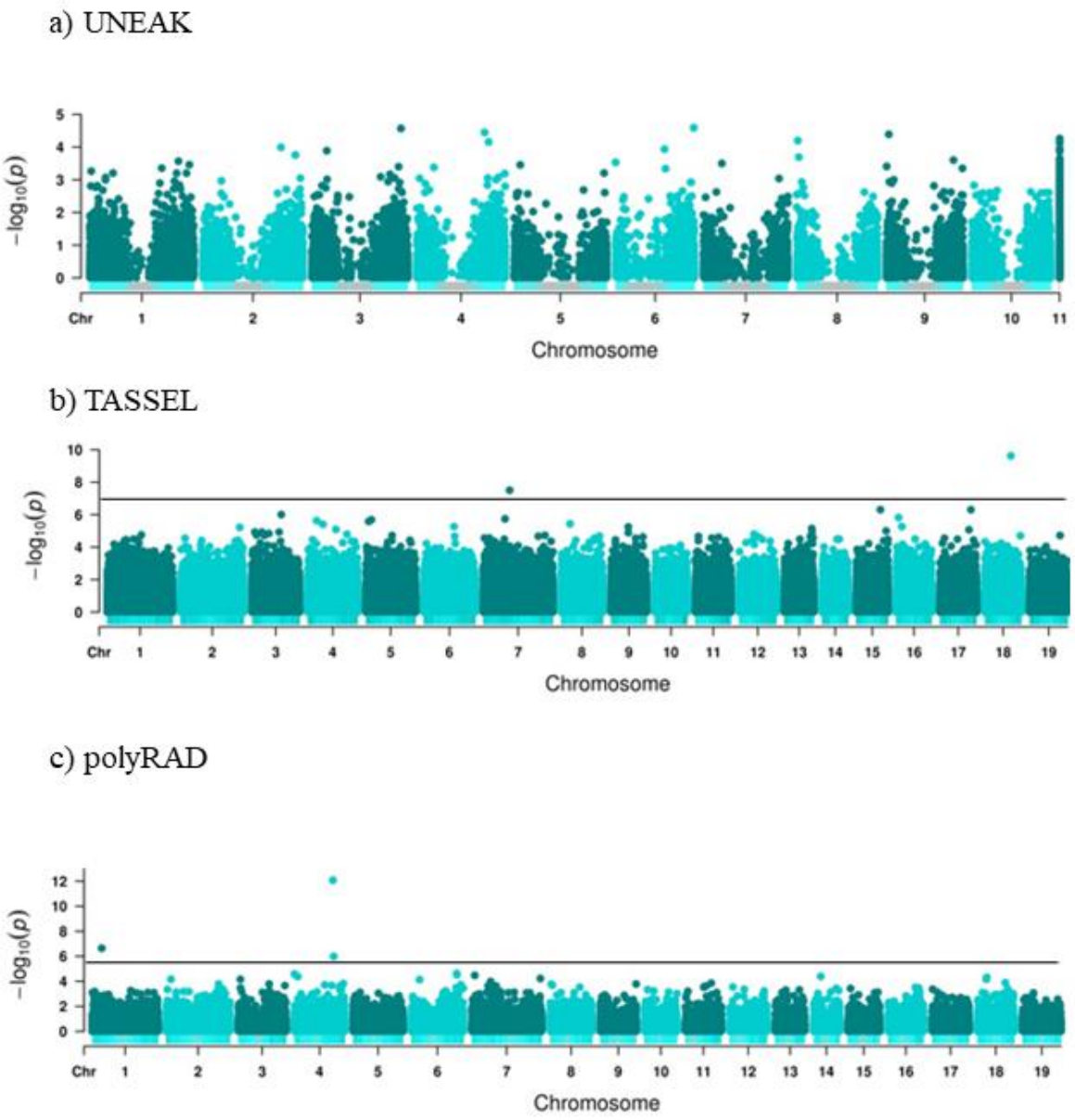

Supplementary Figure 22. A genome wide association study assessing the diameter of the basal internode, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 5 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 2 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).



