diversity

MDPI

Article

Sequence Diversity and Identification of Novel Puroindoline and Grain Softness Protein Alleles in *Elymus*, *Agropyron* and Related Species

Mark D. Wilkinson [1,*], Robert King [2] and Roberta Grimaldi [3]

1 Plant Science Department, Rothamsted Research, Harpenden AL5 2JQ, UK
2 Computational and Analytical Sciences, Rothamsted Research, Harpenden AL5 2JQ, UK; robert.king@rothamsted.ac.uk
3 Department of Food and Nutritional Sciences, University of Reading, Reading RG6 6AP, UK; Roberta.grimaldi@reading.ac.uk
* Correspondence: mark.wilkinson@rothamsted.ac.uk; Tel.: +44-1582-938887

Received: 12 September 2018; Accepted: 11 October 2018; Published: 18 October 2018

Abstract: The puroindoline proteins, PINA and PINB, which are encoded by the *Pina* and *Pinb* genes located at the Ha locus on chromosome 5D of bread wheat, are considered to be the most important determinants of grain hardness. However, the recent identification of *Pinb-2* genes on group 7 chromosomes has stressed the importance of considering the effects of related genes and proteins. Several species related to wheat (two diploid *Agropyron* spp., four tetraploid *Elymus* spp. and five hexaploid *Elymus* and *Agropyron* spp.) were therefore analyzed to identify novel variation in *Pina*, *Pinb* and *Pinb-2* genes which could be exploited for the improvement of cultivated wheat. A novel sequence for the *Pina* gene was detected in *Elymus burchan-buddae*, *Elymus dahuricus* subsp. *excelsus* and *Elymus nutans* and novel PINB sequences in *Elymus burchan-buddae*, *Elymus dahuricus* subsp. *excelsus*, and *Elymus nutans*. A novel PINB-2 variant was also detected in *Agropyron repens* and *Elymus repens*. The encoded proteins detected all showed changes in the tryptophan-rich domain as well as changes in and/or deletions of basic and hydrophobic residues. In addition, two new AGP sequences were identified in *Elymus nutans* and *Elymus wawawaiensis*. The data presented therefore highlight the sequence diversity in this important gene family and the potential to exploit this diversity to modify grain texture and end-use quality in wheat.

Keywords: *Elymus species*; Puroindoline genes; *Pinb-2* genes; novel sequence variants

1. Introduction

Grain texture has major effects on the end-use quality of bread wheat (*T. aestivum* L.) as it influences flour's particle size distribution, starch damage, milling/air-classification properties and water absorptivity [1]. Flour from hard wheat is used for making bread whilst the flour from soft wheat is preferred to produce biscuits, cakes and pastries [2]. The major determinants of texture are the puroindoline proteins (Pins), PINA and PINB, which are encoded by *Pin* genes at the *Ha* (hardness) locus on chromosome 5D [3,4] and account for about 60–80% of the variation in hardness in crosses between bread wheat cultivars [5,6]. These two proteins are also the major components of "friabilin", a mixture of proteins present on the surface of starch granules prepared through the aqueous washing of flour from soft wheat but not hard wheats [7]. PINs are being researched extensively due to their crucial and disruptive role in dictating grain hardness [8]. It has been suggested that friabilin/PINs influence grain texture by reducing adhesion between the surface of the starch granule and the surrounding protein matrix, resulting in the requirement of less energy for milling and less starch damage [9,10]. Recently [11] showed that the introduction of puroindoline genes into durum wheat

reduced milling energy and changed milling behavior, which was similar to soft common wheats. Equally, the silencing of the Pin genes led to an increase in grain hardness [12,13].

In addition to the *Pina* and *Pinb* genes, the *Ha* locus contains a third gene called *Grain Softness Protein-1* (*Gsp-1*) [4]. This gene encodes a protein which is post-translationally cleaved to give two products. A short 15-residue peptide corresponding to the protein N-terminus is glycosylated and forms the core of the arabinogalactan peptide (AGP). The second encoded protein, called "grain softness protein" (GSP), is related to the PIN proteins and was initially considered to contribute to the control of hardness [14]. Although this role was disputed by [15], we have recently shown that down-regulation of the expression of *Gsp-1* in transgenic plants resulted in a small but statistically significant increase in hardness [16]. It has also been shown that multiple genes related to *Gsp-1* also occur elsewhere in the wheat genome [14,17,18]. The *Gsp-1* genes now appear to be ubiquitous in grasses, being identified in 65 species from five major grass subfamilies, with over 20 different *Gsp-1* alleles being reported [19].

Genes encoding variant forms of PINB have been identified on the group 7 chromosomes (7A, 7B or 7D) of wheat [20–23]. Four such variants have been described, designated as *Pinb-2v1*, *Pinb-2v2*, *Pinb-2v3* and *Pinb-2v4* [20,24], with numerous single nucleotide polymorphisms (SNPs) for each sequence type [21,24–26]. The *Pinb-2* variants 2 and 3 have been shown to be allelic, segregating as a single bi-allelic locus, with segregation fitting a 1:2:1 ratio [23]. More recently, [23] proposed that the designations of *Pinb-2v2* and *Pinb-2v3* should be changed to *Pinb-B2a* and *Pinb-B2b*, respectively, while *Pinb-2v1 and Pinb-2v4* should be changed to *Pinb-D2a* and *Pinb-A2a*. Although the *Pinb-v2 genes* have been mapped to the group 7 chromosomes where they are associated with a minor quantitative trait (QTL) for hardness [20], a role in determining hardness has not been confirmed, with conflicting reports from [22,26,27].

Several studies have reported sequence diversity in *Pin* genes of wheat and related grass species, aimed at identifying novel variations which could affect their functional properties, and hence be exploited to increase the range of textural characteristics in wheat breeding programmes [4–6,17,28–31] We have also shown that species of the genera *Elymus* and *Agropyron* species are particularly rich sources of sequence in the *Gsp-1* gene, and particularly for the sequence encoding the AGP peptide sequence [19]. We have therefore extended this study to survey sequence diversity in *Pina*, *Pinb*, *Pinb-2v* and *Gsp-1* genes in a range of diploid, tetraploid and hexaploid species of the *Elymus* and *Agropyron* species as well as closely related genera (*Pseudorogneria*, *Psathyrostachys*, *Thinopyrum*). The wheatgrass (*Agropyron*, *Elymus*, *Thinopyrum*, *Pseudorogneria*) and wild ryegrass (*Elymus* and *Psathyrostachys*) species all share common genomic origins even though their taxonomic classification remains controversial. These genetic similarities between the genera *Thinopyrum*, *Agropyron*, *Pseudorogneria*, and hexaploid wheat have been clearly shown using the fluorescent in situ hybridization (FISH) technique [32] and have allowed for hybridization between these species (including barley, durum wheat and rye) to introgress novel variations in breeding programmes [33–35].

Our study has identified novel variations in the protein sequences PINA, PINB, PINBv-2 variant and *Gsp-1* genes in *Elymus*, *Agropyron* and closely related species, highlighting the diversity in this functionally important multigene family.

2. Materials and Methods

2.1. Source of Material

The following seed material was provided by GRIN, (Germplasm Resources Information Network), USA: Seed collections for *Elymus angulatus* (6x; Accession number PI 655176), *Elymus burchan-buddae* (4x; Accession number PI 639854), *Elymus dahuricus* subsp. *excelsus* (6x; Accession number PI 655189), *Elymus nutans* (6x; Accession number PI 639855), *Elymus sibiricus* (4x; Accession number PI 670360), *Elymus trachycaulus* subsp. *subsecundus* (4x; Accession number PI 232148), *Elymus wawawaiensis* (4x; Accession number PI 537313), *Agropyron mongolicum* (2x; Accession number

PI-531543; PI-598482), *Pseudoroegneria spicata* (2x; Accession number PI-563867) and *Psathyrostachys juncea* (2x; Accession number W6-25130) were maintained at the Western Regional PI Station (USDA, ARS, WRPIS), Washington State University, Regional Plant Introduction Station, 59 Johnson Hall, P.O. Box 646402 Pullman, Washington 99164-6402.

Seeds for *Agropyron cristatum* (2x; Accession number PI-429779), *Thinopyrum scripeum* (2x; Accession number C15-21-25, Italy), *Thinopyrum elongatum* (2x; Accession number PI-547326) and *Thinopyrum bessarabicum* (2x; Accession number PI-531712) were kindly provided by Professor Elizabeth Kellogg, University of Missouri-St. Louis, USA. Seeds for *Agropyron repens* (6x) and *Elymus repens* (6x) were obtained from Herbiseed Seed Company, Twyford, UK. Seed stocks of *Triticum aestivum* bread wheat (cv. Cadenza, Paragon and Falcon) were maintained at Rothamsted Research.

2.2. Germination of Seed Material

All the seeds were germinated on wet filter paper in 90 mm Petri dishes, sealed with Parafilm (Fisher Scientific Ltd, UK, Bishop Meadow Road, Loughborough, LE11 5RG) and incubated at 4 °C for 2 days before being transferred to room temperature.

2.3. Sequence Analysis

Genomic DNA was extracted from leaf material of seedlings using a Promega Wizard kit (Promega (UK) Ltd., Southampton, Hampshire, SO16 7NS UK).

Full-length Pina (447 bp) and Pinb (447 bp) were amplified with gene-specific primers: Pina Forward primer: 5′ ATGAAGGCCCTCTTCCTCA 3′ and Reverse primer: 5′ TCACCAGTAATAGCCAATAGTG 3′; Pinb- Forward primer: 5′ ATGAAGACCTTATTCCTCCTAGC 3′ and Reverse primer: 5′ TATAGATATCATCACCAGTAAT 3′. Pinb variant forms (450 bp) were amplified with generic primers: Forward primer PinbvF: 5′ ATGAAGACCTTATTCCTCCTAGCTC 3′ and Reverse primer PinbvR: 5′ TCAGTAGTAATAGCCATTAGTAGGGACG 3′. Full-length AGP/*Gsp-1* products (495 bp) were amplified with gene-specific primers: GSPF Forward primer: 5′ CATGAAGACCTTCTTCCTCC 3′ and GSPR Reverse primer: 5′ TCACAAGTAATATCCGCTAGTGA 3′. The reactions were performed in 25 µL using a 1.1X ReddyMix™PCR Master Mix (1.5 mM $MgCl_2$) from Thermo Scientific (ABgene House, Blenheim Road, Epsom, Surrey, KT19 9AP, UK), also containing 100–150 ng of genomic DNA and 0.8 µM of each primer. The cycling conditions were 96 °C for 5 min followed by 32 cycles of 96 °C for 30 s, 52 °C for 30 s, 72 °C for 1 min 30 s and the extension of 72 °C for 10 min for the *Pinbv* and *Gsp-1* gene PCR reactions. An annealing temperature of 55 °C was used for both the *Pina* and the *Pinb* gene screens, and all other conditions remained the same. 5 µL of the reactions was analyzed on 1.0% (*w*/*v*) agarose gels, stained with ethidium bromide and visualized using UV light.

