



Article

Genetic Diversity of Northern Wheatgrass (*Elymus lanceolatus* ssp. *lanceolatus*) as Revealed by Genotyping-by-Sequencing

Pingchuan Li ^{1,†} , Surendra Bhattarai ^{1,†}, Gregory W. Peterson ², Bruce Coulman ¹, Michael P. Schellenberg ³, Bill Biligetu ^{1,*} and Yong-Bi Fu ^{2,*} 

¹ Department of Plant Sciences, University of Saskatchewan, 51 Campus Drive, Saskatoon, SK S7N 5A8, Canada; lipingchuan@gmail.com (P.L.); surendra.bhattarai@usask.ca (S.B.); bruce.coulman@usask.ca (B.C.)

² Plant Gene Resources of Canada, Saskatoon Research and Development Centre, Agriculture and Agri-Food Canada, 107 Science Place, Saskatoon, SK S9H 3X2, Canada; Gregory.peterson@agr.gc.ca

³ Swift Current Research and Development Center, Agriculture and Agri-Food Canada, Box 1030, Swift Current, SK S9H 3X2, Canada; mike.schellenberg@agr.gc.ca

* Correspondence: bill.biligetu@usask.ca (B.B.); yong-bi.fu@agr.gc.ca (Y.-B.F.); Tel.: +1-306-966-4007 (B.B.); +1-306-385-9298 (Y.-B.F.)

† These authors contributed equally to the research.

Received: 29 March 2018; Accepted: 9 April 2018; Published: 11 April 2018



Abstract: Recent advances in next generation sequencing technologies make genotyping-by-sequencing (GBS) more feasible for the molecular characterization of plant germplasm with complex and unsequenced genomes. This study represents the first preliminary effort using GBS to discover genome-wide genetic variants of northern wheatgrass (*Elymus lanceolatus* ssp. *lanceolatus* (Scribn. and J. G. Sm.) Gould) plants and to assess the genetic diversity present in four cultivated and six wild accessions. The effort generated the first novel set of genomic resources and 5659 single nucleotide polymorphism (SNP) markers for this tetraploid grass. The diversity analysis revealed 8.8% of SNP variation residing among the 10 accessions and 1.9% SNP variation present between cultivated and wild accessions. The Bayesian analysis identified three major clusters of the assayed samples, and the principal coordinates analysis revealed the genetic distinctness of the two accessions collected from Nevada and Wyoming. The flow cytometry analysis confirmed the tetraploid nature of some of the assayed samples and estimated the average genome size to be 9.3–9.4 Gb for this species. These findings are useful for the genetic improvement of this native grass species for forage production and rangeland reclamation. The findings are also encouraging for the broad application of genotyping-by-sequencing in the characterization of genome-wide genetic variability in non-model polyploid plants.

Keywords: genotyping-by-sequencing; *Elymus*; genetic diversity; genetic structure; SNP

1. Introduction

Genotyping-by-sequencing (GBS) has emerged as a powerful genomic approach for characterizing the genetic diversity of non-model plants on a genome-wide scale [1–4]. Generally, the GBS analysis involves five major steps: (1) reducing genome complexity with restriction enzymes; (2) barcoding enzyme-cut genomic DNAs with indexed adaptors; (3) sequencing the barcoded DNA fragments in a high-throughput DNA sequencer; (4) a bioinformatics analysis of indexed sequence reads to identify genetic variants; and (5) a genetic diversity analysis of assayed samples based on a sample-by-variant matrix [5]. A GBS application can produce high-density, low-cost genotype information without the

need for a reference genome sequence [6]. However, recent GBS applications have also revealed some weaknesses, including a large number of missing data points, uneven genome coverage, complex bioinformatics, and issues related to polyploidy [7–10]. To address some of these challenges, Tinker et al. [11] developed a GBS-based pipeline called Haplotag that can generate tag-level haplotype and single nucleotide polymorphism (SNP) data for polyploid organisms. Successful applications of this approach have been reported in a comparative analysis of diploid, tetraploid, and hexaploid genomes of 27 oat species [12] and in an adaptation study of cultivated oat [13].

Northern wheatgrass (NWG; *Elymus lanceolatus* ssp. *lanceolatus* (Scribn. and J. G. Sm.) Gould), also known as thickspike wheatgrass, is one of the 150 species belonging to the largest genus *Elymus* in the Triticeae tribe [14–16]. This grass is mainly an allotetraploid ($2n = 4x = 28$) with a SSHH genome [17] and is drought-tolerant and rhizomatous with excellent germination and early-spring vigor [18]. Its native range extends from Canada south through North Dakota, Nebraska, Colorado, Wyoming, and New Mexico and west to the Pacific coast [19]. It is also commonly found in the northern Rocky Mountains and adjacent Great Plains regions. Established NWG stands can withstand grazing and trampling, but under heavy and prolonged grazing, some NWG stands may decrease and can be replaced by other grasses or shrubs [20]. This native grass can persist as a dominant species on burned sites for over 30 years [21], recover rapidly following a fire [22], and have a major impact on fire cycles on native rangeland. Northern wheatgrass, similar to other *Elymus* species, has been used for the revegetation of rangeland and has played an important role in soil stabilization, erosion control, and forage production [23]. However, some challenges remain in the utilization of this native grass. First, there is lack of commercial seed quantities for this species. Second, interest in its breeding seems to have increased, but the development of NWG varieties has been relatively slow. Third, few genetic studies have been conducted to assess its genetic variability and structure in natural stands, but genetic diversity is important for reclamation practices and forage production. A sufficient level of genetic diversity promotes plant adaptation in a new environment and enhances the long-term sustainability of a plant community [24] and its adaptation for future climatic changes.

Efforts have been made to assess the genetic diversity of many other *Elymus* species using different molecular markers. For example, a number of studies have been conducted on the genetic diversity of *E. caninus* [25,26], *E. fibrosus* [27–29], *E. alaskanus* [30–33], *E. trachycaulus* [34,35], and *E. sibiricus* [36,37]. These studies have shown that each species possesses a unique pattern of genetic variability, and they are useful for understanding the genetic diversity for similar outcrossing grass species. However, no studies have been found to assess the genetic diversity in NWG plants. Research is also needed to assess the morphological, anatomical and molecular characteristics of NWG plants through plant breeding. Some of these assessments, however, require the development of informative genetic markers such as SNP, which is challenging for NWG as a non-model polyploid plant that lacks genomic resources. Fortunately, recent GBS applications to polyploidy plants [12,13] show the feasibility of using GBS to discover genome-wide genetic variability in NWG plants and to characterize NWG germplasm for its breeding and genetic research.

The objectives of our study were to apply GBS in combination with the Haplotag pipeline to (1) generate a novel set of genomic resources for NWG plants, (2) identify genome-wide SNP markers, (3) assess the genetic diversity present in 10 NWG accessions, and (4) evaluate the utility of the GBS application in the genetic diversity analysis of complex polyploid plants.