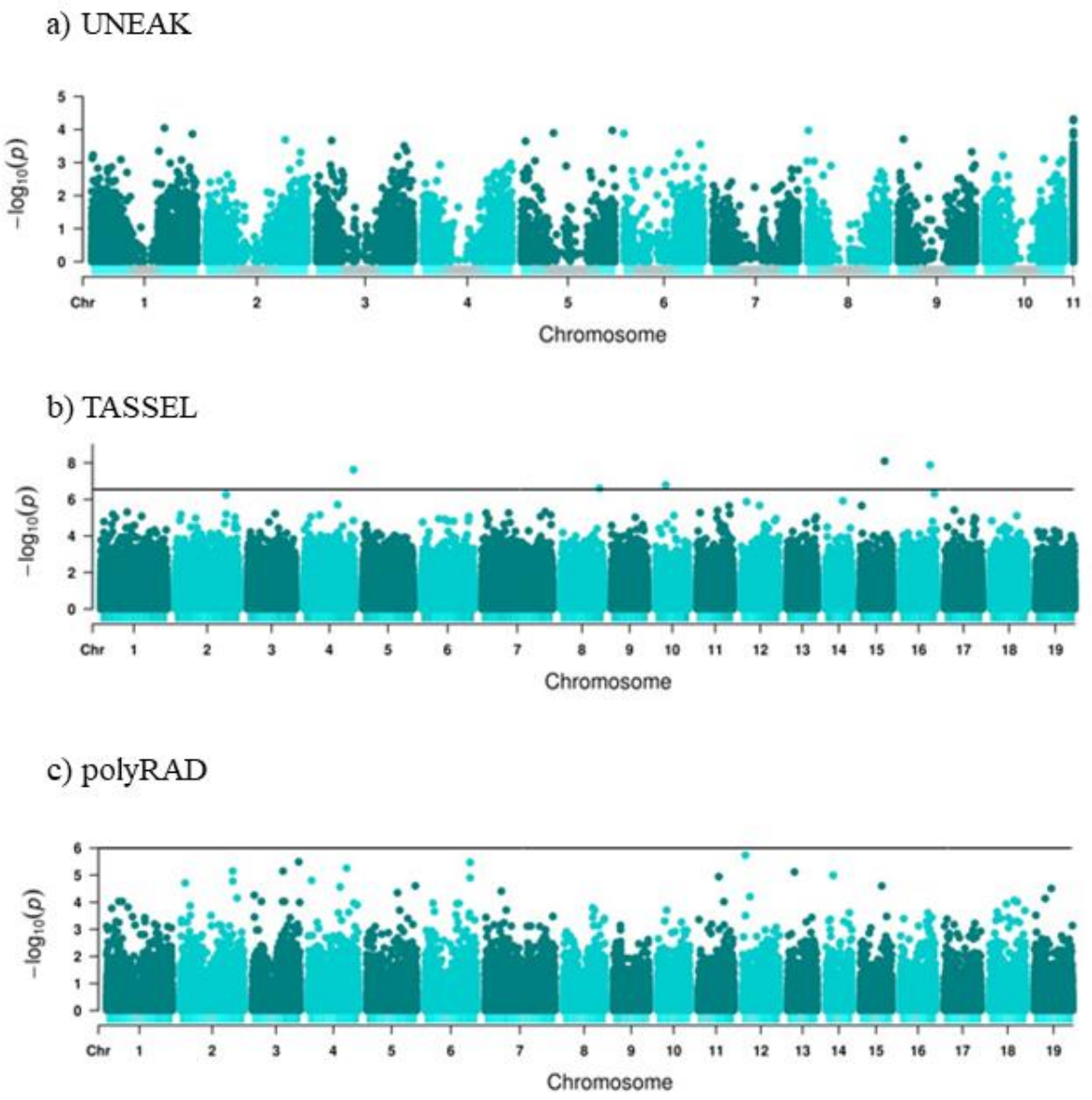

Supplementary Figure 23. A genome wide association study assessing the diameter of the basal internode, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 5 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 3 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline. Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).



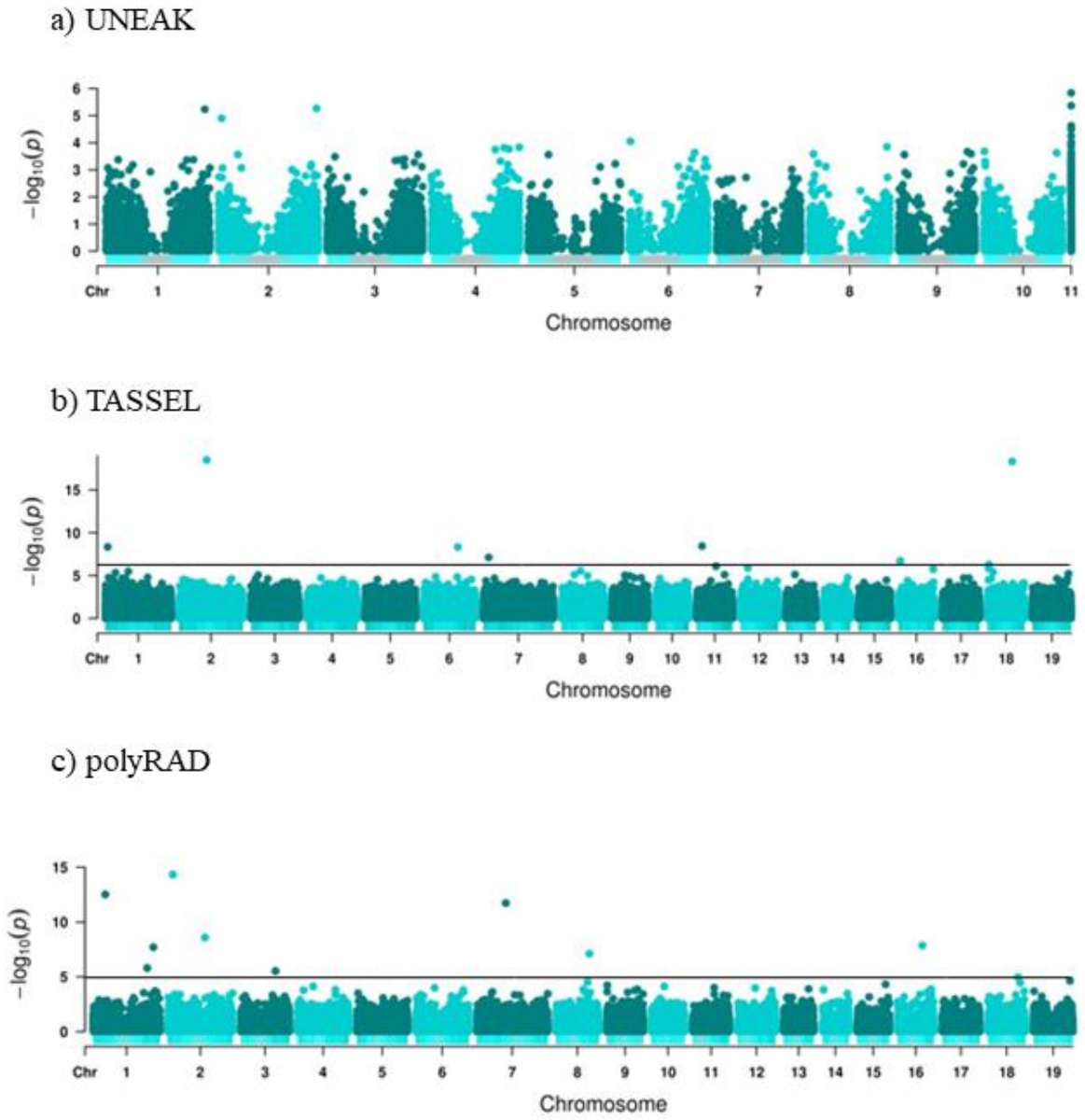

Supplementary Figure 24. A genome wide association study assessing the diameter of the topmost internode, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Zhuji, China by Zhejiang University (ZJU).




Supplementary Figure 25. A genome wide association study assessing the diameter of the topmost internode, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 2 significant SNPs associated with TASSEL pipeline, and c) 3 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).

