agronomy

MDPI

Article

Generation and Characterisation of a Reference Transcriptome for Phalaris (*Phalaris aquatica* L.)

Rebecca C. Baillie [1], Michelle C. Drayton [1], Luke W. Pembleton [1], Sukhjiwan Kaur [1], Richard A. Culvenor [2], Kevin F. Smith [3], German C. Spangenberg [1,4], John W. Forster [1,4,*] and Noel O. I. Cogan [1,4]

1 Agriculture Victoria, AgriBio, the Centre for AgriBioscience, 5 Ring Road, Bundoora, Victoria 3083, Australia; bec.baillie@ecodev.vic.gov.au (R.C.B.); michelle.drayton@ecodev.vic.gov.au (M.C.D.); luke.pembleton@ecodev.vic.gov.au (L.W.P.); sukhjiwan.kaur@ecodev.vic.gov.au (S.K.); german.spangenberg@ecodev.vic.gov.au (G.C.S.); noel.cogan@ecodev.vic.gov.au (N.O.I.C.)

2 CSIRO Agriculture Flagship, Canberra, Australian Capital Territory 2601, Australia; r.culvenor@pi.csiro.au

3 Faculty of Veterinary and Agricultural Sciences, The University of Melbourne, Melbourne, Victoria 3010, Australia; kfsmith@unimelb.edu.au

4 School of Applied Systems Biology, La Trobe University, Bundoora, Victoria 3086, Australia

* Correspondence: john.forster@ecodev.vic.gov.au; Tel.: +61-03-90327054

Academic Editor: Chengdao Li

Received: 24 October 2016; Accepted: 25 January 2017; Published: 15 February 2017

Abstract: *Phalaris aquatica* is a cool-season perennial grass species that is extensively cultivated in Australia, with additional usage in other areas of the world. Phalaris displays a number of desirable agronomic characteristics, although unfavourable traits include excessive seed shattering, sensitivity to aluminium toxicity, and several toxicosis syndromes. Varietal development has to date been based on traditional selection methods, but would benefit from the application of genomics-based approaches, which require the development of large-scale sequence resources. Due to a large nuclear DNA content, methods that target the expressed component of the genome and reduce the complexity of analysis are most amenable to current sequencing technologies. A reference unigene set has been developed by transcriptome sequencing of multiple tissues from a single plant belonging to the variety Landmaster. Comparisons have been made to gene complements from related species, as well as reference protein databases, and patterns of gene expression in different tissues have been evaluated. A number of candidate genes relevant to removal of undesirable attributes have been identified. The reference unigene set will provide the basis for detailed studies of differential gene expression and identification of candidate genes for potential transgenic deployment, as well as a critical resource for genotypic analysis to support future genomics-assisted breeding activities for phalaris improvement.

Keywords: RNA-Seq; harding grass; de novo assembly; sequence annotation; tissue-specific gene expression; *P. tuberosa*; Poaceae

1. Introduction

Phalaris aquatica L. (syn. *P. tuberosa* L.), known as phalaris in Australia and harding grass elsewhere, is a cool-season perennial grass species that has been exploited for forage production. The largest zones of cultivation are in Australia, conservatively estimated as c. 2.7 million ha in 1994 [1]. Following recognition of the value of phalaris, usage spread to New Zealand, USA, several countries in South America, such as Argentina, and also, to a limited extent, northern Africa, southern Europe, and South Africa.

Phalaris displays high productivity throughout the cooler seasons, and nutritive quality comparable to other temperate grasses during this period [2]. Phalaris also tolerates heavy grazing following establishment, grows well on a variety of soil types, is tolerant of waterlogging and moderate levels of soil salinity, and is relatively unaffected by pests and diseases. An excellent ability to survive periods of summer drought is a major favourable characteristic of phalaris, attributable to partial dormancy of buds located at the bases of reproductive tillers, combined with a deep root system capable of accessing sub-soil moisture [3]. Under current scenarios of climate change for southern Australia, with the likelihood of longer and more severe droughts, increased adoption of phalaris may be anticipated. A number of unfavourable agronomic traits, however, are also observed, including relatively slow seedling establishment and productivity in the establishment year; sensitivity to soil acidity and associated aluminium toxicity [4]; susceptibility to seed shattering [5]; and several toxicosis syndromes including 'phalaris staggers' and 'phalaris sudden death' [6]. The former has been attributed to the neurotoxic effects of indole alkaloids such as mono- and dimethylated tryptamines and tyramines, as well as β-carbolines derived from tryptamine [7,8], while the causative agent for the latter is still unknown, although cyanogenic glycosides may be responsible for some cases [9].