2. Materials and Methods

2.1. Plant Materials

We acquired 31 presumed NWG accessions from the plant germplasm system of the United States Department of Agriculture—Agricultural Research Service (USDA-ARS) and one accession from the Plant Gene Resources of Canada (PGRC) for the joint forage breeding program of the University of Saskatchewan and Agriculture and Agri-Food Canada (AAFC). For this study, we selected seven wild

accessions that were collected from native stands to represent the areas where this species is found in the United States of America (USA) and one wild accession collected originally from a native stand in Iran. We also included three registered cultivars and one breeding line in the USA and Canada, and these cultivated accessions represent superior plants that have been selected for certain traits from wild populations and have or will have seeds available for planting. Seeds were randomly selected from each accession and grown for six weeks in a greenhouse at the AAFC Saskatoon Research and Development Center under the following growth conditions: 16 h photoperiod at 22 °C and 8 h dark at 16 °C. Young leaf tissues were collected from 12 randomly selected plants for each of the 12 accessions and stored at −80 °C in a freezer prior to DNA extraction. The sampled plants, after 3–4-months growth, were assessed for species identity, and one accession TMP24017 was identified as western wheatgrass (*E. smithii*). To help the identity verification of the assayed samples, efforts were also made to assess the genome size and ploidy of some of the samples (see Section 2.2). The accession TMP24008 was identified as an octoploid, but not a tetraploid, northern wheatgrass. These identifications were done after genotyping-by-sequencing of the 144 samples for the 12 accessions as described below, but only the samples of the 10 accessions listed in Table 1 were used for the bioinformatics and genetic diversity analysis.

Table 1. List of the 10 northern wheatgrass (*E. lanceolatus* ssp. *lanceolatus*) accessions studied and their information and genome size estimation.

Accession ^a	Alt Acc ^a	Germplasm Type	Collected/Received ^a	Label	GsrTa ^a	GsrTd ^a
CN37154		cultivar ‘Elbee’	AB, Canada (1980)	AB	9.195	8.841
TMP24006	PI632756	cultivar ‘Schwendimar’	WA, USA (2003)	WA	9.531	9.397
TMP24007	PI469235	cultivar ‘Critana’	MT, USA (1982)	MT1	9.574	9.478
TMP24013	PI562038	Cultivated	MT, USA (1992)	MT2		
TMP24010	W614632	wild accession	WY, USA (1993)	WY		
TMP24011	W616745	wild accession	NV, USA (1995)	NV	9.412	9.294
TMP24012	PI564552	wild accession	UT, USA (1992)	UT		
TMP24015	PI562037	wild accession	ID, USA (1992)	ID		
TMP24016	PI552794	wild accession	CO, USA (1991)	CO		
TMP24018	PI380620	wild accession	Iran (1972)	IR		

^a Accession is numbered by the Plant Gene Resources of Canada. Alt Acc is the alternative accession number, maintained in the Germplasm Resources Information Network (GRIN) system of USDA-ARS. Collected/received shows the germplasm origin and the date of collection or acquisition. WA, Washington; WY, Wyoming; IR, Iran; NV, Nevada; AB, Alberta; UT, Utah; MT, Montana; ID, Idaho; CO, Colorado. GsrTa is the estimation of genome size (in Gb) relative to *Triticum aestivum*. GsrTd is the estimation of genome size (in Gb) relative to *T. durum*.

2.2. Genotyping-by-Sequencing

For each of the 144 samples, DNA was extracted from 0.1 g of well-ground tissue powder, following the protocols of the NucleoSpin® Plant II Kit (Macherey-Nagel, Bethlehem, PA, USA), and was eluted in a 2-mL Eppendorf tube using Elution Buffer. The DNA quality was assessed with a NanoDrop 8000 (Thermo Fisher Scientific, Waltham, MA, USA) by comparing the 260 and 280 nm absorption. The qualified DNA samples were quantified through the Quant-iT™ PicoGreen® dsDNA assay kit (Invitrogen, Carlsbad, CA, USA) and subsequently diluted to 60 ng/μL with 1× TE buffer prior to further treatment. A total of 144 DNA samples were acquired for sequencing.

Multiplexed GBS libraries were prepared following a minor modified gd-GBS protocol [5]. In brief, each library preparation was started with 200 ng of purified genomic DNA by restriction enzyme digestion. A combination of *HinfI* + *HpyCH4IV* was applied to digest the DNA, as this combination displayed better performance over the conventional *PstI* + *MspI* combination in terms of genome coverage and SNP genotyping accuracy [9]. A ligation-specific customized 5'/3' adapter and inserts by T4 ligase were applied to all the samples. Ligation fragments were purified by using an AMPure XP kit and subsequently amplified with Illumina TruSeq HT multiplexing primers. Prior to the pooling of samples into a library, amplicon fragments were pre-selected by a Pippin Prep (Sage Science, Beverly, MA, USA) instrument for an insert size range between 250 and 450 bp; that is, fragment size is between

400 and 600 bp. All three well-pooled sample libraries were diluted to 6 pM and denatured with 5% of sequencing-ready Illumina PhiX Library Control that can serve for calibration. The sequencing was performed in the Illumina MiSeq Instrument with the option of paired-end reads of 251-bp length. All the raw pair-end sequencing data in FASTQ format were deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject ID PRJNA392957. The Supplementary Materials Table S1 lists the sequencing information for all 144 assayed samples.

2.3. Bioinformatics Analysis

The sequence analysis started with removing the 24 FASTQ files associated with the two accessions, TMP24008 and TMP24017, and one failed sample (no. 6) from accession TMP24016. This removal generated a list of FASTQ files for 119 samples from the 10 accessions. The selected sequence data were cleaned with Trimmomatic v0.32 [38] to remove any sequenced-through Illumina adapters, low-quality sequence (sliding window of 10 bases, average Phred of 20), and fragments under 64 bases long.

Efforts were made to separate a FASTQ file of 250 bp into three separate fragment sets, as the UNEAK-GBS pipeline [39] considers only sequences of 64 bp (after barcode removal) with an intact 4-base *HinfI* residue (ANTC) at the beginning. A custom Perl script *fasta184CutandCode.pl* was developed to separate each input sequence of 250 bp into three sets: the first 64 bases where the script replaced the *HinfI* residual sequence with a *Sau3AI* residual sequence (GATC) that is recognized by UNEAK and the next two 60-base portions with an added 4-base *Sau3AI* residual sequence. Also, because the UNEAK pipeline expects to deconvolute barcoded sequence reads, which are not already separated by sample, the script added an arbitrary barcode sequence (CATCAT) in front of each sequence fragment. The resulting three fragments were each 70 bases long and were recognized by the UNEAK-GBS pipeline [39]. Note that the relationship between the three fragment sets was not preserved going into UNEAK, and each fragment set was passed into UNEAK as an independent data set.

UNEAK and the Haplotag pipeline [11] were run for each 70-base fragment set resulting in the analysis of a total of 180 bases of genetic sequence. Running UNEAK with the conditions (described in the Supplementary Materials Section B) generated two types of meta data files: a single mergedAll.txt (all tags observed more than 10 times) and a set of individual tagCount files (one per sample) required for the Haplotag pipeline.