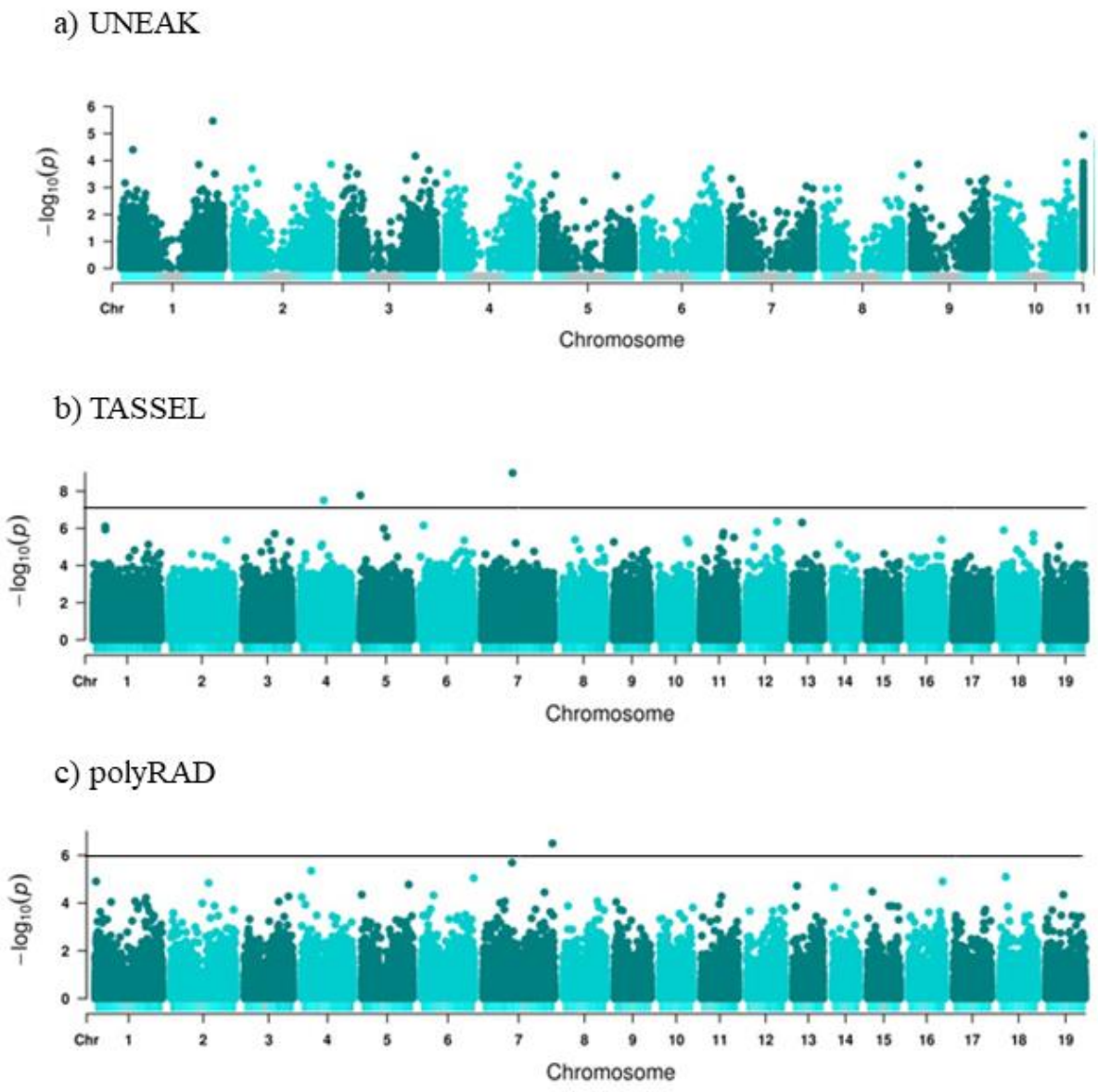



Supplementary Figure 26. A genome wide association study assessing the diameter of the topmost internode, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 4 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline. Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).