The genus *Phalaris* (colloquially known as canary grasses) belongs to the tribe Aveneae of the Pooideae (cool-season grass sub-family) of the grass and cereal family Poaceae [10–12]. *P. aquatica* is a tetraploid taxon with a chromosomal constitution of $2n = 4x = 28$. The nature of tetraploidy in *P. aquatica* has been the subject of conflicting reports. A cytogenetic study of two 4x taxa, *P. arundinacea* (reed canary grass) and *P. aquatica*, identified predominant bivalent formation in each instance, supporting allopolyploid ($P_xP_xP_yP_y$) structures. However, continued bivalent formation in the F_1 hybrid between the two species suggested a more complicated situation, being suggestive of partial homology within genomes [13]. Meiotic studies of a monoploid ($2n = 14$) plant derived from cv. Australian detected an unusually high proportion of bivalents (average of 4.3 per pollen mother cell, as compared to 5.4 univalents) [14], indicating a segmental allopolyploid constitution [15]. Further chromosome pairing studies in interspecies hybrids supported a close relationship between the putative sub-genomes of *P. aquatica* [16], implying that regularised bivalent formation during meiosis in *P. aquatica* may be dependent on active control of synapsis, despite close sequence similarities between at least some members of the basic chromosome complement. Genome size estimates based on microdensitometry have been performed for tetraploid (4x) races of *P. arundinacea* [17]. Mean values are in the vicinity of 9.5 pg nuclear DNA (2C). On this basis, each haploid genome would contain c. 2.38 pg. Using a common conversion factor of 1 pg = 9.78×10^8 bp [18], the haploid genome size would be 2.3×10^9 bp. Similar results were obtained by flow cytometry [19]. Assuming a similar DNA content for *P. aquatica*, each haploid sub-genome component may be of similar size, comparable to values for other Aveneae species such as oat (*Avena sativa* L.) (c. 2.1×10^9 bp [20]) and members of the closely related Poeae tribe such as perennial ryegrass (*Lolium perenne* L.: 2×10^9 bp). On this basis, the genome of *P. aquatica* is likely to be large and complex, due to a prevalence of moderately to highly repetitive DNA, and the likelihood of significant homoeologous (between sub-genomes) and paralogous (between duplicated gene copies) sequence variation [21]. In addition, the outbreeding reproductive habit of phalaris will generate allelic diversity within and between individuals [22], contributing to homologous sequence variation.

To date, phalaris improvement and varietal development has been based on classical selection methods and has not incorporated molecular breeding technologies such as marker-assisted selection or transgenesis. Nonetheless, methods such as genomic selection, based on assessment of sequence diversity at high-density across whole genomes, are highly attractive for implementation in forage species such as phalaris [23,24]. Such approaches, however, depend on the availability of large-scale genomic resources. For species such as phalaris, which are of limited international significance despite regional importance, insufficient resources are currently available for the generation and assembly of a full genome sequence. In contrast, sequencing of the transcriptome, corresponding to the expressed portion of the genome, provides an attractive option, especially through the use of RNA-Seq technology

on second-generation sequencing platforms [25]. Despite a c. 10^3-fold variation in genome size across the angiosperms [26], transcriptomes of diploid species typically vary over a much narrower range, from 50–80 Mbp [27]. Sampling of RNA samples from multiple developmental stages or environmental conditions can allow construction of a transcription atlas, supporting gene isolation, development of gene-associated molecular genetic markers, comparative genomics, identification of differentially regulated gene sets, and quantification of gene expression, as well as ultimate annotation of genome sequences [25,28]. A transcriptome study has previously been performed for another member of the *Phalaris* genus, *P. arundinacea*, in order to provide data on differential gene expression in response to environmental stress [29].