PCR products were purified with the remainder of the PCR reaction using a Wizard® SV Gel and PCR Clean-Up System (Promega (UK) Ltd.). Fragments were then ligated into the pGem®-T Easy system (Promega (UK) Ltd.) and clones obtained using the Wizard® Plus SV Minipreps DNA purification system (Promega (UK) Ltd.). Sequencing reactions were performed with the BigDye Terminator Version 3.1 Cycle Sequencing Kit (Applied Biosystems, Life Technologies Ltd., 3 Fountain Drive, Inchinnan Business Park, Paisley PA4 9RF, UK) and all the reactions were analyzed at Source Bioscience (Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK). Annotations of sequences were carried out using Bioedit (PC) (Copyright Tom Hall @ 1997–2011).

2.4. Construction of Protein Models and Phylogenetic Trees

3D models of protein structures were generated using the Phyre2 web portal for protein modelling, prediction and analysis (Structural Bioinformatics Group, Imperial College London, London, UK). The InterPro database 66.0 software; website: https://www.ebi.ac.uk/interpro/ was used to provide a functional analysis of protein sequences by classifying them into families and predicting domains and important sites.

For the phylogenetic tree analysis nucleotide sequences for all the species screened, with known *Triticum aestivum* sequences for all genes, were aligned using MAFFT v.7.308, in Geneious 10.0.9. For the phylogenetic reconstruction, the K80 + G nucleotide substitution model was selected by AIC in jModeltest 2.1.10. The Maximum Likelihood phylogeny was reconstructed using MrBayes, with the substitution model selected in jModeltest; MrBayes v3.2.6 with the following settings: 100,000 generations, sampling every 100 generations with a burnin fraction of 0.25 [36–38].

3. Results

The sequences of *Pina*, *Pinb*, *Pinb-2v* and *Gsp-1* genes were determined for species of *Elymus* (eight species, eight accessions), *Agropyron* (three species, three accessions), *Thinopyrum* (three species, three accessions), *Psathyrostachys* (one species) *and Pseudorogneria* (one species). Individual alignments of protein sequence types are shown in Figures 1–3, while a phylogenetic tree for all sequences analyzed shown in Figure 4. The sequence results and genome relationships of these species are summarized in Table 1, while the full nucleotide and protein sequences for all genes are given in the supplementary data: *Pina* S2–S7; *Pinb* S8–S15; *Pinb-2* variant S16–S19 and *Gsp-1* S20–S23. All novel sequences were submitted to the European Nucleotide Archive.

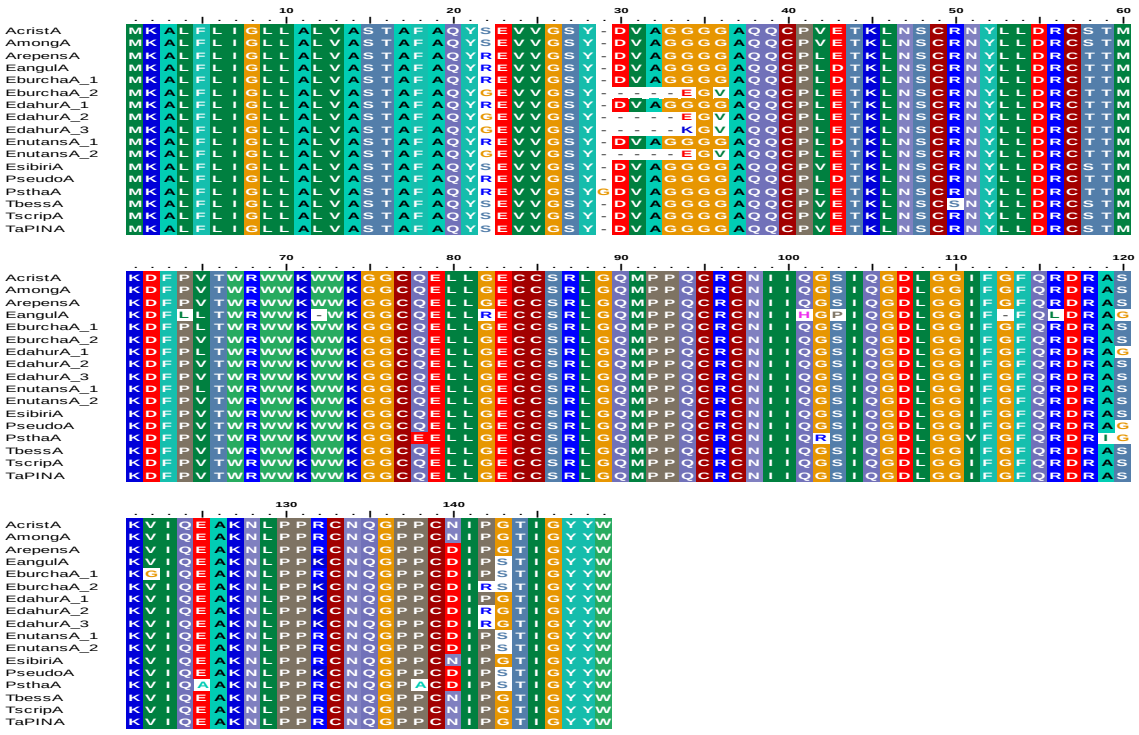

Figure 1. Protein alignment of examples of PINA forms detected in various *Elymus*, *Agropyron* and close relative species compared with 'wild-type' from *Triticum aestivum* bread wheat cv. Cadenza. (Codes used are: Acrist = *Agropyron cristatum*; Among = *Agropyron mongolicum*; Arepens = *Agropyron repens*; Eangul = *Elymus angulutus;* Eburch = *Elymus burchan-buddae*; Edahur = *Elymus dahuricus* subsp. *excelsus*; Enutans = *Elymus nutans*; Esibir = *Elymus sibiricus*; Pseudo = *Pseudoroegneria spicata;* Pstha = *Psathyrostachys juncea;* Tbess = *Thinopyrum bessarabicum;* Tscrip = *Thinopyrum scirpeum* and TaPINA = *T. aestivum cv. Cadenza* (PINA-D1a). 'A' denotes sequence type, i.e., PINA, and underscore/number refers to the number of different sequence types detected in the species screened.).