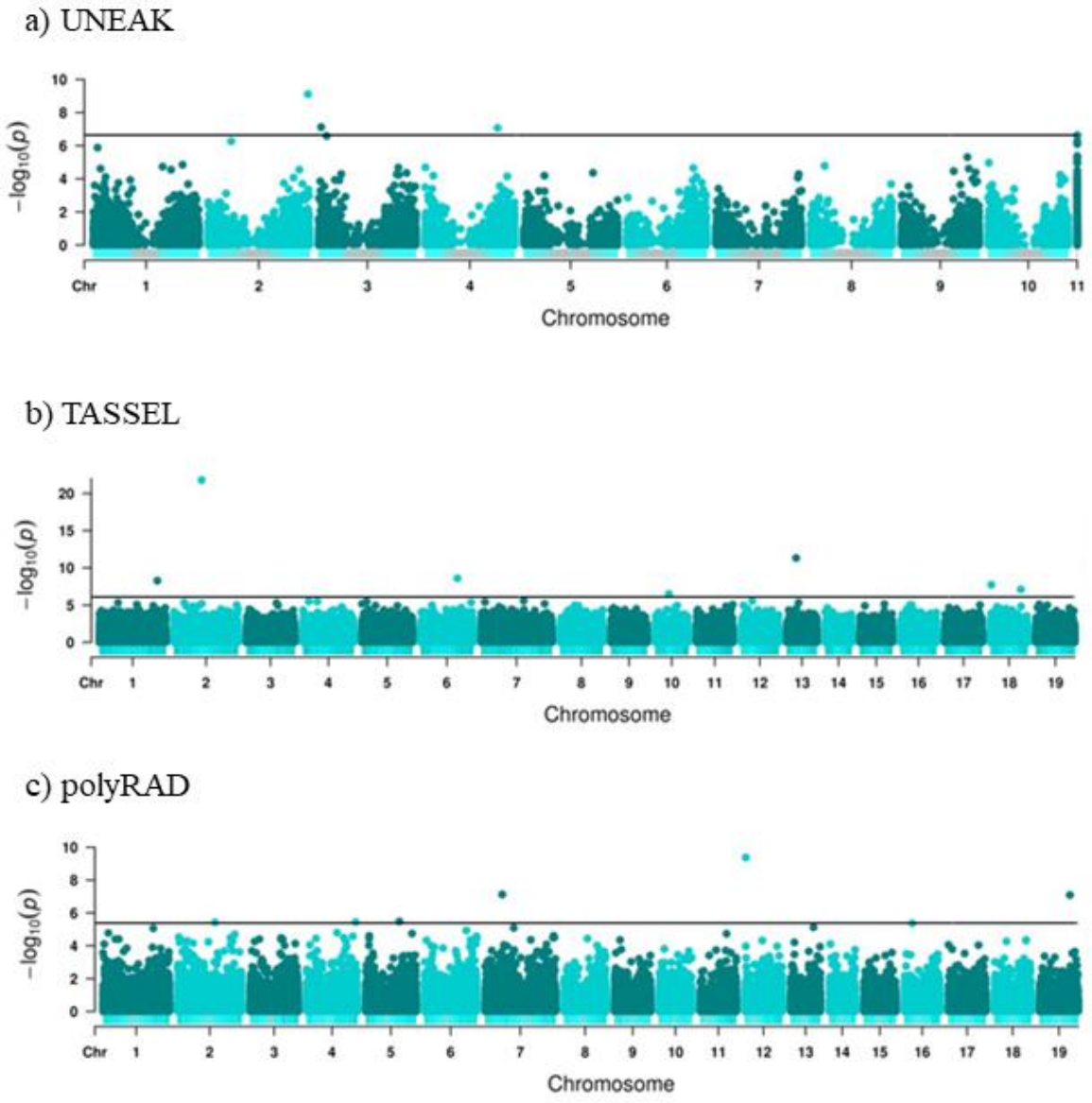



Supplementary Figure 27. A genome wide association study assessing dry biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 7 significant SNPs associated with TASSEL pipeline, and c) 9 significant SNPs associated with polyRAD variant calling pipeline.

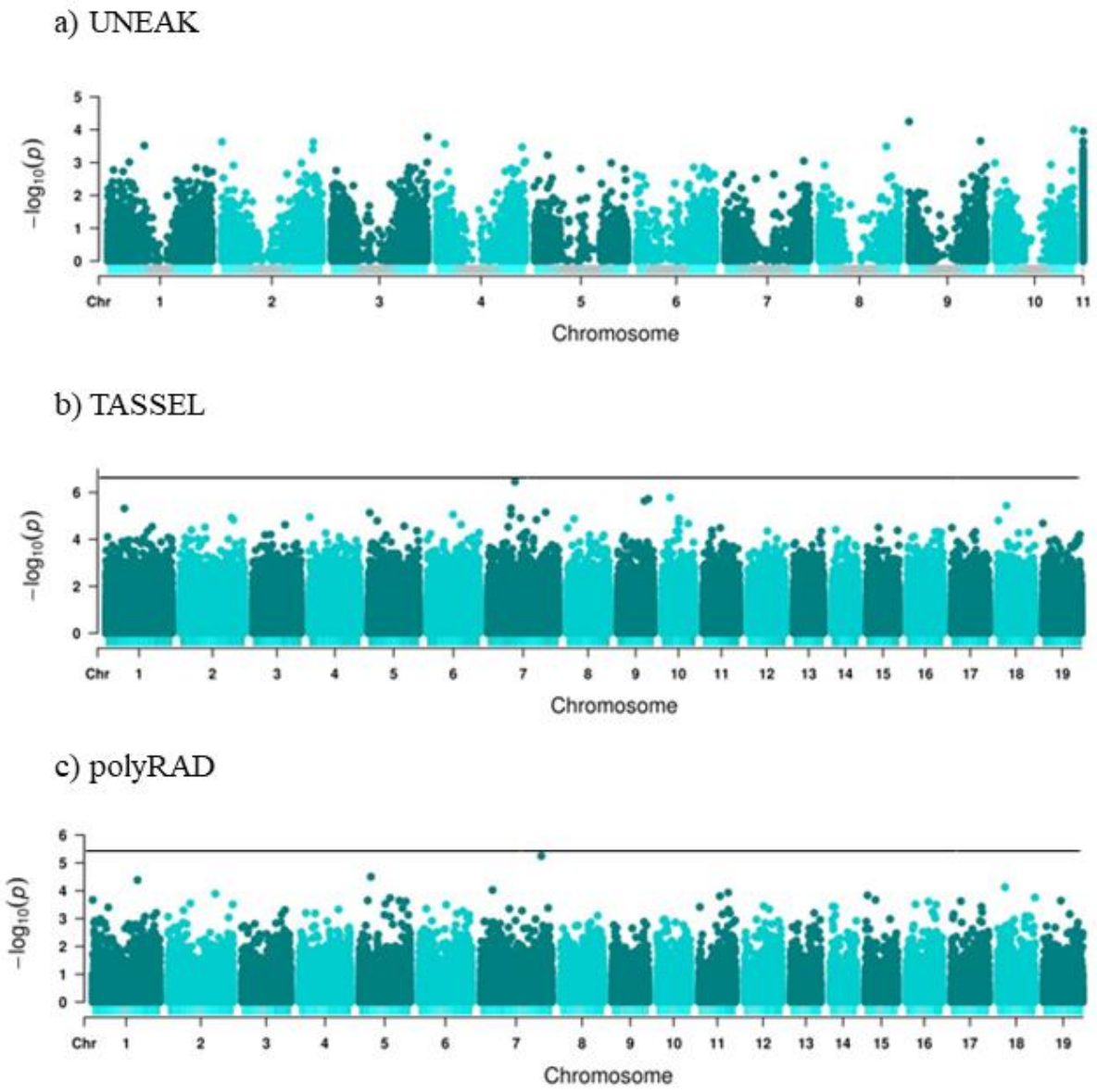
Location(s): Zhuji, China by Zhejiang University (ZJU).




Supplementary Figure 28. A genome wide association study assessing dry biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 3 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).