The present study describes the development of a reference unigene set based on transcriptome sequencing from multiple tissues of a single plant from the variety Landmaster, representing the first comprehensive genomic resource for phalaris. Comparisons were made to gene complements from related species. Unigenes were annotated, and tissue-specific expression patterns have been identified. The value of the dataset was exemplified through detection of candidate genes for a number of important agronomic traits. The unigene set will provide a valuable resource for future genomics-assisted breeding activities in phalaris.

2. Results

2.1. De Novo Sequence Assembly of the Phalaris Transcriptome

The reference unigene set for the selected cv. Landmaster genotype reference was obtained from a total of 553,566,274 sequence reads. For the initial short oligonucleotide analysis package (SOAP) denovo assembly, a range of *k*-mers were empirically tested and a value of 71 was identified as delivering the optimal assembly, based on size of assembly compared to mean and median contig and scaffold size. The initial 71 *k*-mer assembly was 233,054,090 bp in length (not including Ns) from 437,776 contigs and scaffolds, with a mean length of 577 bp and N50 values of 1304 bp from 52,492 contigs. Following this initial assembly, contigs that were <149 bp were removed, as the single sequence read length was 150 bp and so these contigs are likely to be spurious features within the complete data set. Contigs that were <250 bp were required to have >10 sequence reads associated with the assembly, otherwise they were also removed from the assembly. This filtering step removed a large number of contigs, and left 217,707 contigs and scaffolds (49.7% of total).

The remaining contigs and scaffolds were then filtered by performing a basic local alignment search tool using a translated nucleotide query (BLASTX analysis in comparison to the uniref90 database. A total of 107,463 scaffolds and contigs were identified as being of plant origin (78,713 scaffolds relating to 26,467 loci and 28,750 contigs), while 18,470 contigs were of non-plant origin and 91,774 failed to return any match. Any contig or scaffold that failed to return a match to a known plant protein was removed. These included 9154 contigs identified as deriving from the *Puccinia* genus of foliar rust-causing fungal pathogens (which includes the stem rust pathogen *P. graminis*, known to infect phalaris), as well a substantial number of insect and bacterial origin. The assembled scaffolds identified as forked bubble or complex in nature were individually assembled with CAP3, after manual sequence assembly and evaluation was performed on a limited set to assess the degree of sequence variation between locus-related contigs. Manual evaluation was performed to evaluate the potential for co-assembly of homoeologous variants of a gene locus, given the predicted allopolyploid constitution of phalaris. However, minimal sequence divergence was identified from the range of examined loci. If the observed loci were generated as a result of coalescence between homoeologous variants this limited sequence divergence would make sub-genome sequence-specific assembly technically impossible. Processing of these scaffolds through the CAP3 assembler generated 14,324 scaffolds relating to 12,668 loci, as well as 18,020 scaffolds that were unable to be assembled. Representatives of a range of complex loci that were unassembled were therefore manually analysed and aligned. In all instances a single contig was able to be generated for the locus from the scaffolds.

A common issue was identified as the presence of large predicted gaps in the scaffolds filled with Ns. As a result, the longest scaffold derived from the unassembled complex locus was entered into the reference.

Following extensive sequence analysis and filtering, a final reference of 56,873 sequences, corresponding to 58,174,765 bp, was created. The assembled reference has N50 values of 11,945 sequences $\geq$ 1558 bp in length, with a GC content of 49.11%. The final assembled reference represents only 24.9% of the sequence length of total unfiltered assemblies, and only 13% of the initial contig and scaffold number (Table 1). Examination of the Uniref90 results relating to the final reference identified a total of 37,687 different proteins. A further examination of the taxonomic distribution of uniref90 proteins revealed that >69% of all retained sequences displayed a highest BLASTX match to a protein from Poaceae species (either *B. distachyon* or the cereal species *T. aestivum* L. (bread wheat), *Hordeum vulgare* L. (barley), or *Aegilops tauschii* L., all of which belong to the sub-family Pooideae of the Poaceae family, along with phalaris (Figures S1 and S2)). Comparison of the final 56,873 reference transcripts to a core set of 956 single copy plantae orthologs revealed 64% as complete, 24% fragmented, and only 11% missing (BUSCO notation: C:64%[D:17%], F:24%, M:11%, n:956).