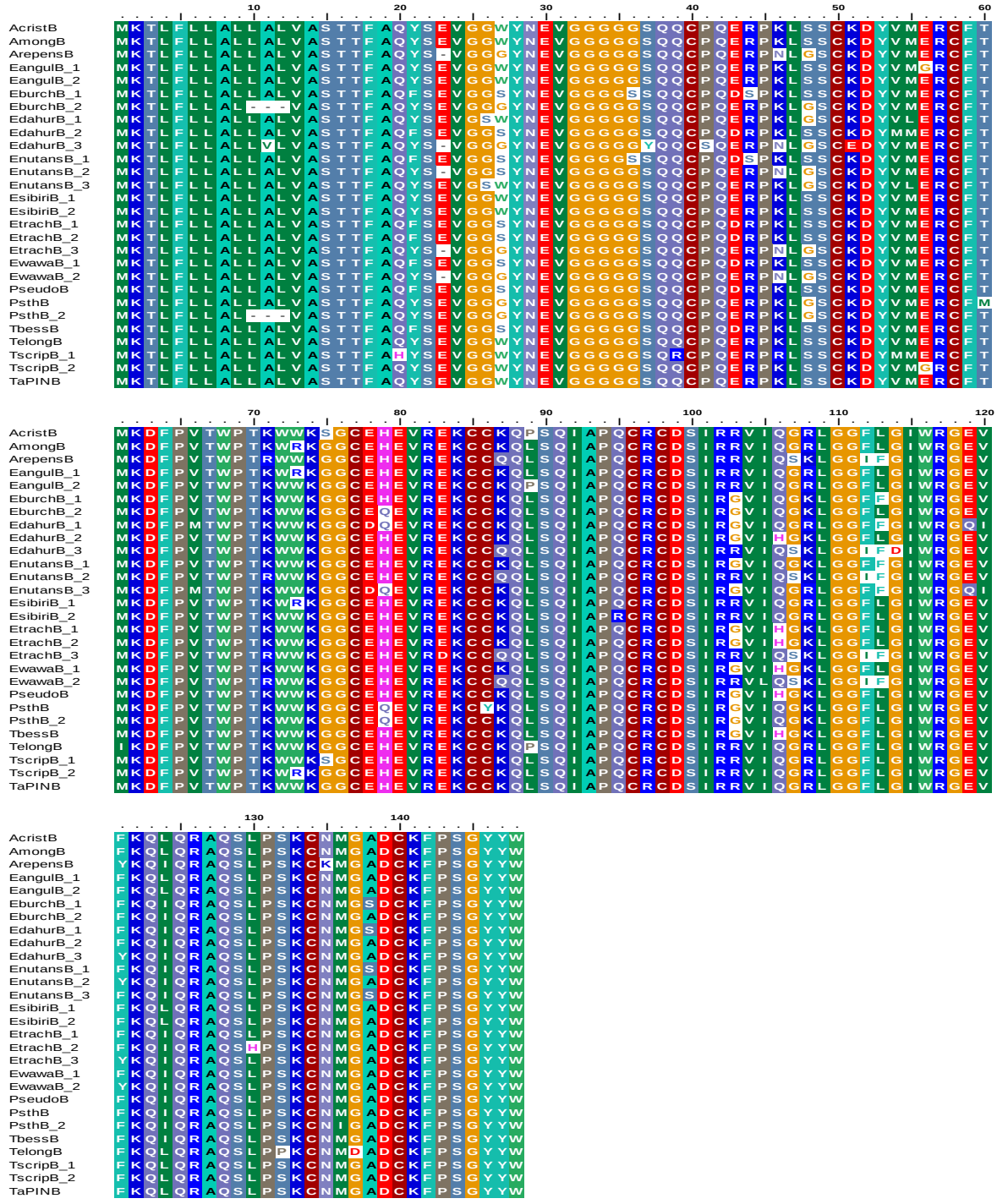

Figure 2. Protein alignment of examples of PINB forms detected in various *Elymus*, *Agropyron* and close relative species compared with 'wild-type' from *Triticum aestivum* bread wheat cv. Falcon. (Codes used are: Acrist = *Agropyron cristatum*; Among = *Agropyron mongolicum*; Arepens = *Agropyron repens*; Eangul = *Elymus angulutus*; Eburch = *Elymus burchan-buddae*; Edahur = *Elymus dahuricus*; Enuta = *Elymus nutans*; Esibir = *Elymus sibiricus*; Etrach = *Elymus trachycaulus* subsp. *subsecundus;* Ewawa = *Elymus wawawaiensis* Pseudo = *Pseudoroegneria spicata;* Pstha = *Psathyrostachys juncea;* Tbess = *Thinopyrum bessarabicum;* Telong = *Thinopyrum elongatum;* Tscrip = *Thinopyrum scirpeum and* TaPINB = *T. aestivum* cv. Falcon (PINB-D1). 'B' denotes sequence type, i.e., PINB, and underscore/number refers to the number of different sequence types detected in the species screened.).

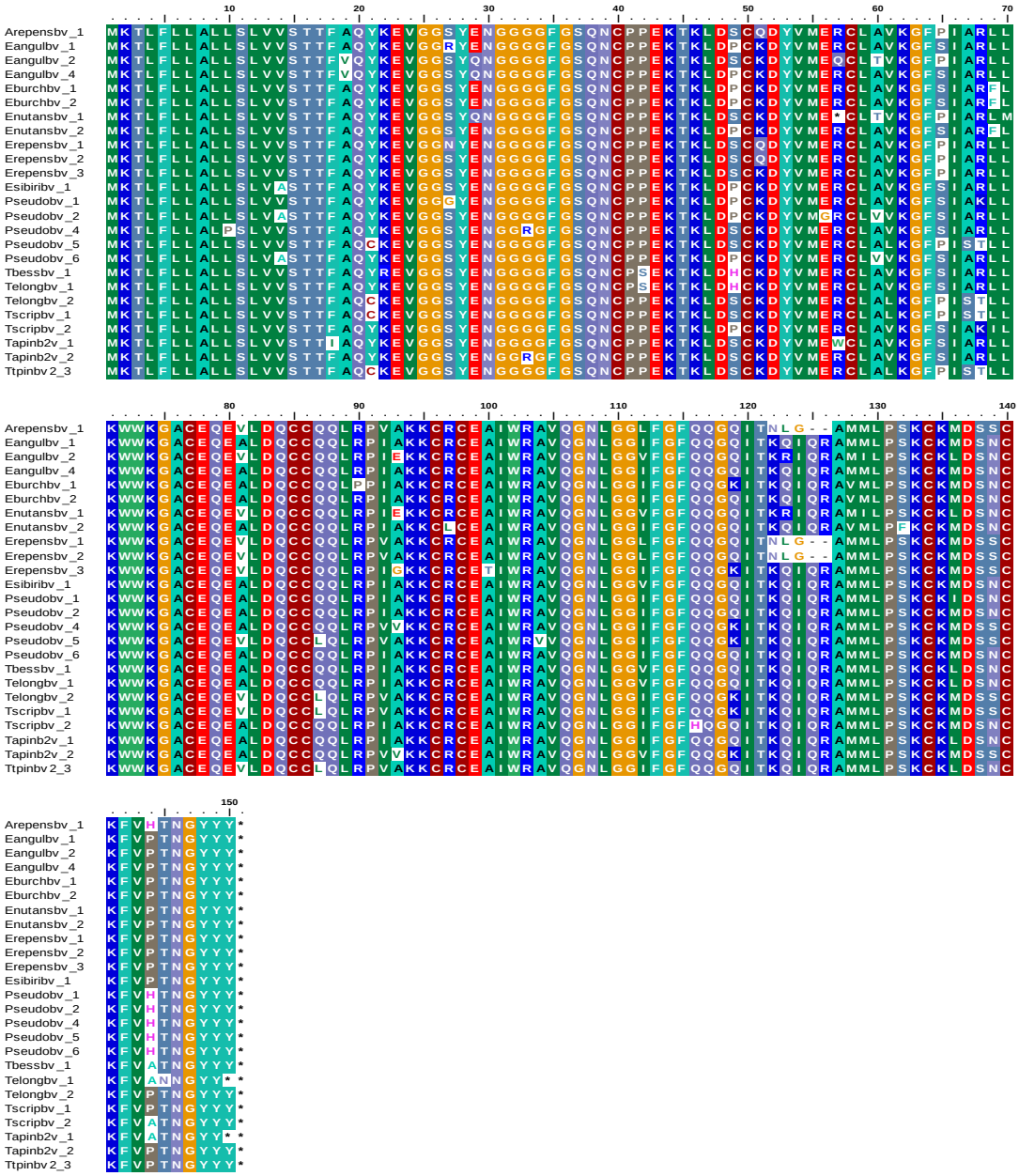

Figure 3. Protein alignment of examples of PINB-2 variant forms detected in various *Elymus*, *Agropyron* and close relative species compared with variant forms from *Triticum aestivum* bread wheat cv. Hereward and *Triticum turgidum* subsp. *durum*. (Codes used are: Arepens = *Agropyron repens*; Eangul = *Elymus angulutus*; Eburch = *Elymus burchan-buddae*; Enutans = *Elymus nutans;* Erepens= *Elymus repens*; Esibir = *Elymus sibiricus*; Pseudo = *Pseudoroegneria spicata;* Tbess = *Thinopyrum bessarabicum;* Telong = *Thinopyrum elongatum;* Tscrip = *Thinopyrum scirpeum and* TaPINV1; V2 = *T. aestivum cv. Hereward;* TtPINBv3 = *Triticum turgidum* subsp. *durum*. 'Bv' denotes sequence type, i.e., PINB variant (*Pinb-2* genes), and underscore/number refers to the number of different sequence types detected in the species screened.).

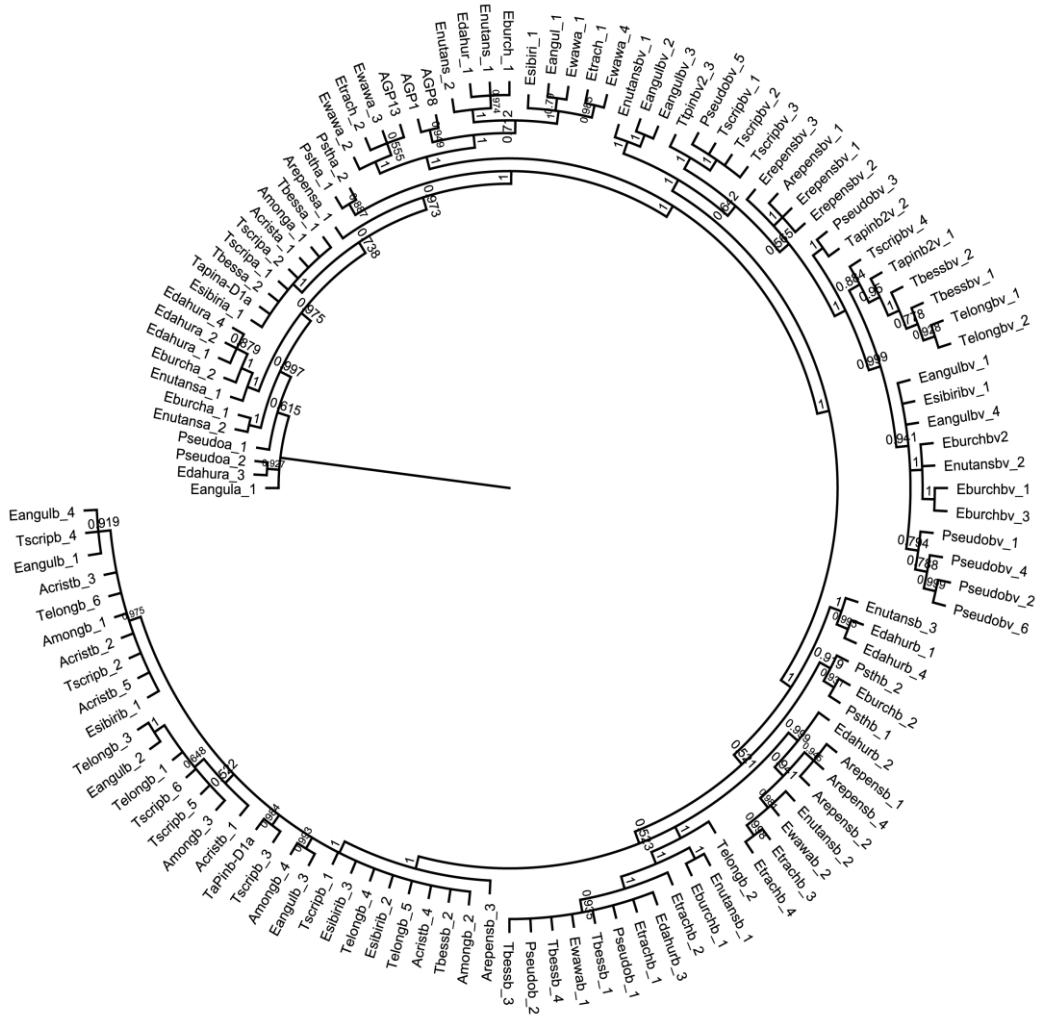

Figure 4. Phylogenetic tree of the nucleotide sequences of Pina, Pinb, Pinb-2v and gsp-1 for all species screened, with known *Triticum aestivum* sequences. Separations between *Gsp-1* and *Pin* genes are shown, with *Pinb-2* variant genes arising from the ancestral *Pinb* gene, after *Pinb* and *Pina* separation. The branches were transformed equally, and branches ordered in increasing order. The species codes are those used in the Figure legends from the alignments shown in Figures 1–3.