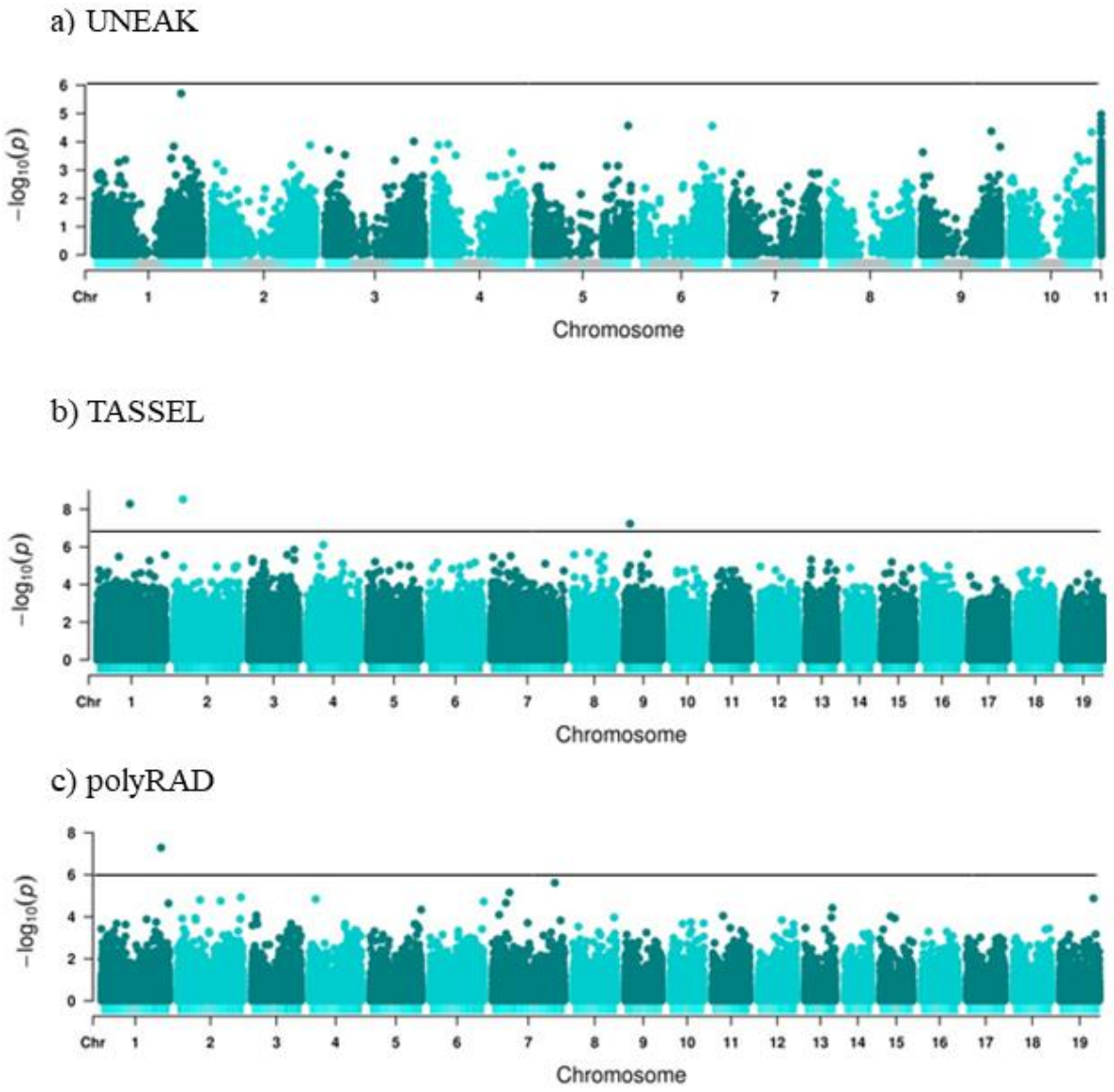



Supplementary Figure 29. A genome wide association study assessing dry biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 26 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 7 significant SNPs associated with TASSEL pipeline, and c) 3 significant SNPs associated with polyRAD variant calling pipeline.
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).