Table 1. Overview of sequencing outputs and assembly.

Primary Assembly—SOAPdenovo-Trans	
Total number of filtered reads	553,566,274
Total number of contigs	437,776
N50 length	1304
Total base pairs	233,054,090
Secondary Assembly—CAP3 and Filtering	
Total number of scaffolds and contigs	56,873
N50 length	1558
Total base pairs	58,174,765

2.2. Sequence Annotation of the Phalaris Transcriptome

Following the initial assembly, a targeted comparison was made to both the *B. distachyon* and rice gene complements. A total of 31,771 of the phalaris reference sequences generated a significant match to 17,573 *B. distachyon* CDSs (coding DNA sequences), while comparison to CDSs of rice (which is taxonomically more distant from phalaris) only identified 23,513 sequences matching 14,200 rice genes. A total of 20,860 of phalaris sequences identified significant matches to both Poaceae model genomes.

The phalaris unigene set was assigned gene ontology (GO) terms based on sequence similarity to the Nr databases. BLAST searches showed highest similarity to rice, followed by maize (*Zea mays* L.), *Aegilops tauschii and Brachypodium distachyon* (Figure 1), with 72.5% of transcripts (41,286) being allocated at least one GO term. Within this group, assignments to the biological process category was highest (42%), followed by cellular function (40%) and molecular function (18%; Figure 2). Among the biological process sub-categories, metabolic process (27%) and cellular process (23%) were prominently represented (Figure 2, Table S2), indicating that tissues used in this study were undergoing extensive metabolic activity. A moderate number of transcripts were also involved in the single-organism process (17%), biological regulation (8%), response to stimulus (7%), regulation of biological process (8%), and biogenesis and localisation (5%) categories. Under the molecular function category, catalytic activity (51%) and binding (49%) were the most common (Figure 2, Table S1). For the cellular component category, the majority of the transcripts were assigned to the cell (26%), cell part (26%), organelle (23%), and membrane (10%) categories, while much smaller proportions (<5%) were assigned to membrane part, organelle part, and macromolecular complex (Figure 2, Table S1).

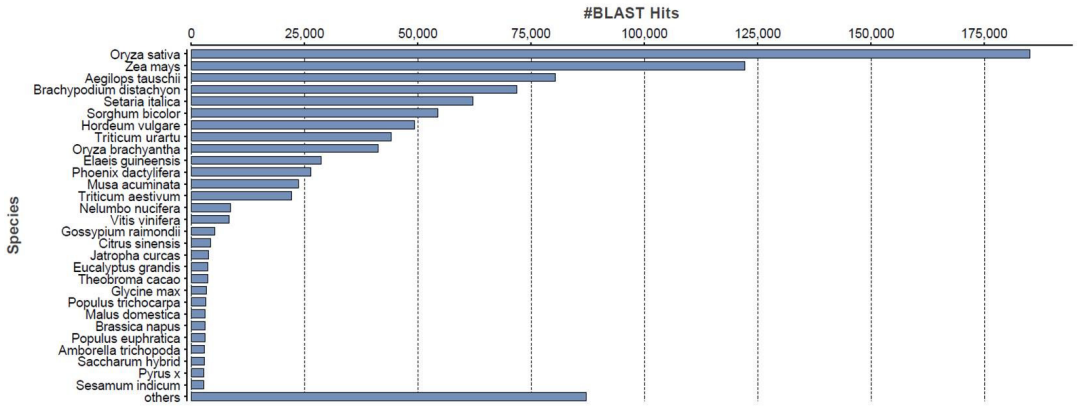


Figure 1. Species-specific distribution of highest matches for gene ontology (GO).