Running Haplotag with the parameters and filtering thresholds shown in the HTIndex.txt file (described in the Supplementary Materials Section B) generated a matrix of samples by SNP loci. It should be noted that Haplotag first generates a set of tag-level haplotypes ('HTgenos') followed by a set of SNP data derived from these haplotypes ('HTSNPgenos'). These two types of data are technically redundant, and the choice between them depends on the application and choice of software. We determined that most (97.5%) of haplotypes in this work contained only a single SNP, and we chose to analyze the SNP data set for simplicity and for compatibility with downstream analysis software.

A filtered SNP file was generated by the Character by Taxa (CbyT) program supplied by N. Tinker. Briefly, three separate 'HTSNPGenos' files generated by Haplotag were concatenated prior to running CbyT. The CbyT "minimum presence" value was set to 80%, 70%, 60%, and 50% for the associated 20%, 30%, 40%, and 50% missing data, respectively. The output files contained a SNP-by-sample matrix used in further analyses. An additional description of the SNP data matrix, along with the custom Perl and Shell scripts, is available in the Supplementary Materials. Note that the analyses from the FASTQ file separation to SNP generation were conducted using Microsoft Windows 7 64-bit OS with an Intel Core i7-3930K CPU @ 3.20 GHz (12 threads) and 32 GB RAM.

2.4. Genetic Diversity Analysis

SNP data with less than 50% missing values was used for diversity analysis, and they consisted of 119 samples from 10 accessions \times 5659 SNP markers. This data was first analyzed for the minor allele

frequency and the extent of the missing SNP data in a Microsoft Excel® file, followed by the separate diversity analyses at the individual and accession levels.

At the individual level, three types of diversity analyses were performed. First, the genetic structure was inferred using a model-based Bayesian method available in the program STRUCTURE version 2.2.3 [40,41]. The STRUCTURE program was run 30 times with 60-core parallel computing in a Linux server for each subpopulation (K) value, ranging from 1–10, using the admixture model with 10,000 replicates for burn-in and 10,000 replicates during analysis. The final population subgroups were determined based on (1) the likelihood plot of these models, (2) the change in the second derivative (ΔK) of the relationship between K and the log-likelihood [42], and (3) the stability of grouping patterns across 30 runs. For a given K, one of the 30 runs with the highest likelihood value was selected to assign the posterior membership coefficients to each sample. A graphical bar plot was then generated with the posterior membership coefficients. The size and composition of each optimal cluster with respect to population were analyzed. Second, a neighbor-joining (NJ) analysis of all 119 samples was conducted using PAUP* [43] based on the original data of the 5659 markers, and a radiation tree was displayed using MEGA 5.05 [44]. Third, a principal coordinates analysis (PCoA) of all 119 samples was also performed using the R routine *AveDissR* for assessing genetic distinctness and redundancy [45,46], and plots of the first three resulting principal components were generated to assess the sample associations. The resulting PCoA plots and NJ trees were individually labeled for the inferred structures for comparison.

For the accession level, an analysis of molecular variance (AMOVA) was performed with Arlequin version 3.01 [47] based on the 5659 markers to quantify the genetic variation present among the 10 accessions and to generate the pairwise genetic distances among these accessions and accession-specific F_{st} values for each accession. Additional AMOVA was also made to quantify the genetic variation between the cultivated and wild accessions and among the clusters identified from the STRUCTURE analysis. The pairwise genetic distances among the 10 accessions were used to generate a Unweighted Pair Group Method with Arithmetic Mean (UPGMA) dendrogram using MEGA 5.05 to assess accession differentiation and distinctness.

2.5. Flow Cytometry Analysis

The genome size and ploidy of NWG samples were determined by using CyFlow® Ploidy Analyzer manufactured by Sysmex Partec (Görlitz, Germany). This analysis is based on relative fluorescence intensity and requires one or two references to anticipate the ploidy and roughly interpret the genome size by comparing their average peak channel values. For this analysis, the tetraploid wheat (*T. durum*, $2n = 4x = 28$) and hexaploid wheat (*T. aestivum*, $2n = 6x = 42$) were chosen as the internal references, mainly because previous studies indicated that the majority of NWG plants are tetraploid, including the cultivar ‘Elbee’ [17,18], and the same family could give rise to the closest value and lowest bias. The genome sizes for durum and bread wheat were obtained from the Kew Botanical Garden website. The published protocol to identify the NWG ploidy [48] was applied.

Three plants grown in the greenhouse were randomly selected from each accession and wheat samples, and their fresh young leaves ($\sim 1 \text{ cm}^2$) were randomly collected for nuclei extraction, as the fresh young leaves consist predominantly of cells in the G1 phase. In contrast, cells with S-phase nuclei are rare during cell cycling. A total of 12 NWG accessions and internal references were chopped up, and intact nuclei suspensions were prepared by using Partec buffer supplemented with DNA-staining fluorochrome DAPI (4',6-diamidino-2-phenylindole). The delivery rate was adjusted at between 20–50 nuclei/s on flow cytometry. A threshold to filter out undesirable background from cell debris or the auto fluorescent compound was set by keeping the gain within the manufacturer recommended range. A total of 5000 particles were measured for both references and accessions to assess the analysis quality. The coefficient variation rate ($CV = SD/\text{mean channel value} \times 100$) was kept lower than 5%. The *T. durum* or *T. aestivum* reference G1 peak was positioned on approximately one-fifth of the linear abscissa (which was set as 200 as the control and reported as arbitrary units, A.U.), and then,

all the other samples' G1 peak positions were measured on the abscissa. Values for an accession were obtained by averaging the A.U. from three plants per accession. The genome size for each sample was estimated using the following formula:

$$\text{Sample 1C value} = \text{Reference 1C value} \times \frac{\text{sample 2C mean peak position}}{\text{reference 2C mean peak position}},$$

where 1C value represents the amount, in pictograms (or Mbp), of DNA contained within a haploid nucleus or one half the amount in a diploid cell of an eukaryotic organism. The estimated genome sizes were used to confirm presumed tetraploidy if the estimated genome sizes did not deviate significantly among all the assayed samples. The sample from TMP24008 with an estimated genome size of 17.1 Gb was deemed not to be tetraploid, and the accession was removed from the bioinformatics and diversity analyses, while the other assayed samples were confirmed as being tetraploid.

3. Results

Three MiSeq runs, each with 48 multiplexed samples, generated 23.3, 20.6, and 22.1 million raw forward (R1) sequence reads of 250 bp, respectively, for 144 samples, totaling 66 million sequences. Excluding the samples for the two accessions (TMP24017 and TMP24008) and one failed sample from TMP24016, approximately 56.1 million raw forward sequence reads were obtained for 119 samples of the 10 accessions for further bioinformatics analysis. The number of raw forward sequence reads per sample ranged from 274,794 to 728,999 with an average of 466,126 over all three MiSeq runs. The combined UNEAK and Haplotag analysis generated 213, 1009, and 5659 SNPs at the 30%, 40%, and 50% levels of missing data across 119 samples, respectively. The analysis also produced many meta genomic files associated with the SNP discovery, and these files are described and accessible in the Supplementary Materials. Assessing the distribution of minor allele frequency for the data set of the 5659 SNP markers revealed a wide range of minor allele frequencies from 0.05 to 0.50 with the average of 0.184 and displayed a gradual reduction of minor alleles with increased occurrence frequencies from 0.125 to 0.5 (Figure 1A). Also, increased SNPs were associated with increased levels of missing SNP data (Figure 1B). For example, 80 SNPs and 2017 SNPs had roughly 30% and 50% levels of missing data across 119 samples, respectively.