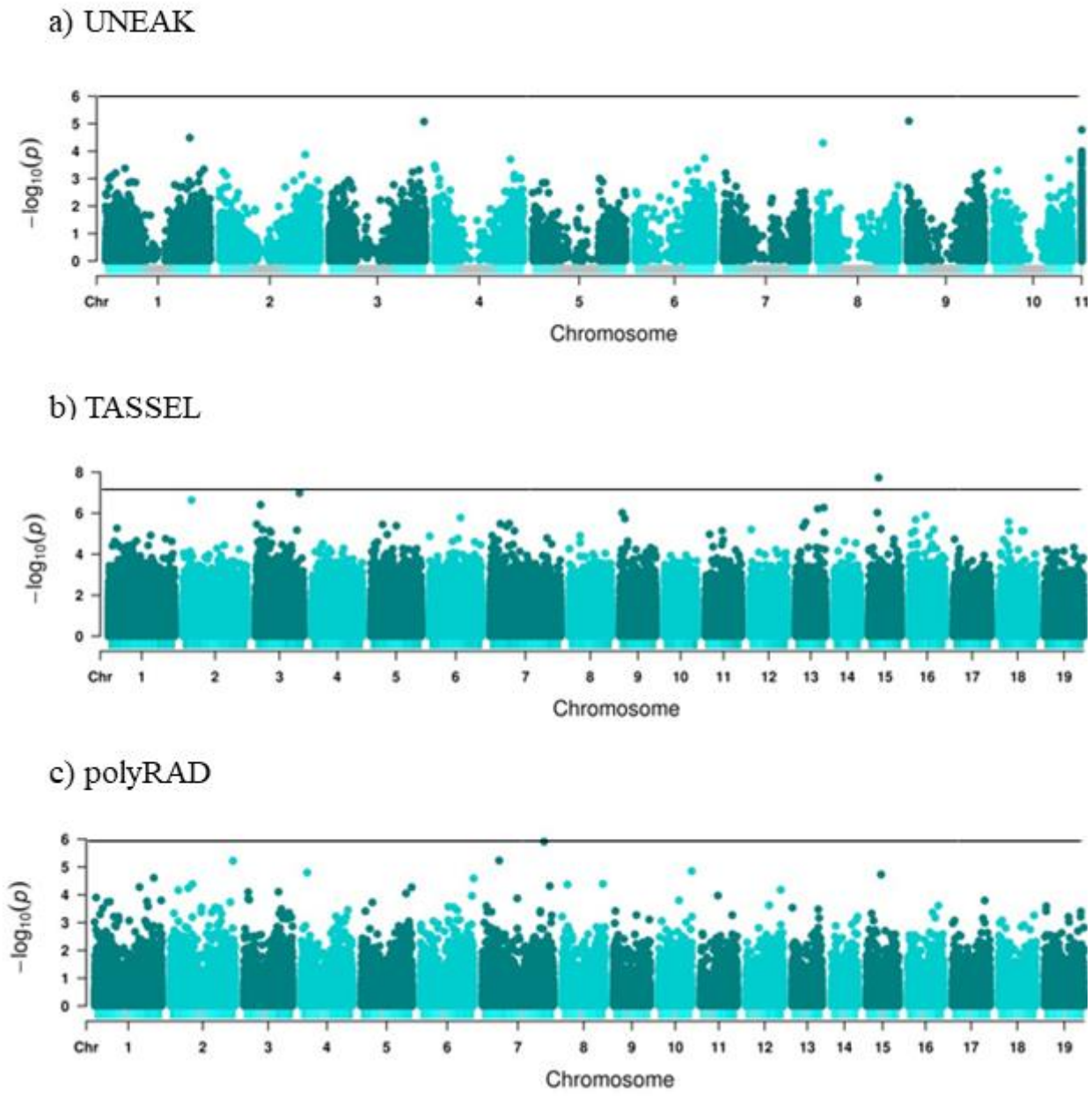



Supplementary Figure 30. A genome wide association study assessing the internode length, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 0 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.

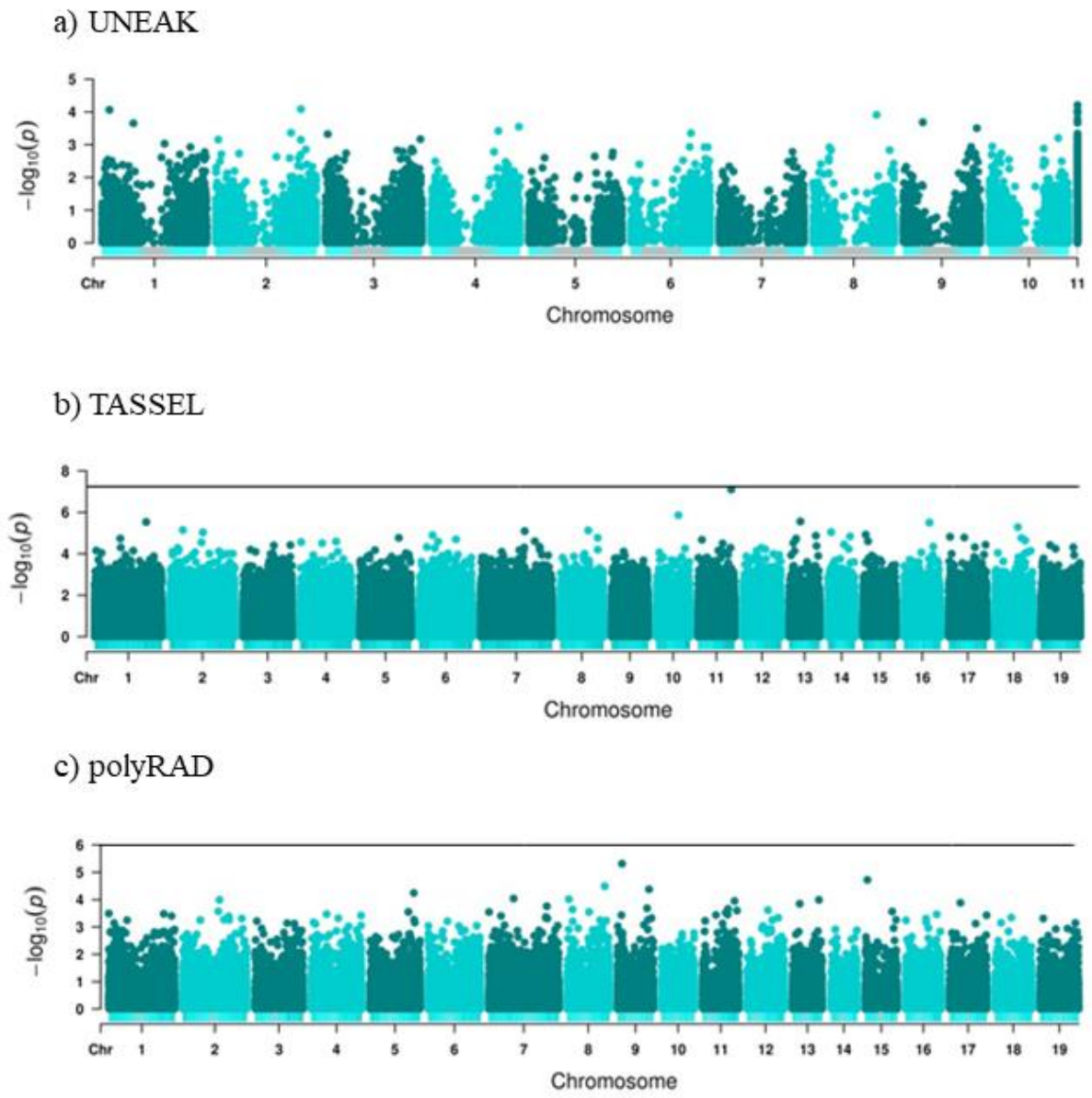
Location(s): Zhuji, China by Zhejiang University (ZJU).