Table 1. Summary of the frequency of occurrence of *Pina*, *Pinb*, *Pinb-2* and *Gsp-1* genes in *Elymus* and *Agropyron* species and related species. (* = indicates already published information from [19]). +/− indicates the presence or absence of the genes in that species screened. The first number represents the number of different protein sequences detected; the number in brackets indicates the number of clones sequenced. BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) searches were carried out on all sequences and an example of relevant accession numbers related to sequence identity for genes screened is given, indicated by the superscripts 1, 2, 3, 4. Genome designations from [34].

Species and Genome	*Pina* Gene [1]	*Pinb* Gene [2]	*Pinb-2* Gene [3]	*Gsp-1* Gene [4]	Summary of Sequence Information
Agropyron cristatum (2x)-**P**	+1 (3)	+1 (6)	-	+5 (14 *)	100% identity P33432 [1]; 99% identity AHC54604 [2]; [19] [4]
Agropyron mongolicum (2x)-**P**	+1 (6)	+1 (6)	-	+9 (22 *)	100% identity P33432 [1]; 100% identity ALI87022 [2]; [19] [4]
Thinopyrum bessarabicum (2x)-**E^{b}** or **J**	+1 (6)	+1 (4)	+1 (3)	+2 (10 *)	99% identity P33432 [1]; 97% identity AER62841 [2]; 98% identity AFB35605 [3]; [19]
Thinopyrum elongatum (2x)-**J**	-	+1-(6)	+2-(14)	+2-(11 *)	98% identity AHC54604 [2]; 98% identity AFB35605 [3] & 100% identity CCH14734 [3]; [19] [4]
Thinopyrum scripeum (2x)-**E** or **J^{e}**	+1 (4)	+1 (6)	+2 (7)	+1 (2)	100% identity P33432 [1]; 97% identity CAH10199 [2]; 97% identity AFB35605 [3] & 100% identity CCH14734 [3]; 95% identity XP_020170444 [4]
Pseudoroegneria spicata (2x)-**St**	+1 (7)	+1 (9)	+2 (21)	+3 (11 *)	97% identity AER62831 [1]; 97% identity AER62841 [2]; 93% identity CAQ16391 [3] & 99% identity AFB35608 [3]; [19] [4]
Psathyrostachys juncea (2x)-**Ns**	+1 (8)	+2 (4)	-	+4 (9 *)	99% identity AER62828 [1]; 97–99% identity AER62849 [2]; [19] [4]
Elymus burchan-buddae (4x)-**StY**	+2 (9)	+2 (2)	+1 (4)	+1 (8)	98% identity AER62831 [1] & Novel Pin a sequence GenBank accession number LT669797; & 100% AER62849 [2] & Novel Pin b sequence Genbank accession number LR025202 [2]; 97% identity AFB35605 [3]; AGP1 detected [19] [4]
Elymus sibiricus (4x)-**StH**	+1 (6)	+2 (7)	+1 (2)	+1 (2)	100% identity P33432 [1]; 100% identity ALI87022 [2] & 99% Q10464 [2]; 97% identity AFB35605 [3]; AGP3 detected [19]
Elymus trachycaulus subsp. *subsecundus* (4x)-**StH**	-	+3 (10)	-	+2 (11)	97% identity AER62841 [2] and 98% identity BAK64220 [2]; AGP3 and AGP13 [19] [4]
Elymus wawawaiensis (4x)-**StH**	-	+2 (5)	-	+3 (9)	97% identity AER62841 [2] & 98% identity BAK64220 [2]; AGP3 and AGP13 detected [19] [4]; Novel Gsp-1 sequence GenBank accession number LT669800 [4]
Elymus angulatus (6x)-**StYH**	+1 (3)	+2 (5)	+2 (4)	+1 (6)	92% identity AER62832 [1] (two stop codons present); 100% identity AHC54604 & 99% identity ALI87022 [2]; 97% identity AFB35605 [3] & 97% BAM43258 [3]; AGP1 detected [19] stop codon present [4]
Elymus dahuricus subsp. *excelsus* (6x)-**StYH**	+2 (8)	+3 (12)	-	+1 (2)	97% identity AER62831 [1] & Novel Pin a sequence GenBank accession number LT669797; sequence ID SHD75392 [1]; 98% identity BAK64220 [2] & Novel Pin b sequence GenBank accession number LT669796 [2]; AGP1 detected [19] [4]
Elymus nutans (6x)-**StYH**	+2 (8)	+3 (6)	+1 (6)	+2 (8)	99% identity AER62831 [1] & Novel Pin a sequence GenBank accession number LT669797 [1]; 91% identity CAC33506 [2]; 98% identity BAK64220 [2] & 2 Novel Pin b sequence GenBank accession number LT669796 [2] & LS991245 [2]; 93% identity AFB35607 [3]; AGP1 detected [19]; Novel Gsp-1 sequence GenBank accession number LT669799 [4]
Agropyron repens (6x)-**StStH**	+1 (6)	+1 (6)	+1 (8)	+1 (4 *)	98% identity AER62831 [1]; 98% identity BAK64220 [2]; Novel Pinb-2 sequence GenBank accession number LT669798 [3]; [19] [4]
Elymus repens (6x)-**StStH**	+(not sequenced)	+(not sequenced)	+2 (9)	+2 (3 *)	95% identity AFB35607 [3] & Novel Pinb-2 sequence GenBank accession number LT669798 [3]; [19] [4]

3.1. Sequence Diversity in PINA Proteins

PCR screens using *Pina*-specific primers produced amplicons of approximately 450 base pairs from all *Elymus* and *Agropyron* species except for *Elymus trachycaulus* subsp. *subsecundus*, *Elymus wawawaiensis* and *Thinopyrum elongatum*. Two types of PINA sequence were detected in *Elymus burchan-buddae*, *Elymus dahuricus* subsp. *excelsus* and *Elymus nutans*. One was a novel form (GenBank accession number LT669797; sequence ID SHD75392), encoding a protein with 89% identity to PINA sequences from *Aegilops kotschyi* [29]. Alignment with wild-type PINA from *T. aestivum* cultivar Cadenza (PINA-D1a) (Figure 1) shows unique amino acid substitutions and deletions: Ser22Gly, four residues deleted corresponding to 29–32 and Asp33Gly in all species and additional substitutions Asp33Lys, Gly35Val, Val41Leu, Ser57Thr, Arg131Lys, Asn139Asp, and Pro141Arg substitutions in *Elymus dahuricus* subsp. *excelsus*. The second sequence type encoded a protein with 99–97% identity with published sequences from *Elymus libanoticus* [39] with unique Ser22Arg and Ser119Gly substitutions compared to wild-type protein sequences.

The PINA sequences from *Elymus sibiricus*, *Agropyron cristatum*, *Agropyron mongolicum*, *Thinopyrum scripeum* (100%), *Thinopyrum bessarabicum* (99%), *Agropyron repens* (98%) and *Pseudoroegneria spicata* (97%) showed high identity with the bread wheat sequence (accession number P33432; [40]). The sequence from *Psathyrostachys juncea* showed 99% identity to published sequences from *Psathyrostachys juncea* (accession number AER62828; [39]) with a glycine insertion at position 29 of the mature protein. The two *Elymus angulutus* sequences both had two stop codons at positions 72 and 113 of the mature proteins, meaning that they are unlikely to be expressed. Examples of the sequences detected in the species are shown in an alignment in Figure 1. The effect of these shorter PINA sequences from *Elymus burchan-buddae*, *Elymus dahuricus* and *Elymus nutans* on the protein structure, especially the tryptophan-rich loop region which is characteristic of PINs, was investigated using the Phyre2 web portal for protein modelling, prediction and analysis [41] (Figure 5 and Supplementary Figure S1).