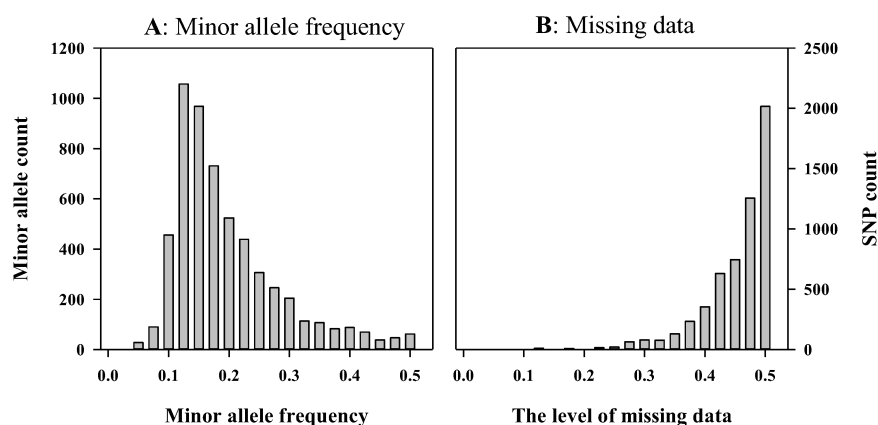


Figure 1. The minor allele frequency distribution (A) and the extent of missing data (B) for 5659 SNP markers in 119 samples of the 10 northern wheatgrass accessions.

The STRUCTURE analysis did not consider prior accession information and revealed three optimal clusters (Figure 2A) with supports from the change in $\text{LnP}(k)$ variance (Figure 2B) and the largest absolute value of the second-order rate of change of the likelihood distribution $|\text{Ln}''(K)|$ (Figure 2C). Cluster 1, highlighted in red, consisted of 24 samples (12 from WA accession, 11 from IR

accession, and one from WY accession). Cluster 2, highlighted in green, was the largest group, which consisted of 72 samples from seven accessions. Cluster 3, highlighted in blue, had 23 samples (12 from NV accession, 10 from WY accession and one from IR accession). The neighbor-joining (NJ) clustering revealed close genetic relationships of 119 samples from the 10 accessions (Figure 3). It appeared to have five distinguishable groups, but these groups were genetically extremely close. The NJ clustering largely matched with the Bayesian inferences of genetic structure (Figure 2). However, a few mismatches also existed from the Bayesian inferences. For example, some members of Cluster 1, highlighted in red, such as IR.10 and IR.11 were spread into Cluster 2, highlighted in green (Figure 3).

The PCoA plot revealed the genetic relationships of the 119 samples that were not in support of the Bayesian inferences from the STRUCTURE analysis (Figure 4A). Three Bayesian clusters were not distinguishable and were overlapping over the PCoA plot. However, the PCoA plot clearly showed the distinctness of the samples from the NV and WY accessions from the rest of the samples (Figure 4B). Also, the samples from three accessions (NV, WY, and IR) displayed a wider spread over the plot than the rest, showing the larger genetic diversity present in these accessions.

Partitioning the genetic variation of the 119 samples of the 10 accessions was made through AMOVA. It was found that 8.8% SNP variance resided among these 10 accessions, and 91.2% variance was present within the accessions (Table 2). Assessing the accession-specific F_{st} values showed the range of 0.0832 (AB accession) to 0.0983 (WY accession) with a mean of 0.0877, as illustrated in Figure 5B. The pairwise genetic distances among the 10 accessions ranged from 0.0264 (between the AB and MT1 accessions) to 0.1519 (between the NV and CO accessions) with a mean of 0.0871. Further AMOVAs showed that 1.7% SNP variation resided between the cultivated and wild accessions, and 7.9% variance was present among the three Bayesian clusters (Table 2).

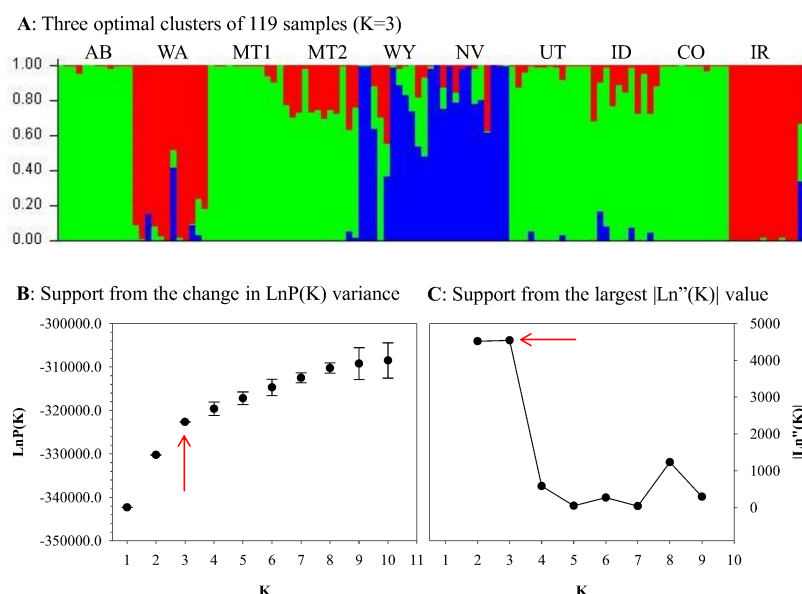


Figure 2. Three genetic clusters of the 119 northern wheatgrass samples inferred by the STRUCTURE analysis based on the 5659 SNP markers. (A) The mixture coefficients of 119 samples with $K = 3$, presented in the original order of samples from the 10 accessions (see Table 1 for accession label; the first four accessions from the left are for cultivated accessions and the rest are for wild accessions). (B) Support from the $\text{LnP}(K)$ estimation, highlighted by the red arrow. (C) Support from the estimation of the largest absolute value of the second-order rate of change of the likelihood distribution $|\text{Ln}''(K)|$, highlighted by the red arrow.