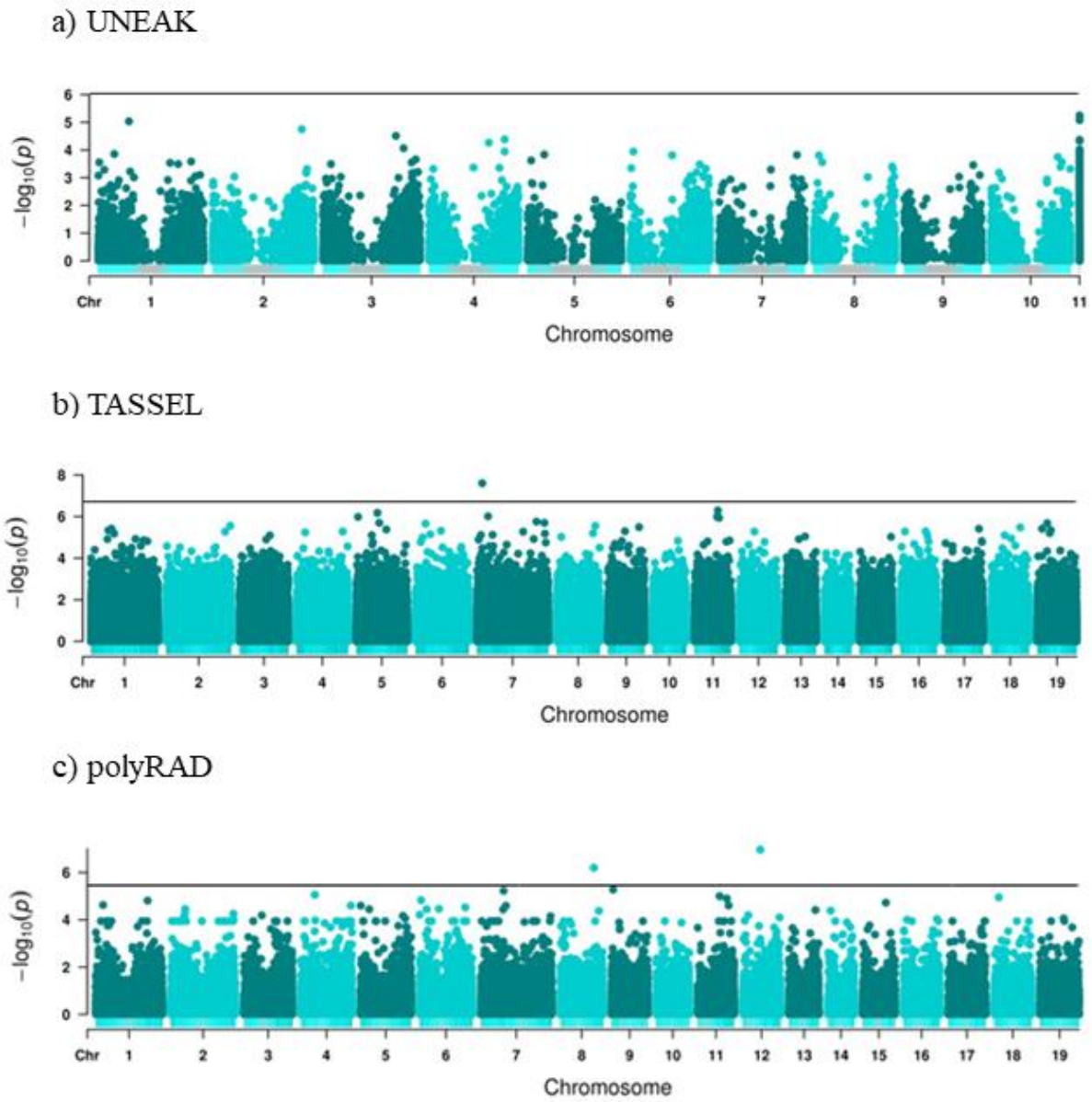

Supplementary Figure 31. A genome wide association study assessing the internode length, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 3 significant SNPs associated with TASSEL pipeline, and c) 1 significant SNPs associated with polyRAD variant calling pipeline. Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).




Supplementary Figure 32. A genome wide association study assessing the internode length, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).

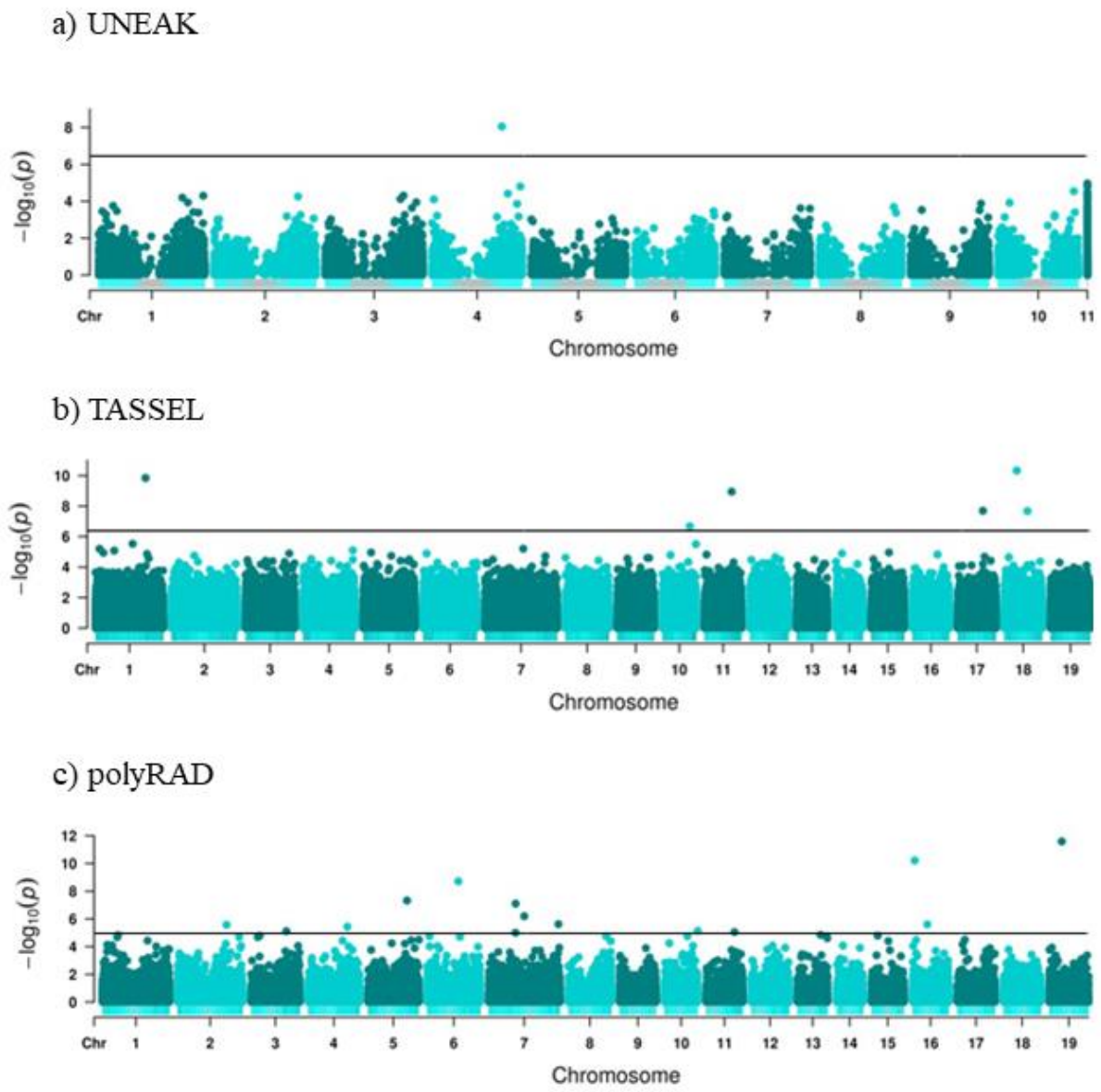



Supplementary Figure 33. A genome wide association study assessing the proportion of reproductive culms, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 0 significant SNPs associated with TASSEL pipeline, and c) 0 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Zhuji, China by Zhejiang University (ZJU).