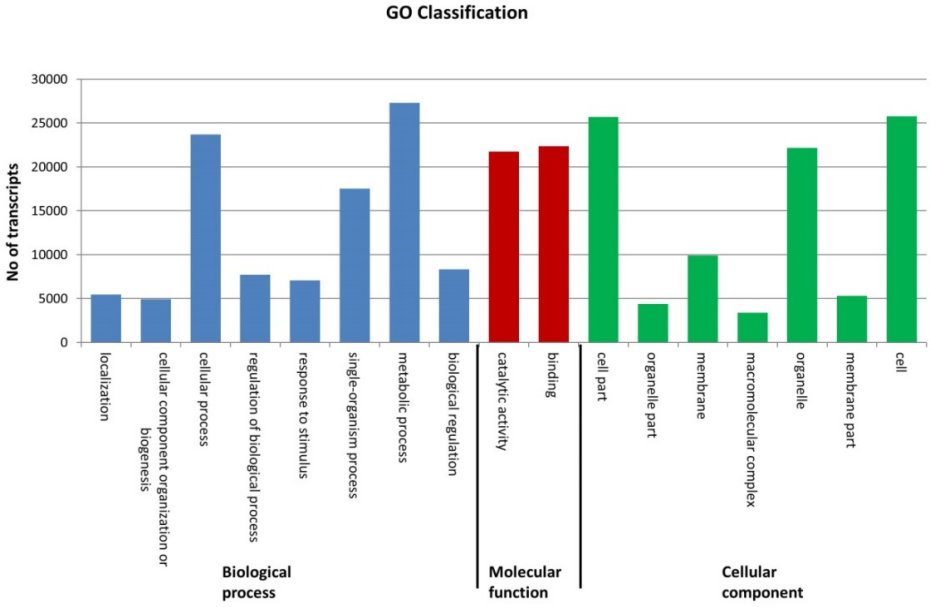


Figure 2. GO classification categories for phalaris unigenes.

2.3. Tissue-Specific Gene Expression Analysis in Phalaris

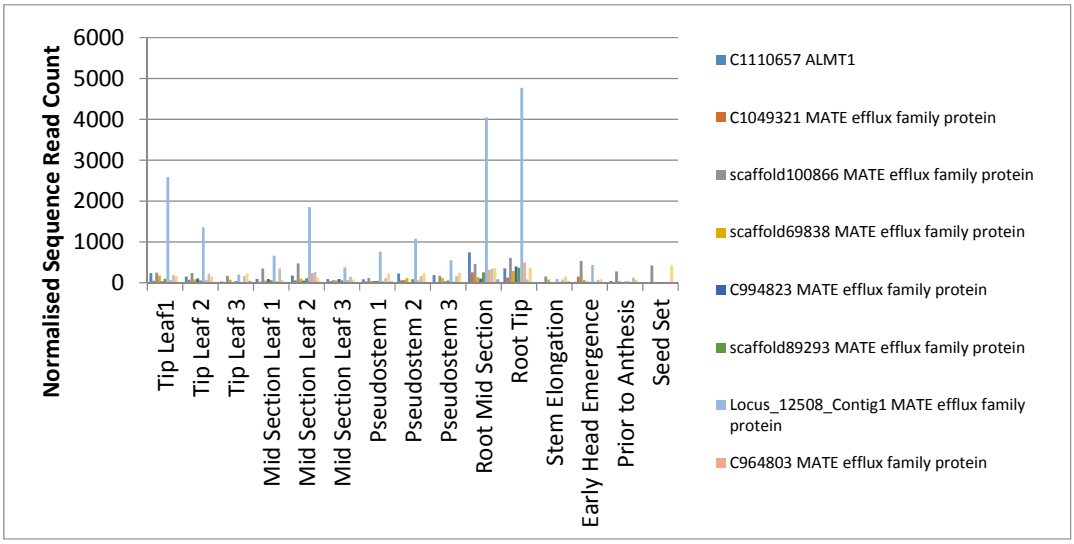
The reference unigene set was used to analyse global gene expression. Individual samples were reference aligned against the assembled sequences, and a normalised count was generated. A total of 20,593 unigenes (36.2%) were identified as being expressed in all tissue samples, and a further 18,874 (33.2%) were identified in all samples with the exception of one. From the 15 tissue-specific samples that contributed to transcriptome assembly, the lowest level of expression, or the most tissue-specific pattern, was displayed by a cohort of three unigenes that were only detected in two of the tissues. A tissue-type analysis was subsequently performed, in which data was combined into three groups: vegetative tissues, root tissues, and reproductive tissues. A total of 53,978 (94.9%) unigenes were detected in all three groups, while a further 2804 were detected in both the vegetative and reproductive groups and 64 were detected in both the vegetative and root groups. A single unigene was detected only in reproductive tissues, and no gene was detected only in root tissues. The normalised read counts per sample are presented in Table S1.