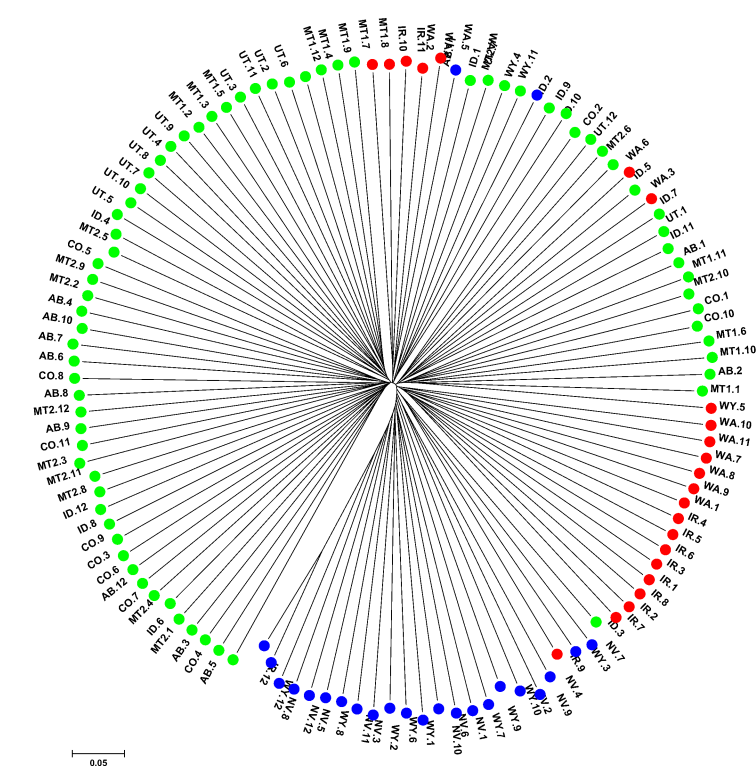


Figure 3. Genetic relationships of the 119 samples of the 10 northern wheatgrass accessions as revealed by neighbor-joining clustering with the 5659 SNP markers. Each sample is numbered after its accession label (see Table 1) with a dot. Each node for a sample is represented by a colored circle. Red, green, and blue represent Clusters 1, 2 and 3 inferred from the STRUCTURE analysis (see Figure 2A), respectively.

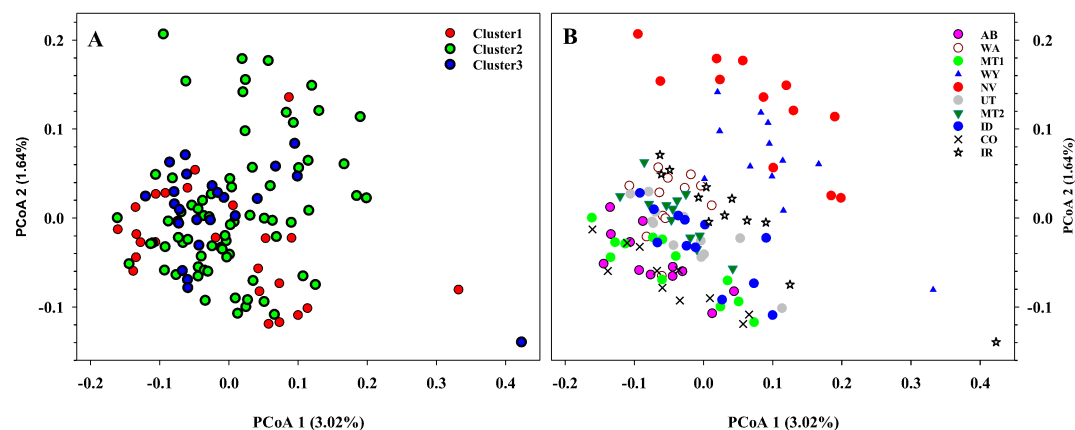


Figure 4. Genetic relationships of the 119 samples of the 10 northern wheatgrass accessions revealed by principal coordinates analysis (PCoA) with the 5659 SNP markers. Two panels are the same, but the left panel (A) labels samples for clusters obtained from the STRUCTURE analysis, while the right panel (B) labels samples for the 10 accessions (see Table 1 for accession label). Note accessions AB, WA, MT1, and MT2 are cultivated accessions, while the rest are wild accessions.

Table 2. Results of the analysis of molecular variance for three models of genetic structure (10 accessions, cultivated vs wild accessions, and the three clusters from the STRUCTURE analysis) based on the 5659 SNP markers.

Model/Source of Variation	df	Sum of Squares	Variance Explained	Variance (%) ^a
<i>10 accessions</i>				
Among accessions	9	6854.6	22.4	8.77
Within accessions	226	52,675.3	233.1	91.24
<i>Cultivated vs. wild accessions</i>				
Between groups	1	742.6	4.3	1.69
Within groups	234	58,787.3	251.2	98.31
<i>Three clusters from STRUCTURE</i>				
Among clusters	2	3176.5	20.7	7.88
Within clusters	233	56,353.5	241.1	92.12

^a These variances were statistically significant from zero at $p < 0.0001$.

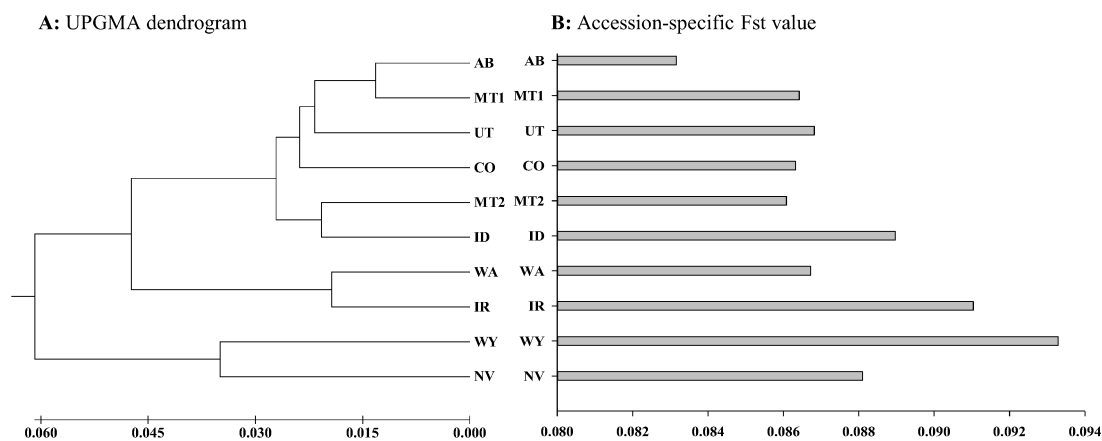


Figure 5. Genetic diversity and genetic relationships of the 10 northern wheatgrass accessions. The left panel (A) displays their genetic relationships in a UPGMA dendrogram based on the Phi statistics obtained from the AMOVA. The right panel (B) shows the accession-specific Fst values for the 10 accessions. The accession labels are given in Table 1.

An UPGMA dendrogram based on the AMOVA results revealed three genetically distinct groups of accessions at the Phi statistic of 0.035 or larger (Figure 5A). The most distinct group was from the NV and WY accessions, as supported in the PCoA plot (Figure 4). The second distinct group consisted of samples from the WA and IR accessions. Interestingly, the accession relationships in the dendrogram matched well with the accession-specific Fst values (Figure 5B).