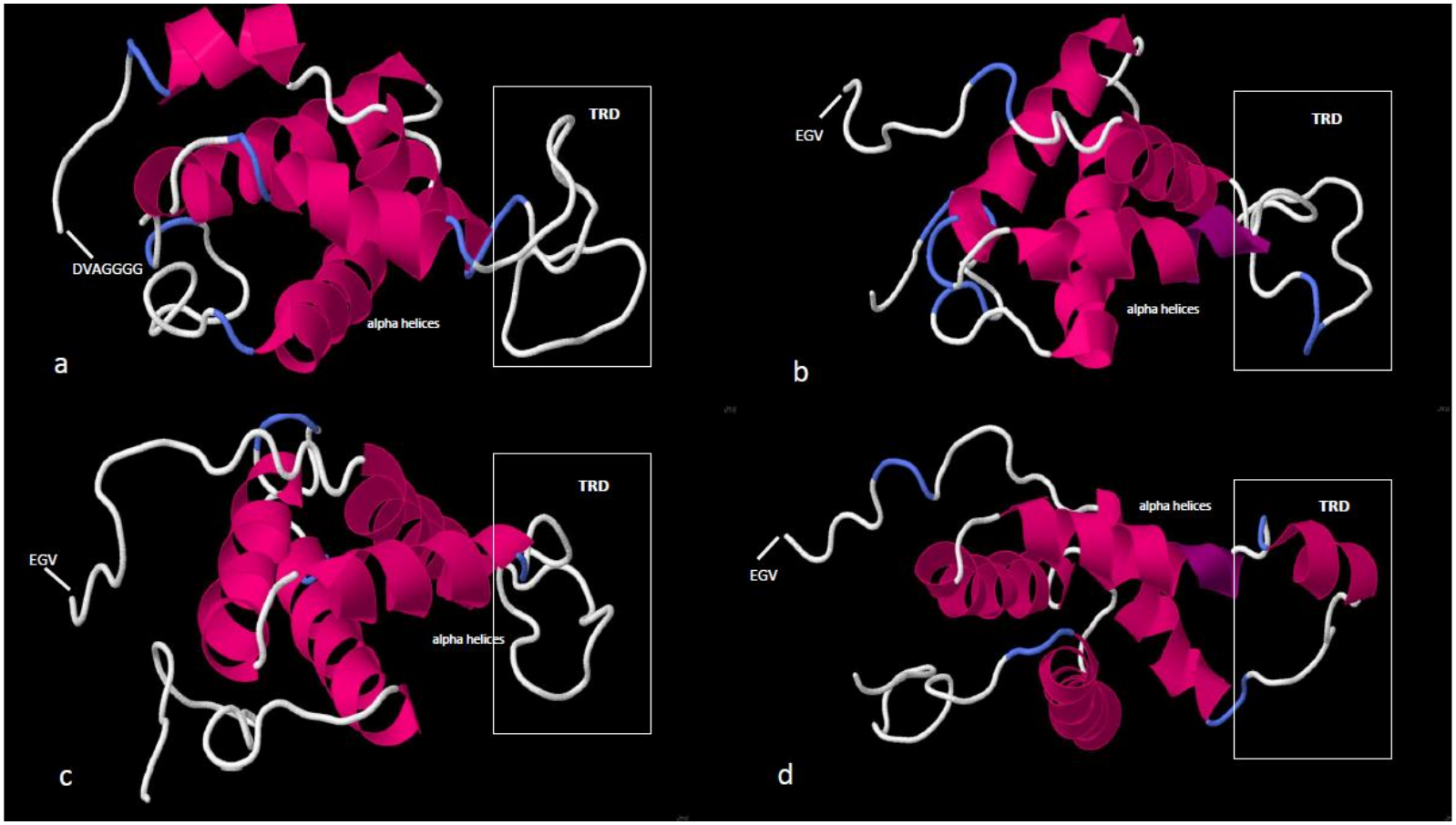

Figure 5. Predicted 3D structures of PINAs from: (**a**) *T. aestivum* (wild-type); (**b**) *Elymus burchan-buddae*; (**c**) *Elymus dahuricus*; and (**d**) *Elymus nutans* using the Phyre2 web portal for protein modelling. The amino acid location of the shorter sequences (i.e., EGV) compared to the wild-type sequence (i.e. DVAGGGG), tryptophan-rich domain (TRD) and alpha helices are indicated. The region in purple always indicates a pi-helix. Regions in blue indicate a protein turn. A turn is a secondary structure, structural motif where the C^{α} atoms of two residues separated by 1–5 peptide bonds are closer than normal (less than 7 Å [0.70 nm]), leading to the formation of an inter main chain hydrogen bond between the corresponding residues. The signal peptide (residues 1–28/24) has been removed as it would be in the post-translational processing.

3.2. Sequence Diversity in PINB Proteins

PINB sequences from *Agropyron cristatum* had 99% identity with published sequences from bread wheat [4] while those from *A. mongolicum*, *Elymus angulatus*, *Elymus sibiricus* and *Thinopyrum scripeum* showed 99–100% sequence identity with published sequences for bread wheat ([42]; ALI87022; direct submission). These sequences contained a Trp73Arg substitution. The sequences from *Agropyron repens* and *Elymus nutans* had 98% identity with sequences from *Hordeum patagonicum* subsp. *patagonicum* [43]. The second sequence type detected in *Elymus angulatus* differed only in this Trp73Arg substitution and was 100% identical with the PINB from bread wheat (accession number AHC54604; [44]). A second PINB sequence in *Elymus sibiricus*, was identical to Accession Q10464 from bread wheat [43] (Figure 2).

Elymus burchan-buddae and *Elymus nutans* both had two types of PINB sequence, one of which showed 91% identity with published sequences from *Aegilops speltoides* (accession number CAC33506; [45]). The-novel sequence type in *Elymus nutans*, *Elymus dahuricus* subsp. *excelsus* and *Elymus burchan-buddae* (GenBank accession number LS991245, LT669796; sequence ID SHD75391 and LR025202 respectively) show 91% identity with the *Aegilops speltoides* PINB type, with a range of substitutions: Gly26Ser, Trp27Ser/Gly, Glu43Asp, Ser48Gly, Val54Met, Met55Leu, His79Gln, Arg103Gly, Gln106His, Arg108Lys, Leu113Phe, Glu119Gln, Val120Ile, Leu125Ile, and Ala138Ser (see Figure 4). The second PINB sequence from *Elymus burchan-buddae* was identical to *Psathyrostachys juncea* PINB [36]. The second PINB sequence type in *Elymus dahuricus* subsp. *excelsus* had a 97% sequence identity with sequences from *Agropyron mongolicum* [39].

PINB sequences from *Elymus trachycaulus*, *Elymus wawawaiensis*, *Pseudoroegneria spicata* and *Thinopyrum bessarabicum* had 97% identity with published sequences for *A. mongolicum* [39] while second sequence types in *Elymus trachycaulus* and *Elymus wawawaiensis* showed 99% identity with hordoindoline sequences from *H. brachyantherum* subsp. *californicum* [43].

Two PINB sequence types were detected in *Psathyrostachys juncea*, showing 97% identity with PINB sequences from *Psathyrostachys huashanica* (accession number ADO17545; [46] direct submission) and 99% sequence identity with published sequences for *Psathyrostachys juncea* [39], respectively. Sequences from *Thinopyrum elongatum* and *Thinopyrum scripeum* showed 97% sequence identity with published sequences for bread wheat [4,40] (Figure 2).

3.3. Sequence Diversity in PINB-2 Proteins

Pinb-2 gene sequences were not detected in all species. However, a new PINB-2 sequence type was present in both *Agropyron repens* and *Elymus repens* (GenBank accession number LT669798; sequence ID SHD75393). This form contains a Ser65Pro substitution, as observed in several barley species [47], and three residues (Asn-Leu-Glu) at positions 122–124 instead of the five residues (Lys-Gln-Ile-Gln-Arg) in the barley proteins. The Ser65Pro substitution was also present in sequences from *E. angulutus* and three sequences from *Elymus repens* but these also contained the Lys-Gln-Ile-Gln-Arg sequence rather than Asn-Leu-Glu (Figure 3). PINB-2v1 type sequences containing single amino acid substitutions were detected in *E. angulutus*, *E. nutans*, *E. burchan-buddae*, and *E. sibiricus*, (Figure 3). PinB-2v1 type sequences from *E. angulutus*, *E. burchan-buddae*, and *E. sibiricus* had 97% sequence identity with published sequences for *T. aestivum* Accession number AFB35605 [26] while *Elymus nutans* sequences had 93% identity with PinB-2v2-2 Accession number AFB35607 [26]. A second sequence type found in *E. nutans* had 97% sequence identity with hordoindoline sequences from *H. brachyantherum* subsp. *californicum* [44], although it contained a stop codon at position 57 of the protein sequence.

Pseudoroegneria spicata contained sequences encoding PINB-2v1, PINB-2v2 and PINB-2v3, the former two being 99% identical to the published sequences for bread wheat [20]. The PINB-2v3 sequence for *Pseudoroegneria spicata* had 99% identity with published sequences for bread wheat (e.g., AFB35608; [26]). One PINB-2v1 sequence type contained a unique Arg68Lys substitution, which was also detected in *Thinopyrum scirpeum* showing 97% sequence identity with published sequences for PINB-2v1 [26]. The *Thinopyrum* species contained both PINB-2v1 and PINB-2v3 type protein sequences. The PINB-2v1 sequences in *Thinopyrum elongatum* and *Thinopyrum bessarabicum* had 98% identity with

PINB-2v1-3 sequences reported by [26] while PINB-2v3 sequences from both *Thinopyrum elongatum* and *Thinopyrum scirpeum* were identical to *Aegilops searsii* (accession number CCH14734, [48]; direct submission).

Variant Pinb-2 PCR products were not obtained from *Agropyron cristatum*, *Agropyron mongolicum*, *Elymus dahuricus* subsp. *excelsus*, *Elymus trachycaulus* subsp. *subsecundus*, *Elymus wawawaiensis* and *Psathyrostachys juncea*.

The PINB-2 sequences from *Elymus* species, *Agropyron* and related species are aligned with PINB-2v1 (CAQ16390), PINB-2v2 (CAQ16391) from *T. aestivum* bread wheat cultivar Hereward [20] and PINB-2v3 from *Triticum turgidum* subsp. *durum* (CAQ16392) [20] in Figure 3.

3.4. Gsp-1 Sequences in Elymus Species

Two new AGP sequence motifs were detected in *Elymus nutans* and *Elymus wawawaiensis*, respectively (Supplementary Figure S22). One of these, in *E. nutans* (GenBank accession number LT669799; sequence ID 75394) was like the AGP7 and AGP8 motifs reported by [19] with 93% identity with *Australopyrum retrofractum* and *Elymus libanoticus* sequences [36]. This AGP motif differs from AGP7 and AGP8 in Tyr21Phe, Ala22Val and Ala34Gly amino acid substitutions while the GSP-1 protein encoded by the same sequence had substitutions including Ser79Tyr, Phe81Val, Leu128Phe, Ile149Met and Phe153Asp, which have been reported before in GSP-1 protein sequences [19].

The AGP motif in *E. wawawaiensis* (GenBank accession number LT669800; sequence ID SHD75395) is like AGP13 [19], differing only in an Ala29Ser amino acid substitution, and has 96% identity with sequences in *H. bogdanii* Accession ADH94955 (Cenci et al., 2009, direct submission). Novel amino acid substitutions within the associate GSP-1 sequence included Glu88Lys and Leu106Pro substitution, not reported before [19]. The two new motifs are aligned with AGP8 and AGP13 in Supplementary Figure S26.

The major AGP motifs present in the other species had all been described before: AGP1 (*Elymus nutans*, *Elymus burchan-buddae*, and *Elymus dahuricus* subsp. *excelsus*); AGP3 (*Elymus angulatus*, *Elymus sibiricus*, *Elymus trachycaulus* subsp. *subsecundus* and *Elymus wawawaiensis*); and AGP13 (*Elymus trachycaulus* subsp. *subsecundus* and *Elymus wawawaiensis*). AGP1 was also detected in *Elymus angulatus* but had a stop codon present. Nucleotide and protein alignments are shown in the supplementary data, Figures S20–S23.

4. Discussion

For more than 20 years, gene mining of the *Ha* locus has provided valuable information on the diversity of the puroindoline genes and has identified variations in the sequences of the encoded proteins that have had an impact on their functionality, e.g., [4–6,17,28,49–52]. However, these studies have focused on the *Pin* gene family and on the progenitors of bread wheat and related wild grass species, e.g., [29–31,39,53].