2.4. Identification of Candidate Genes for Agronomic Traits in the Phalaris Transcriptome

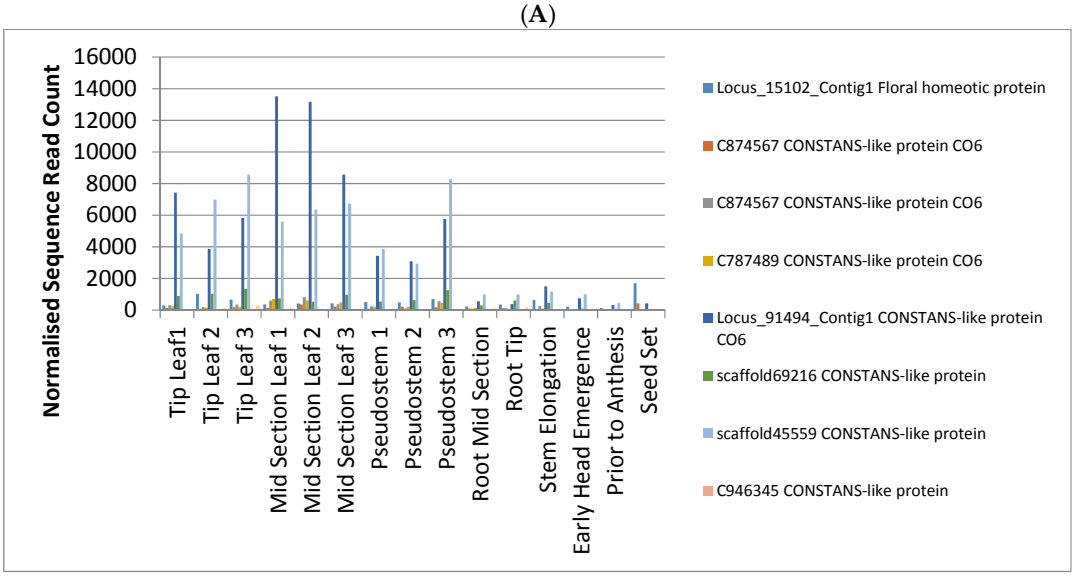
In order to exemplify the value of the dataset in terms of candidate gene identification, a text-based search of the reference unigene set was made for specific genes that related to agronomic traits of interest, on the basis of the highest recorded BLASTX match to the Uniref protein database, and associated description term. Specific candidates were identified with high confidence that relate to processes of flowering (including *CONSTANS*-like genes), tolerance to aluminium toxicity (including organic acid transporter and MATE (multidrug and toxic compound extrusion) efflux protein family genes), herbage quality (including genes for lignin biosynthesis) and toxin production (Table 2, Table S1). The relative gene expression levels for each candidate within a given class were determined, revealing tissue-specific expression patterns capable of correlation with anticipated biological activity (Figure 3, Table S1).

Table 2. Summary information on identification of candidate genes for key agronomic traits. MATE, multidrug and toxic compound extrusion.