Efforts were also made to estimate the genome size of four northern wheatgrass plants representing four accessions (Table 1). Using the genome size of *T. aestivum* as a reference, the estimates of the genome size ranged from 9.195 to 9.574 Gb with an average of 9.428 Gb. Similarly, the genome size estimates ranged from 8.841 to 9.478 Gb with a mean of 9.253 Gb when durum wheat was used as a genome size reference. These genome size estimates confirmed the presumed tetraploidy of the assayed samples.

4. Discussion

The study presented here represents the first preliminary effort through GBS to capture genome-wide genetic variants of northern wheatgrass plants and to assess the genetic diversity

present in this species. The effort generated the first novel set of genomic resources and 5659 SNP markers for diversity analysis of this tetraploid species. The diversity analysis revealed 8.8% of SNP variation residing among the 10 accessions and 1.9% SNP variation was present between the cultivated and wild accessions. The Bayesian analysis identified three major clusters of the assayed samples, and the principal coordinates analysis revealed the genetic distinctness of the two accessions collected from Nevada and Wyoming. The flow cytometry analysis confirmed the tetraploidy of the assayed plants and estimated an average genome size of 9.3–9.4 Gb for this species. These findings are useful for the genetic improvement of this native grass species for forage production and rangeland reclamation. They also suggest that there can be a broad GBS application in the characterization of genome-wide genetic variants in non-model polyploidy plants.

The extent of the genetic diversity present among and within assayed accessions (Table 2) are not surprising given that northern wheatgrass is a highly outcrossing species [49]. These results are in general agreement with those findings reported in similar studies of other native grass species using amplified fragment length polymorphism (AFLP) markers (e.g., see [50,51]) and other markers (e.g., [52]). An interesting finding is the genetic distinctness of the accessions collected from Nevada and Wyoming, as these plants were collected from the middle range of the species distribution [19]. Further studies are needed to assay samples representing the whole species distribution for a better understanding of the diversity distribution and its associations with evolutionary forces in natural populations. Also, a comparative phenotypic study of these accessions under the same growth environment may yield useful information on their adaptive variability in photoperiod sensitivity and flowering time to verify the observed genetic distinctness.

It was evident that the cultivated samples had lower diversity than those collected from the natural stands (Figures 4 and 5). For example, the samples representing the wild accessions displayed a wider spread on the PCoA plot than those of the cultivated accessions (see AB vs. NV accessions in Figure 4B). This finding is expected, given that the selection acted for biomass or high seed yield on the cultivated accessions, such as ‘Elbee’, the first variety of northern wheatgrass released in Canada in 1980 [18]. However, the genetic differentiation between the cultivated and wild accessions was relatively small, accounting for only 1.7% SNP variation (Table 2). This finding is encouraging for NWG breeding as ‘Elbee’, a synthetic of only eight clones, did not show much genetic shift, while still maintaining high genetic diversity. Another interesting result is the genetic association of the cultivated accessions with the wild accessions (Figure 5A). The cultivar ‘Schwendimar’ was developed from a wild stand of northern wheatgrass collected from Oregon, United States, but it was genetically closer to the wild accession collected from Iran in 1972. Similarly, the cultivated accession from Montana was closely related to the wild accession from Idaho. These findings, particularly those revealing genetic relatedness, are useful for northern wheatgrass breeding and genetic analysis. In addition, the finding of high within-accession genetic variation (91%; Table 2) implies that the seed collection from a single site may provide an adequate genetic diversity for creating an improved germplasm used for nearby site reclamation. This differs from the previous recommendation of using multi-site composite materials to maintain high levels of genetic diversity at new restoration sites [53].

The GBS application presented here is encouraging for sampling genome-wide variants to assess genetic diversity in non-model polyploid species. The archived FASTQ data in NCBI SRA and the generated meta genomic files in the Supplementary Materials represent the first set of genomic resources ever obtained for NWG plants, and they can be used for genomic and genetic investigations of this and other grass species. Analyzing these NWG genomic data revealed thousands of genome-wide SNP markers for its genetic diversity analysis. Thus, it is technically possible to sample genome-wide genetic variability in plants with complex genomes through GBS, even though SNP calling from sequences of a complex, polyploid genome is confounded by the presence of homeologs [54], orthologs, and paralogs [55]. Thousands of SNP markers would be more informative for plant genetic diversity analysis with higher resolution than the several hundreds of AFLP markers normally screened in previous studies [50,51,56,57]. More importantly, the total laboratory and sequencing cost for

generating these genomic data on this set of studied samples was roughly \$12,000, indicating the feasibility of a wider application of GBS to characterize native grass species. However, bias in SNP calling exists due to incomplete allele sampling [10]. It is difficult to separate true null tag-level haplotypes from missing sequence data, particularly at the sequence depth typically employed in GBS studies [12]. Improving Haplotag with the flexibility to relax the restraints on the sequenced fragment length and library barcoding is still needed [11].

5. Conclusions

The first preliminary GBS application has generated a novel set of genomic resources and 5659 SNP markers for the diversity analysis of northern wheatgrass. It was found that 8.8% of SNP variation resided among the 10 accessions and 1.9% SNP variation was present between the cultivated and wild accessions. Further analyses revealed three genetic clusters of the assayed samples and the genetic distinctness of the plants collected from Nevada and Wyoming. The estimated genome size for this species ranged from 8.8 to 9.6 Gb and averaged 9.3 Gb. These findings are useful for northern wheatgrass breeding and are also encouraging for broad GBS application in the characterization of genome-wide genetic variability in non-model polyploidy plants.

Supplementary Materials: The supplementary materials are available online at <http://www.mdpi.com/1424-2818/10/2/23/s1>.

Acknowledgments: The authors would like to thank the following: Carolee Horbach for her technical assistance in the research; Tina Bundrock for her assistance in the flow cytometry analysis; Isobel Parkin for the access to and the use of the Illumina MiSeq instrument; Nick Tinker for his assistance on the Haplotag analysis; and the two anonymous journal reviewers for their helpful comments on the early version of this manuscript. This work was financially supported by the Beef Cattle Research Council of Canada and Agriculture and Agri-Food Canada GF2 Funds.

Author Contributions: Y.-B.F., B.B., B.C. and M.P.S. conceived of the project. Y.-B.F. designed the research. B.C. prepared the study materials. G.W.P. conducted the sequencing. P.L., S.B., G.W.P. and Y.-B.F. performed data analysis. P.L., S.B. and Y.-B.F. wrote the manuscript. Y.-B.F., B.B., B.C. and M.P.S. revised the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Ethical Standards: The experiment complies with the current laws of Canada where it was performed.