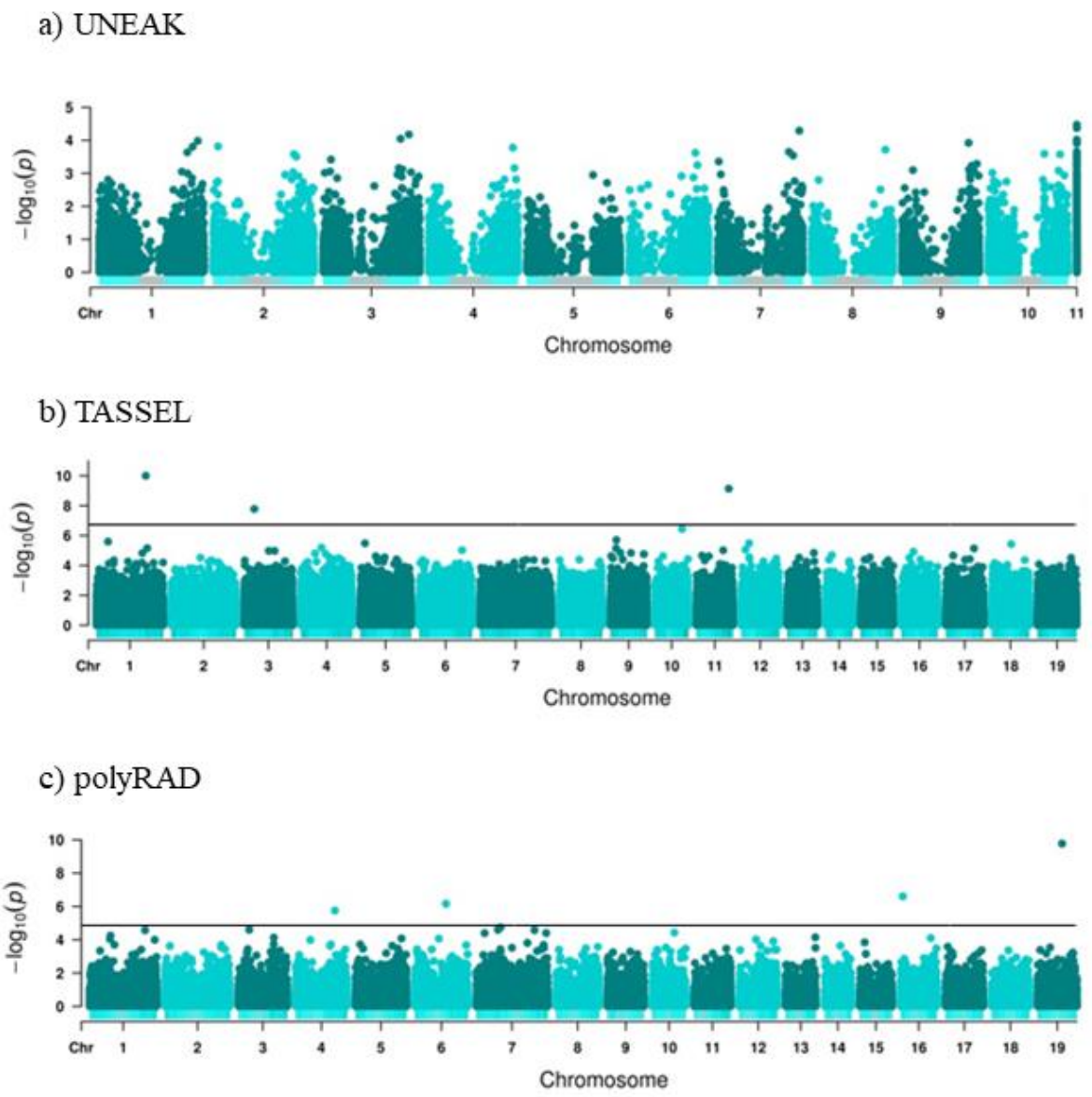



Supplementary Figure 34. A genome wide association study assessing the total number of culms, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from one location. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 1 significant SNPs associated with TASSEL pipeline, and c) 2 significant SNPs associated with polyRAD variant calling pipeline.

Location(s): Zhuji, China by Zhejiang University (ZJU).




Supplementary Figure 35. A genome wide association study assessing the total number of culms, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from 5 locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 1 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 6 significant SNPs associated with TASSEL pipeline, and c) 10 significant SNPs associated with polyRAD variant calling pipeline.
Location(s): Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU).




Supplementary Figure 36. A genome wide association study assessing the total number of culms, a trait associated with biomass yield, within 568 *Miscanthus sinensis* accessions collected from six locations. Comparison of the novel sorting pipeline was conducted with the standard TASSEL pipeline and UNEAK to evaluate the significance of polyRAD variant calling pipeline in the downstream analysis of genomic studies. Positions of SNPs included in the analysis were aligned with respect to the *Sorghum bicolor* v. 3.0 reference genome and resulted in the identification of a) 0 significant SNPs associated with UNEAK non-reference pipeline. Positions of SNPs included in the analysis were aligned with respect to the *Miscanthus sinensis* v. 7.1 reference genome and resulted in the identification of b) 3 significant SNPs associated with TASSEL pipeline, and c) 4 significant SNPs associated with polyRAD variant calling pipeline.
Locations: Sapporo, Japan by Hokkaido University (HU); Leamington, ON by New Energy Farms (NEF); Fort Collins, CO by Colorado State University (CSU); Urbana, IL by the University of Illinois (UI); Chuncheon, Korea by Kangwon National University (KNU); Zhuji, China by Zhejiang University (ZJU).