This study characterized *Pina*, *Pinb*, *Pinb-2* and *Gsp-1* genes, focusing on diploid, tetraploid and hexaploid species of the genera *Elymus* and *Agropyron* species which are less closely related to wheat, but also include more closely related species. *Pinb* and *Gsp-1* genes were identified in all sixteen species that were screened, *Pina* genes in thirteen of the sixteen and *Pinb-2* genes in ten of the sixteen. Although the levels of sequence conservation were generally high, as reported by [19], some amino acid positions showed high levels of variation in all protein sequences. The differences between forms of PINA and PINB-2 proteins included in-frame deletions of amino acids while some PINB proteins had changes within the tryptophan-rich loop which is suggested to be essential for their role in determining grain texture (possibly by direct interaction with the starch granule surface) [48].

A combination of infrared and Raman spectroscopy and showed that PINA and PINB proteins have a secondary structure consisting of 30% α-helix, 30% β-sheet with 40% unordered structure [54]. The PIN and GSP proteins belong to a large family of seed storage proteins called the "prolamin superfamily" [55] and have been predicted to have similar structures to other members of this family,

namely the 2S storage albumins occurring in dicotyledonous seeds [56]. The InterPro database 66.0 identified all the proteins analyzed as part of this bifunctional inhibitor/lipid transfer protein/seed storage 2S albumin superfamily. A comparison with related proteins in this database suggests that residues 49–139 of the PINA, PINB and PINB-2 proteins form a bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain (named the IPR016140 domain), and PINB protein residues 50–59, 79–90, 92–101 and 125–140 form a cereal seed allergen/grain softness/trypsin and alpha-amylase inhibitor domain (named the IPR006106 domain). The AGP/GSP1 proteins analyzed were also shown to have IPR016140 domains (i.e., residues 63–154) and, similarly to PINB proteins, have IPR006106 domains (i.e., residues 63–72, 92–103, 105–114 and 138–153). These domains represent a structural region consisting of 4-helices with a folded leaf topology, and forming a right-handed super helix. 1H and 15N NMR-Spectroscopy confirmed that these finger millet (ragi) proteins form this globular 4-helix motif with a simple 'up-and-down' topology, including a short anti-parallel beta-sheet [57].

The PINA sequences from *Elymus sibiricus*, *Agropyron repens*, *Agropyron cristatum*, *Agropyron mongolicum*, *Thinopyrum bessarabicum*, *Thinopyrum scripeum* and *Pseudorogneria spicata* all showed high sequence identity (97–100%) with the published sequence from bread wheat [30,40,45] with only eight out of 148 residues (5.4%) showing variation in more than one sequence analyzed (Figure 1). Figure 1 shows that, of these eight residues, four substitutions showed no change in amino acid type, i.e., Val42Leu and Val64Leu both being hydrophobic, Glu43Asp being acidic and Ser58Thr being polar in nature. The four remaining substitutions were changes from polar to basic (Serr22Arg), polar to hydrophobic (Ser120Gly), polar to acidic (Asn140Asp) and hydrophobic to polar (Gly143Ser). There is no variation in the signal peptide (1–19 residues) and all the sequences maintained their ten cysteine residues. High conservation of PINA sequences was also observed by [30], who suggested that this was consistent with a role in plant defense. However, although in vitro studies have suggested that Pin proteins may contribute to plant defense [26], it should be noted that the biological roles of Pins, GSP and AGP have not been established in planta, and hence discussions of structure: functional relationships are speculative.

A new PINA sequence type, with the deletion of four amino acids at position 29-32 of the wild-type protein, was detected in *Elymus burchan-buddae*, *Elymus dahuricus* subsp. *excelsus* and *Elymus nutans*. The most closely related sequence is one reported for *Aegilops speltoides* (89% identity) [29] and the phylogenetic analysis (Figure 4) shows a clear separation of these sequences from other types of PINA. 3D model structures were generated using the Phyre2 web portal for protein modelling, prediction and analysis [41] and to predict any putative if any effects of these deletions/shorter sequences and amino acid variations on overall protein structure. Figure 5a is the best fit alignment of the wild type PINA to those detected in the *Elymus* species, and shows changes to the overall protein structure and to the tryptophan loop. We can only speculate that these changes may be due to amino acid changes detected in these sequences shown in Supplementary Figure S1. [8] used similar software, i.e., I-TASSER, to generate 3D structures to predict changes to their PINA proteins from synthetic hexaploid wheat lines. The structures generated were similar to the ones shown in Figure 5 with the four characteristic alpha helices, with the tryptophan-rich-domain (TRD) likely to form a loop also confirmed [8].

The PINB sequences were less conserved than the PINA sequences, with twenty-six of the 148 amino acid residues showing two or more amino acid substitutions (17.6%), some of which were specific to sequence type (Figure 2). Completely novel sequence types were identified which were 91% identical to sequences from *Aegilops speltoides* [45] (GenBank accession number LT669796; sequence ID SHD75391, LS991245 and LR025202), while sequences from several species with the H genome (shared with *Hordeum* species) (*Agropyron repens*, *Elymus nutans*, *Elymus trachycaulus* and *Elymus wawawaiensis*) were more like those of hordothionins in the *Hordeum* species [33,43]. The sequences from the *Elymus* species included several novel variants of PINB sequences within the *Elymus* genus and two hexaploid

Elymus species (*Elymus dahuricus* subsp. *excelsus and Elymus nutans*) each had three distinct PINB sequence types.

Of the 26 residues showing substitutions, ten were conserved in that they did not result in a change in the type of amino acid (Trp27Gly, Val54Met, Met55Leu, Val66Met, Leu89ProPhe112Ile, Leu113Phe, Val120Ile, Leu124Ile, Arg108Lys). However, the Leu89Pro mutation is known to confer the hard phenotype (PinB-1c (Rahman et al., 1994)). A change from a hydrophobic to a basic amino acid at position 73 (which is within the tryptophan loop region) occurred in sequences from *Agropyron mongolicum*, *Elymus angulutus* and *Elymus sibiricus*. Because of its position, this substitution could be expected to affect grain hardness and we have indeed found the same mutation in the hard *T. aestivum* cultivar Mercia (authors' unpublished results). Other mutations resulted in substitutions of polar for hydrophobic residues (Gly26Ser, Try27Ser, Gly36Ser, Gly75Ser, Gly107Ser, Phe121Tyr, Ala138Ser) and polar for basic residues (Glu43Asp, Lys46Asn, His79Gln, Lys87Gln). Domains IPR016140 and IPR006106 showed a lot more variation than for PINA proteins (i.e., Gly75Ser, His79Gln, Lys87Gln, Leu89Pro, Gly107Ser, Phe121Tyr, and Ala138Ser substitutions observed). All PINB sequences contained ten cysteine residues except for one sequence from *Psathyrostachys juncea* which contained a Cys86Try mutation (Figure 2): this could have a deleterious effect on protein stability as Cys86 is predicted to form a disulphide bond with Cys134 [26].

Pinb-2v1 genes occurred in more species than the other three gene types of *Pinb-v* genes, being detected in eight of the 16 species. Although they all had 97% sequence identity with sequences detected in bread wheat [26], unique amino acid substitutions were observed in the individual species (Figure 3). Some of the substitutions also occurred in the tryptophan loop, for example, Ser65Pro, Arg68Lys, Leu69Phe mutations and Leu69Ile (Domain IPR016140). Pinb-2v3 sequences were only found in *Pseudorogneria spicata*, *Thinopyrum bessarabicum* and *Thinopyrum elongatum* while *Pinb-2v2* was only found in *Pseudorogneria spicata*. However, a new PINB-2 variant form was detected in *Agropyron repens* and *Elymus repens* (GenBank accession number LT669798; sequence ID SHD75393), which contained the tripeptide motif Arg-Leu-Gly instead of the pentapeptide Arg-Gln-Ile-Gln-Arg at positions 122 to 12. This study has identified two new AGP/GSP sequences from the *Elymus* species.

The phylogenetic analysis in Figure 4 shows a clear grouping of the different sequence types with the *Gsp-1* gene sequences being quite separate from the sequences of all the *Pin* genes. The results also clearly show that the *Pinb-2* variant genes arose from the ancestral *Pinb* gene after the *Pinb* and *Pina* ancestral lines had separated.

In most cases, the number of different sequences detected in the species was directly related to ploidy levels. This was especially true for the *Pinb* gene where two to three different sequences were detected in the tetraploid *Elymus trachycaulus* subsp. *subsecundus*, *Elymus burchan-buddae* and *Elymus sibiricus*, the hexaploid *Elymus dahuricus* subsp. *excelsus* and *Elymus nutans.* Two different Pina sequences were also detected for *Elymus burchan-buddae*, *Elymus dahuricus* subsp. *excelsus* and *Elymus nutans*. The results reported here therefore contribute new information on the sequence diversity of *Pin* and *Gsp-1* genes in wild-grass species related to wheat, showing clear differences between the degree of sequence conservation of the encoded proteins, and within the individual protein sequences. Because all the species studied can be hybridized with wheat, it will be possible to exploit variations in traits resulting from this diversity in wheat improvement programmes. However, our ability to predict the biological significance of this diversity is currently limited by our lack of knowledge of the biological roles of the proteins *in planta.* In particular, the biological relevance of the effects on grain texture is unclear, with no obvious selective advantage, while other biological properties demonstrated in vitro (such as defense against pathogens and lipid binding) have failed to be confirmed *in planta*.