References

1. Huang, X.; Feng, Q.; Qian, Q.; Zhao, Q.; Wang, L.; Wang, A.; Guan, J.; Fan, D.; Weng, Q.; Huang, T.; et al. High throughput genotyping by whole-genome resequencing. *Genome Res.* **2009**, *19*, 1068–1076. [[CrossRef](#)] [[PubMed](#)]
2. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **2011**, *6*, e19379. [[CrossRef](#)] [[PubMed](#)]
3. Fu, Y.B.; Peterson, G.W. Genetic diversity analysis with 454 pyrosequencing and genomic reduction confirmed the eastern and western division in the cultivated barley gene pool. *Plant Genome* **2011**, *4*, 226–237. [[CrossRef](#)]
4. Peterson, B.; Weber, J.N.; Kay, E.H.; Fisher, H.S.; Hoekstra, H.E. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS ONE* **2012**, *7*, e37135. [[CrossRef](#)] [[PubMed](#)]
5. Peterson, G.W.; Dong, Y.B.; Horbach, C.; Fu, Y.B. Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity* **2014**, *6*, 665–680. [[CrossRef](#)]
6. Poland, J.A.; Rife, T.W. Genotyping-by-sequencing for plant breeding and genetics. *Plant Genome* **2012**, *5*, 92–102. [[CrossRef](#)]
7. Poland, J.A.; Brown, P.J.; Sorrells, M.E.; Jannink, J.L. Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE* **2012**, *7*, e32253. [[CrossRef](#)] [[PubMed](#)]

8. Huang, Y.F.; Poland, J.A.; Wight, C.P.; Jackson, E.W.; Tinker, N.A. Using genotyping-by-sequencing (GBS) for genomic discovery in cultivated oat. *PLoS ONE* **2014**, *9*, e102448. [[CrossRef](#)] [[PubMed](#)]
9. Fu, Y.B.; Peterson, G.W.; Dong, Y. Increasing genome sampling and improving SNP genotyping for genotyping-by-sequencing with new combinations of restriction enzymes. *G3* **2016**, *6*, 845–856. [[CrossRef](#)] [[PubMed](#)]
10. Fu, Y.B.; Yang, M.H. Genotyping-by-sequencing and its application to oat genomic research. In *Oat—Methods and Protocols*; Gasparis, S., Ed.; Springer Science + Business Media: New York, NY, USA, 2017; pp. 169–187.
11. Tinker, N.A.; Bekele, W.A.; Hattori, J. Haplotag: Software for haplotype-based genotyping-by-sequencing analysis. *G3* **2016**, *6*, 857–863. [[CrossRef](#)] [[PubMed](#)]
12. Yan, H.; Bekele, W.A.; Wight, C.P.; Peng, Y.; Langdon, T.; Latta, R.G.; Fu, Y.B.; Diederichsen, A.; Howarth, C.J.; Jellen, E.N.; et al. High-density markers profiling confirms ancestral genomes of *Avena* species and identifies D-genome chromosomes of hexaploid oat. *Theor. Appl. Genet.* **2016**, *129*, 2133–2149. [[CrossRef](#)] [[PubMed](#)]
13. Bekele, W.A.; Wight, C.P.; Chao, S.; Howarth, C.J.; Tinker, N.A. Haplotype based genotyping-by-sequencing in oat genome research. *Plant Biotechnol. J.* **2018**. [[CrossRef](#)] [[PubMed](#)]
14. Dewey, D.R. The genomic system of classification as a guide to intergeneric hybridization with the perennial Triticeae. In *Gene Manipulation in Plant Improvement*; Gustafson, J.P., Ed.; Columbia University Press: New York, NY, USA, 1984; pp. 209–279.
15. Love, A. Conspectus of the Triticeae. *Feddes Repert.* **1984**, *95*, 425–521. [[CrossRef](#)]
16. Jensen, K.B. Cytology and fertility of advanced populations of *Elymus lanceolatus* (Scribn. & Smith) Gould × *Elymus caninus* (L.) hybrids. *Crop Sci.* **2005**, *45*, 1211–1215.
17. Dewey, D.R. Cytogenetics of *Agropyron pringlei* and its hybrids with *A. spicatum*, *A. seribneri*, *A. vidaceum*, and *A. dasystachyum*. *Bot. Gazette* **1976**, *137*, 179–185. [[CrossRef](#)]
18. Smoliak, S.; Johnston, A. Elbee northern wheatgrass. *Can. J. Plant Sci.* **1980**, *60*, 1473–1475. [[CrossRef](#)]
19. Cronquist, A.; Holmgren, A.H.; Holmgren, N.H.; Reveal, J.L.; Holmgren, P.K. *Intermountain Flora: Vascular Plants of the Intermountain West, U.S.A. Vol. 6: The Monocotyledons*; Columbia University Press: New York, NY, USA, 1977; p. 584.
20. Kowalenko, B.L.; Romo, J.T. Regrowth and rest requirements of northern wheatgrass following defoliation. *J. Range Manag.* **1998**, *51*, 73–78. [[CrossRef](#)]
21. Harniss, R.O.; Murray, R.B. 30 years of vegetal change following burning of sagebrush-grass range. *J. Range Manag.* **1973**, *26*, 322–325. [[CrossRef](#)]
22. Wright, H.A.; Bailey, A.W. *Fire Ecology: United States and Southern Canada*; John Wiley and Sons: New York, NY, USA, 1982; p. 501.
23. Knapp, E.E.; Rice, K.J. Genetic structure and gene flow in *Elymus glaucus* (blue wildrye): Implications for native grassland and management. *Restor. Ecol.* **1996**, *4*, 40–45. [[CrossRef](#)]
24. Rogers, D.L.; Montalvo, A.M. *Genetically Appropriate Choices for Plant Materials to Maintain Biological Diversity*; Report to the USDA Forest Service, Rocky Mountain Region, CO, USA; University of California: Davis, CA, USA, 2004; p. 343.
25. Diaz, O.; Salomon, B.; Bothmer, R.V. Genetic diversity and structure in populations of *Elymus caninus* (L.) L. (Poaceae). *Hereditas* **1999**, *131*, 63–74. [[CrossRef](#)]
26. Sun, G.L.; Diaz, O.; Salomon, B.; Bothmer, R.V. Genetic diversity in *Elymus caninus* as revealed by isozyme, RAPD and microsatellite markers. *Genome* **1999**, *42*, 420–431. [[CrossRef](#)] [[PubMed](#)]
27. Sun, G.L.; Diaz, O.; Salomon, B.; Bothmer, R.V. Microsatellite variation and its comparison with isozyme and RAPD variation in *Elymus fibrosus*. *Hereditas* **1999**, *129*, 275–282. [[CrossRef](#)]
28. Diaz, O.; Sun, G.L.; Salomon, B.; Bothmer, R.V. Level and distribution of allozyme and RAPD variation in populations of *Elymus fibrosus* (Poaceae). *Genet. Resour. Crop Evol.* **2000**, *47*, 11–24. [[CrossRef](#)]
29. Wu, D.C.; He, D.M.; Gu, H.L.; Wu, P.P.; Yi, X.; Wang, W.J.; Shi, H.F.; Wu, D.X.; Sun, G. Origin and evolution of allopolyploid wheatgrass *Elymus fibrosus* (Schrenk) Tzvelev (Poaceae: Triticeae) reveals the effect of its origination on genetic diversity. *PLoS ONE* **2016**, *11*, e0167795. [[CrossRef](#)] [[PubMed](#)]
30. Diaz, O.; Salomon, B.; Bothmer, R.V. Genetic variation and differentiation in Nordic populations of *Elymus alaskanus* (Scrib. ex Merr.) LoÈve (Poaceae). *Theor. Appl. Genet.* **1999**, *99*, 210–217.
31. Sun, G.L.; Salomon, B.; Bothmer, R.V. Microsatellite polymorphism and genetic differentiation in three Norwegian populations of *Elymus alaskanus* (Poaceae). *Plant Syst. Evol.* **2002**, *234*, 101–110. [[CrossRef](#)]