Supplementary Materials: The following are available online at http://www.mdpi.com/1424-2818/10/4/114/s1, Figure S1: Protein alignment of novel PINA sequences from *Elymus* species with wild type from *T. aestivum* (bread wheat). Black boxes indicate main amino acid motif differences between the sequences. Those in the green boxes indicate same amino acid type (i.e. hydrophobic or hydrophilic) change; Figure S2: Nucleotide alignment of Pin a sequences for *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae*, *Elymus sibiricus and Elymus dahuricus*. Number refers to clone number; Figure S3: Protein alignment of PINA sequences for *Elymus angulutus*, *Elymus*

nutans, *Elymus burchan-buddae*, *Elymus sibiricus and Elymus dahuricus*. Number refers to clone number; Figure S4: Nucleotide alignment of Pin a sequences for *Agropyron cristatum*, *Agropyron mongolicum*, and *Agropyron repens*. Number refers to clone number; Figure S5: Protein alignment of PINA sequences for *Agropyron cristatum*, *Agropyron mongolicum*, and *Agropyron repens*. Number refers to clone number; Figure S6: Nucleotide alignment of Pin a sequences for *Pseudoroegneria spicata*, *Psathyrostachys juncea*, *Thinopyrum bessarabicum* and *Thinopyrum scripeum*. Number refers to clone number; Figure S7: Protein alignment of PINA sequences for *Pseudoroegneria spicata*, *Psathyrostachys juncea*, *Thinopyrum bessarabicum* and *Thinopyrum scripeum*. Number refers to clone number; Figure S8: Nucleotide alignment of Pin b sequences for *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae*, *Elymus sibiricus and Elymus dahuricus*. Number refers to clone number; Figure S9: Protein alignment of PINB sequences for *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae*, *Elymus sibiricus* and *Elymus dahuricus*. Number refers to clone number; Figure S10: Nucleotide alignment of Pin b sequences for *Agropyron cristatum*, *Agropyron mongolicum*, and *Agropyron repens*. Number refers to clone number; Figure S11: Protein alignment of PINB sequences for *Agropyron cristatum*, *Agropyron mongolicum*, and *Agropyron repens*. Number refers to clone number; Figure S12: Nucleotide alignment of Pin b sequences for *Elymus trachycaulus* subsp. *subsecundus* and *Elymus wawawaiensis*. Number refers to clone number; Figure S13: Protein alignment of PINB sequences for *Elymus trachycaulus* subsp. *subsecundus* and *Elymus wawawaiensis*. Number refers to clone number; Figure S14: Nucleotide alignment of Pin b sequences for *Pseudoroegneria spicata*, *Psathyrostachys juncea*, *Thinopyrum bessarabicum Thinopyrum elongatum* and *Thinopyrum scripeum*. Number refers to clone number; Figure S15: Protein alignment of PINB sequences for *Pseudoroegneria spicata*, *Psathyrostachys juncea*, *Thinopyrum bessarabicum Thinopyrum elongatum* and *Thinopyrum scripeum*. Number refers to clone number; Figure S16: Nucleotide alignment of Pinb-2 (variant) sequences for *Agropyron repens*, *Elymus repens*, *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae* and *Elymus sibiricus*. Number refers to clone number; Figure S17: Protein alignment of PINB-2 (variant) sequences for *Agropyron repens*, *Elymus repens*, *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae* and *Elymus sibiricus*. Number refers to clone number; Figure S18: Nucleotide alignment of Pinb-2 variant sequences for *Pseudoroegneria spicata*, *Thinopyrum bessarabicum*, *Thinopyrum elongatum*, and *Thinopyrum scripeum*. Number refers to clone number; Figure S19: Protein alignment of PINB-2 variant sequences for *Pseudoroegneria spicata*, *Thinopyrum bessarabicum*, *Thinopyrum elongatum*, and *Thinopyrum scripeum*. Number refers to clone number; Figure S20: Nucleotide alignment of *Gsp-1* sequences for *Elymus dahuricus*, *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae*, *and Elymus sibiricus*. Number refers to clone number; Figure S21: Protein alignment of GSP1 sequences for *Elymus dahuricus*, *Elymus angulutus*, *Elymus nutans*, *Elymus burchan-buddae*, *and Elymus sibiricus*. Number refers to clone number; Figure S22: Nucleotide alignment of *Gsp-1* sequences for *Elymus trachycaulus* subsp. *subsecundus* and *Elymus wawawaiensis*. Number refers to clone number; Figure S23: Protein alignment of GSP1 sequences for *Elymus trachycaulus* subsp. *subsecundus* and *Elymus wawawaiensis*. Number refers to clone number.

Author Contributions: The main experimental design, laboratory research and writing of the manuscript was carried out by M.D.W. R.G. carried out the cloning and plasmid preparations for some of the sequence products used in the study. R.K. carried out the phylogenetic analysis of all the sequences and the generation of the phylogenetic tree figure used in the manuscript.

Funding: This research was funded by the Biotechnology and Biological Sciences Research Council (BBSRC) of the UK and the work forms part of the Designing Future Wheat strategic programme, grant number (BB/PO16855/1).

Acknowledgments: Rothamsted Research receives grant-aided support from the Biotechnology and Biological Sciences Research Council (BBSRC) of the UK and the work forms part of the Designing Future Wheat strategic programme (BB/P016855/1). The authors are very grateful for the editorial and intellectual contributions made to this manuscript by Professor Peter R. Shewry.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Greenham, T.J.; Altosaar, I. Wheat puroindolines tether to starch granule surfaces in puroindoline-null (Pin-null) plants. *J. Cereal Sci.* **2018**, *79*, 286–293. [CrossRef]
2. Morris, C.F.; Rose, S.P. Wheat. In *Cereal Grain Quality*; Chapman Hall: London, UK, 1996.
3. Morris, C.F. Puroindolines: The molecular genetic basis of wheat grain hardness. *Plant Mol. Biol.* **2002**, *48*, 633–647. [CrossRef] [PubMed]
4. Chantret, N.; Salse, J.; Sabot, F.; Rahman, S.; Bellec, A.; Laubin, B.; Dubois, I.; Dossat, C.; Sourdille, P.; Joudrier, P.; et al. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (*Triticum* and *Aegilops*). *Plant Cell* **2005**, *17*, 1033–1045. [CrossRef] [PubMed]
5. Turner, A.S.; Bradburne, R.P.; Fish, L.; Snape, J.W. New quantitative trait loci influencing grain texture and protein content in bread wheat. *J. Cereal Sci.* **2004**, *40*, 51–60. [CrossRef]
6. Weightman, R.M.; Millar, S.; Alava, J.; Foulkes, M.J.; Fish, L.; Snape, J.W. Effects of drought and the presence of the 1BL/1RS translocation on grain vitreosity, hardness and protein content in winter wheat. *J. Cereal Sci.* **2008**, *47*, 457–468. [CrossRef]

7. Greenwell, P.; Schofield, J.D. A starch granule protein associated with endosperm softness in wheat. *Cereal Chem.* **1986**, *63*, 379–380.
8. Ali, I.; Saradar, Z.; Rasheed, A.; Mahmood, T. Molecular characterization of the puroindoline-a and b alleles in synthetic hexaploid wheats and in in silico functional and structural insights into Pina-D1. *J. Theor. Biol.* **2015**, *376*, 1–7. [CrossRef] [PubMed]
9. Beecher, B.; Bettge, A.; Smidansky, E.; Giroux, M.J. Expression of wild-type pinB sequence in transgenic wheat complements a hard phenotype. *Theor. Appl. Genet.* **2002**, *105*, 870–877. [PubMed]
10. Martin, J.M.; Meyer, F.D.; Smidasky, E.D.; Wangugi, H.; Blechl, A.E.; Giroux, M.J. Complementation of the pina (null) allele with the wild type Pina sequence restores a soft phenotype in transgenic wheat. *Theor. Appl. Genet.* **2006**, *113*, 1563–1570. [CrossRef] [PubMed]
11. Heinze, K.; Kiszonas, A.M.; Murray, J.C.; Morris, C.F.; Lullien-Pellerin, V. *Puroindoline* genes introduced into durum wheat reduce milling energy and change milling behaviour similar to soft common wheats. *J. Cereal Sci.* **2016**, *71*, 183–189. [CrossRef]
12. Xia, L.; Geng, H.; Chen, X.; He, Z.; Lillemo, M.; Morris, C.F. Silencing of puroindoline a alters the kernel texture in transgenic bread wheat. *J. Cereal Sci.* **2008**, *47*, 331–338. [CrossRef]
13. Gasparis, S.; Orczyk, W.; Zalewski, W.; Nadolska-Orcyzk, A. The RNA-mediated silencing of one the Pin genes in allohexaploid wheat simultaneously decreases the expression of the other and increases grain hardness. *J. Exp. Bot.* **2001**, *62*, 4025–4036. [CrossRef] [PubMed]
14. Jolly, C.J.; Glenn, G.M.; Rahman, S. *GSP-1* genes are linked to the grain hardness locus (*Ha*) on wheat chromosome 5D. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 2408–2413. [CrossRef] [PubMed]
15. Tranquilli, G.; Heaton, J.; Chicaiza, O.; Dubcovsky, J. Substitutions and deletions of genes related to grain hardness in wheat and their effect on grain texture. *Crop Sci.* **2002**, *42*, 1812–1817. [CrossRef]
16. Wilkinson, M.D.; Tosi, P.; Lovegrove, A.; Corol, D.I.; Ward, J.L.; Palmer, R.; Powers, S.; Passmore, D.; Webster, G.; Marcus, S.E.; et al. The *Gsp-1* genes encode the wheat arabinogalactan peptide. *J. Cereal Sci.* **2017**, *74*, 155–164. [CrossRef]
17. Turner, M.; Mukai, Y.; Leroy, P.; Charaf, B.; Appels, R.; Rahman, S. The *Ha* locus of wheat: identification of a polymorphic region for tracing grain harness in crosses. *Genome* **1999**, *42*, 1242–1250. [CrossRef] [PubMed]
18. Gollan, P.; Smith, K.; Bhave, M. *Gsp-1* genes comprise a multigene family in wheat that exhibits a unique combination of sequence diversity yet conservation. *J. Cereal Sci.* **2007**, *45*, 184–198. [CrossRef]
19. Wilkinson, M.D.; Castells-Brooke, N.; Shewry, P.R. Diversity of sequences encoded by the *Gsp-1* genes in wheat and other grass species. *J. Cereal Sci.* **2013**, *57*, 1–9. [CrossRef]
20. Wilkinson, M.; Wan, Y.; Tosi, P.; Leverington, M.; Snape, J.; Mitchell, R.A.C.; Shewry, P.R. Identification and genetic mapping of variant forms of puroindoline b expressed in developing wheat grain. *J. Cereal Sci.* **2008**, *48*, 722–728. [CrossRef]
21. Chen, F.; Beecher, B.; Morris, C.F. Physical mapping and a new variant of *Puroindoline b-2* genes in wheat. *Theor. Appl. Genet.* **2010**, *120*, 745–751. [CrossRef] [PubMed]
22. Geng, H.; Beecher, B.S.; He, Z.; Kiszonas, A.M.; Morris, C.F. Prevalence of *Puroindoline D1* and *Puroindoline b-2* variants in U.S. Pacific Northwest wheat breeding germplasm pools, and their association with kernel texture. *Theor. Appl. Genet.* **2012**, *124*, 1259–1269. [CrossRef] [PubMed]
23. Geng, H.; Beecher, B.S.; Pumphrey, M.; He, Z.; Morris, C.F. Segregation analysis indicates that *Puroindoline b-2 variants* 2 and 3 are allelic in *Triticum aestivum* and that a revision to *Puroindolineb-2* gene symbolization is indicated. *J. Cereal Sci.* **2013**, *57*, 61–66. [CrossRef]
24. Chen, F.; Xu, H.X.; Zhang, F.Y.; Xia, X.C.; He, Z.H.; Wang, D.W.; Dong, Z.D.; Zhan, K.H.; Cheng, X.Y.; Cui, D.Q. Physical mapping of *puroindoline b-2* genes and molecular characterization of a novel variant in durum wheat (*Triticum turgidum* L.). *Mol. Biol.* **2011**, *28*, 153–161. [CrossRef]
25. Chen, F.; Zhang, F.Y.; Cheng, X.Y.; Morris, C.F.; Xu, H.X.; Dong, Z.D.; Zhan, K.H.; Cui, D.Q. Association of Puroindoline b-B2 variants with grain traits, yield components and flag leaf size in bread wheat (*Triticum aestivum* L.) varieties of the Yellow and Huai valleys of China. *J. Cereal Sci.* **2010**, *52*, 247–253. [CrossRef]
26. Ramalingam, A.; Palombo, E.A.; Bhave, M. The *Pinb-2* genes in wheat comprise a multigene family with sequence diversity and important variants. *J. Cereal Sci.* **2012**, *56*, 171–180. [CrossRef]
27. Giroux, M.J.; Kim, K.H.; Hogg, A.C.; Martin, J.M.; Beecher, B. The *Puroindoline b-2* variants are expressed at low levels relative to the *Puroindoline D1* genes in wheat seeds. *Crop Sci.* **2013**, *53*, 833–841. [CrossRef]