32. Zhang, X.Q.; Salomon, B.; Bothmer, R.V. Application of random amplified polymorphic DNA markers to evaluate intraspecific genetic variation in the *Elymus alaskanus* complex (Poaceae). *Genet. Resour. Crop Evol.* **2002**, *49*, 397–407. [[CrossRef](#)]
33. Sun, G.L.; Salomon, B. Microsatellite variability and the heterozygote deficiency in arctic-alpine species *Elymus alaskanus* complex. *Genome* **2003**, *46*, 729–737. [[CrossRef](#)] [[PubMed](#)]
34. Gaudett, M.; Salomon, B.; Sun, G.L. Molecular variation and population structure in *Elymus trachycaulus* and comparison with its morphologically similar *E. alaskanus*. *Plant Syst. Evol.* **2005**, *250*, 81–91. [[CrossRef](#)]
35. Sun, G.L.; Li, W.B. Molecular diversity of *Elymus trachycaulus* complex species and their relationships to non-North American taxa. *Plant Syst. Evol.* **2006**, *256*, 179–191. [[CrossRef](#)]
36. Ma, X.; Zhang, X.Q.; Zhou, Y.H.; Bai, S.Q.; Liu, W. Assessing genetic diversity of *Elymus sibiricus* (Poaceae:Triticeae) populations from Qinghai-Tibet Plateau by ISSR markers. *Biochem. Syst. Ecol.* **2008**, *36*, 514–522. [[CrossRef](#)]
37. Ma, X.; Chen, S.Y.; Bai, S.Q.; Zhang, X.Q.; Li, D.X.; Zhang, C.B.; Yan, J.J. RAPD analysis of genetic diversity and population structure of *Elymus sibiricus* (Poaceae) native to the southeastern Qinghai-Tibet Plateau, China. *Genet. Mol. Res.* **2012**, *11*, 2708–2718. [[CrossRef](#)] [[PubMed](#)]
38. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
39. Lu, F.; Lipka, A.E.; Glaubitz, J.; Elshire, R.; Cherney, J.H.; Casler, M.D.; Buckler, E.S.; Costich, D.E. Switchgrass genomic diversity, ploidy, and evolution: Novel insights from a network-based SNP discovery protocol. *PLoS Genet.* **2013**, *9*, e1003215. [[CrossRef](#)] [[PubMed](#)]
40. Pritchard, J.; Stephens, M.; Donnelly, P. Influence of population structure using multilocus genotype data. *Genetics* **2000**, *155*, 945–959. [[PubMed](#)]
41. Falush, D.; Stephens, M.; Pritchard, J.K. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol. Ecol. Notes* **2007**, *7*, 574–578. [[CrossRef](#)] [[PubMed](#)]
42. Evanno, G.; Regnaut, S.; Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **2005**, *14*, 2611–2620. [[CrossRef](#)] [[PubMed](#)]
43. Swofford, D.L. *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods)*, Version 4; Sinauer Associates: Sunderland, MA, USA, 1998.
44. Tamura, K.; Peterson, D.; Peterson, N.; Stecher, G.; Nei, M.; Kumar, S. MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **2011**, *28*, 2731–2739. [[CrossRef](#)] [[PubMed](#)]
45. Yang, M.H.; Fu, Y.B. AveDissR: An R function for assessing genetic distinctness and genetic redundancy. *Appl. Plant Sci.* **2017**, *5*, 1700018. [[CrossRef](#)] [[PubMed](#)]
46. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016; ISBN 3-900051-07-0. Available online: <http://www.r-project.org/> (accessed on 10 April 2018).
47. Excoffier, L.; Laval, G.; Schneider, S. Arlequin ver. 3.1: An integrated software package for population genetics data analysis. *Evol. Bioinform. Online* **2005**, *1*, 47–50. [[CrossRef](#)]
48. Bennett, R.A.; Séguin-Swartz, G.; Rahman, H. Broadening genetic diversity in Canola using the C-Genome species *Brassica oleracea* L. *Crop Sci.* **2012**, *52*, 2030–2039. [[CrossRef](#)]
49. Hamrick, J.L.; Godt, M.J.W. Allozyme diversity in plant species. In *Plant Population Genetics, Breeding and Genetic Resources*; Brown, A.H.D., Clegg, M.T., Kahler, A.L., Weir, B.S., Eds.; Sinauer Associates: Sunderland, MA, USA, 1989; pp. 43–63.
50. Fu, Y.B.; Coulman, B.E.; Fernandez, Y.S.N.; Cayouette, J.; Peterson, P.M. Genetic diversity of fringed brome (*Bromus ciliatus*) as determined by amplified fragment length polymorphism. *Can. J. Bot.* **2005**, *83*, 1322–1328. [[CrossRef](#)]
51. Biliget, B.; Schellenberg, M.P.; Fu, Y.B. Detecting genetic diversity of side-oats grama grass populations using AFLP Marker. *Can. J. Plant Sci.* **2013**, *93*, 1105–1114. [[CrossRef](#)]
52. Sun, G.L.; Salomon, B.; Bothmer, R.V. Analysis of tetraploid *Elymus* species using wheat microsatellite markers and RAPD markers. *Genome* **1997**, *40*, 806–814. [[CrossRef](#)] [[PubMed](#)]
53. Phan, A.T.; Fu, Y.B.; Smith, S.R., Jr. RAPD variations in selected and unselected blue grama populations. *Crop Sci.* **2003**, *43*, 1852–1857. [[CrossRef](#)]

54. Glover, N.A.; Redestig, H.; Dessimoz, C. Homeologs: What are they and how do we infer them? *Trends Plant Sci.* **2016**, *21*, 609–621. [[CrossRef](#)] [[PubMed](#)]
55. Nielsen, R.; Paul, J.S.; Albrechtsen, A.; Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **2011**, *12*, 443–451. [[CrossRef](#)] [[PubMed](#)]
56. Vos, P.; Hogers, R.; Bleeker, M.; Reijans, M.; van De Lee, T.; Hornes, M.; Frijters, A.; Pot, J.; Peleman, J.; Kuiper, M.; et al. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Res.* **1995**, *23*, 4407–4414. [[CrossRef](#)] [[PubMed](#)]
57. Fu, Y.B.; Phan, A.T.; Coulman, B.; Richards, K.W. Genetic diversity in natural populations and corresponding seed collections of little bluestem as revealed by AFLP markers. *Crop Sci.* **2004**, *44*, 2254–2260. [[CrossRef](#)]



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).