28. Gautier, M.F.; Cosson, P.; Guirao, A.; Alary, R.; Joudrier, P. Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid *Triticum* species. *Plant Sci.* **2000**, *153*, 81–91. [CrossRef]
29. Chen, M.; Wilkinson, M.; Tosi, P.; He, G.; Shewry, P.R. Novel puroindoline and grain softness protein alleles in *Aegilops* species with the C, D, S, M and U genomes. *Theor. Appl. Genet.* **2005**, *111*, 1159–1166. [CrossRef] [PubMed]
30. Massa, A.N.; Morris, C.F. Molecular evolution of the puroindoline-a, puroindoline-b, and grain softness protein-1 genes in the tribe Triticeae. *J. Mol. Evol.* **2006**, *63*, 526–536. [CrossRef] [PubMed]
31. Bhave, M.; Morris, C. Molecular genetics of puroindolines and related genes: Allelic diversity in wheat and other grasses. *Plant Mol. Biol.* **2008**, *66*, 205–219. [CrossRef] [PubMed]
32. Linc, G.; Gaál, E.; Molnár, I.; Icso, D.; Badaeva, E.; Molnár-Láng, M. Molecular cytogenetic (FISH) and genome analysis of diploid wheatgrasses and their phylogenetic relationship. *PLoS ONE* **2017**, *12*, e0173623. [CrossRef] [PubMed]
33. McMillan, E.; Sun, G. Genetic relationships of tetraploid *Elymus* species and their genomic donor species inferred from polymerase chain reaction-restriction length polymorphism analysis of chloroplast gene regions. *Theor. Appl. Genet.* **2004**, *108*, 535–542. [CrossRef] [PubMed]
34. Wang, R.R.C. Wild Crop Relatives: Genomic and Breeding Resources. In *Agropyron and Psathyrostachys*; Springer: Berlin/Heidelberg Germany, 2011.
35. Gazza, L.; Galassi, E.; Ciccoritti, R.; Cacciatori, P.; Pogna, N.E. Qualitative traits of perennial wheat lines derived from different *Thinopyrum* species. *Genet. Res. Crop Evol.* **2016**, *63*, 209–219. [CrossRef]
36. Posada, D. jModelTest: Phylogenetic Model Averaging. *Mol. Biol. Evol.* **2008**, *25*, 1253–1256. [CrossRef] [PubMed]
37. Darriba, D.G.L.; Taboada, R.; Doallo, D. Posada jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **2012**, *9*, 772. [CrossRef] [PubMed]
38. Ronquist, F.M.; Teslenko, P.; van der Mark, D.L.; Ayres, A.; Darling, S.; Hohna, B.; Larget, L.; Liu, M.A.; Suchard, J.P. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **2012**, *61*, 539–542. [CrossRef] [PubMed]
39. Escobar, J.S.; Scornavacca, C.; Cenci, A.; Guilhaunmon, C.; Santoni, S.; Douzery, E.J.; Ranwez, V.; Glemin, S.; David, J. Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol. Biol.* **2011**, *11*, 181. [CrossRef] [PubMed]
40. Gautier, M.F.; Aleman, M.E.; Guirao, A.; Marion, D.; Joudrier, P. *Triticum aestivum* puroindolines, two basic cysteine-rich seed proteins: cDNA sequence analysis and development gene expression. *Plant Mol. Biol.* **1994**, *25*, 43–51. [CrossRef] [PubMed]
41. Kelley, L.A.; Mezulis, S.; Yates, C.M.; Wass, M.N.; Sternberg, M.J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **2015**, *10*, 845–858. [CrossRef] [PubMed]
42. Pouch, M.; Vaculova, K.; Milotova, J.; Stehno, Z.; Curn, V. Molecular and phylogenetic study of puroindoline genes in wild and cultivated wheat species with different ploidy levels. **2018**, in press.
43. Terasawa, Y.; Rahman, S.M.; Takata, K.; Ikeda, T.M. Distribution of Hordoindoline genes in the genus Hordeum. *Theor. Appl. Gene.* **2012**, *124*, 143–151. [CrossRef] [PubMed]
44. Shaaf, S.; Sharma, R.; Baloch, F.S.; Badaeva, E.D.; Knupffer, H.; Kilian, B.; Ozkan, H. The grain Hardness locus characterized in a diverse wheat panel (*Triticum aestivum* L.) adapted to the central part of the Fertile Crescent: genetic diversity, haplotype structure, and phylogeny. *Mol. Genet. Genom.* **2016**, *291*, 1259–1275. [CrossRef] [PubMed]
45. Lillemo, M.; Simeone, M.C.; Morris, C.F. Analysis of puroindoline a and b sequences from *Triticum aestivum* cv. 'Penawawa' and related diploid taxa. *Euphytica* **2002**, *126*, 321–331. [CrossRef]
46. Yan, X.; Wang, Y.-Q.; Zhao, J.-X.; Wu, J.; Chen, X.-H.; Cheng, X.-N. Isolation and sequence analysis on the grain hardness genes in *Psathyrostachys huashanica*. **2018**, in press.
47. Terasawa, Y.; Takata, K.; Anai, T.; Ikeda, T.M. Identification and distribution of *Puroindoline b-2* variant gene homologs in Hordeum. *Genetica* **2013**, *141*, 359–368. [CrossRef] [PubMed]
48. Wilkinson, M.D.; Shewry, P.R. Screening of *Aegilops* species for Puroindoline b variants. **2018**, in press.
49. Bhave, M.; Morris, C. Molecular genetics of puroindolines and related genes: Regulation of expression, membrane binding properties and applications. *Plant Molec. Biol.* **2008**, *66*, 221–231. [CrossRef] [PubMed]

50. Chen, F.; Li, H.; Cui, D. Discovery, distribution and diversity of *Puroindoline-D1* genes in bread wheat from five countries (*Triticum aestivum* L.). *Plant Biol.* **2013**, *13*, 125. [CrossRef] [PubMed]
51. Ma, X.; Sajjad, M.; Wang, J.; Yang, W.; Sun, J.; Li, X.; Zhang, A.; Liu, D. Diversity, distribution of *Puroindoline* genes and their effect on kernel hardness in a diverse panel of Chinese wheat germplasm. *BMC Plant Biol.* **2017**, *17*, 158. [CrossRef] [PubMed]
52. Boehm, J.D.; Ibba, M.I.; Kiszonas, A.M.; See, D.R.; Skinner, D.Z.; Morris, C.F. Genetic analysis of kernel texture (grain hardness) in a hard-red spring wheat (*Triticum aestivum* L.) bi-parental population. *J. Cereal Sci.* **2018**, *79*, 57–65. [CrossRef]
53. Cuesta, S.; Guzman, C.; Alvarez, J.B. Allelic diversity and molecular characterization of *puroindoline* genes in five diploid species of the *Aegilops* genus. *J. Exp. Bot.* **2013**, *64*, 5133–5143. [CrossRef] [PubMed]
54. Bihan, T.L.; Blochet, J.E.; Desormeaux, A.; Marion, D.; Pezolet, M. Determination of the secondary structure and conformation of puroindolines by infrared and Raman spectroscopy. *Biochemistry* **1996**, *35*, 12712–12722. [CrossRef]
55. Shewry, P.R.; Halford, N.G. Cereal seed storage proteins: Structures, properties and role in grain utilization. *J. Exp. Bot.* **2012**, *53*, 947–958. [CrossRef]
56. Elmorjani, K.; Geneix, N.; Dagalarrondo, M.; Branlard, G.; Didier, M. Wheat grain softness protein (*Gsp-1*) is a puroindoline-like protein that displays a specific post-translational maturation and does not interact with lipids. *J. Cereal Sci.* **2013**, *58*, 117–122. [CrossRef]
57. Strobl, S.M.; Mühlhahn, P.; Bernstein, R.; Wiltscheck, R.; Maskos, K.; Wunderlich, M.; Huber, R.; Glockshuber, R.; Holak, T.A. Determination of the three-dimensional structure of the bifunctional alpha-amylase/trypsin inhibitor from ragi seeds by NMR spectroscopy. *Biochemistry* **1995**, *34*, 8281–8293. [CrossRef]