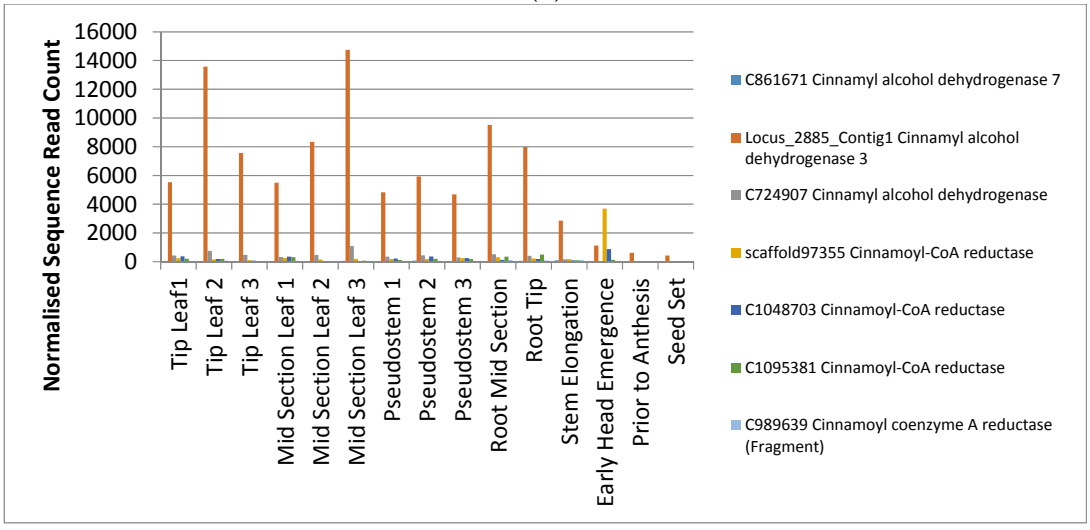
Trait Category/Common Gene Name	Uniref90 Description	*Brachypodium* BLAST Match	Rice BLAST Match
Flowering			
Q gene	Floral homeotic protein	Bradi1g03880.1	Os07g0235800
	CONSTANS-like protein CO6		
	CONSTANS-like protein CO6	Bradi3g05800.1	Os06g0654900
	CONSTANS-like protein CO6		
	CONSTANS-like protein CO6	Bradi1g31280.1	Os04g0497700
	CONSTANS-like protein	Bradi5g14600.1	Os04g0497700
	CONSTANS-like protein	Bradi5g14600.1	Os04g0497700
	CONSTANS-like protein	Bradi1g43670.1	
	CONSTANS-like protein		
	CONSTANS	Bradi1g43670.1	
Aluminium Tolerance			
*ALMT*1	*ALMT*1	Bradi5g09690.1	Os04g0417000
	MATE efflux family protein	Bradi2g17260.1	
	MATE efflux family protein	Bradi1g69120.1	Os03g0227966
	MATE efflux family protein		
	MATE efflux family protein		
	MATE efflux family protein	Bradi1g69120.1	Os03g0227966
	MATE efflux family protein	Bradi2g17260.1	Os05g0554000
	MATE efflux family protein		Os02g0676400
	Aluminum-activated malate transporter 12	Bradi3g33980.1	Os10g0572100
	Aluminum resistance transcription factor 1		
	Aluminum resistance transcription factor 1		
Toxin Biosynthesis			
	Putative Cyanogenic beta-glucosidase		
	(R)-mandelonitrile lyase 2	Bradi1g31250.1	
Herbage Digestibility			
	Cinnamyl alcohol dehydrogenase 7		
	Cinnamyl alcohol dehydrogenase 3	Bradi4g29770.1	Os09g0400400
	Cinnamyl alcohol dehydrogenase	Bradi4g29770.1	Os09g0399800
	Cinnamoyl-CoA reductase	Bradi3g36890.1	
	Cinnamoyl-CoA reductase	Bradi3g36890.1	Os08g0441500
	Cinnamoyl-CoA reductase	Bradi3g36890.1	Os08g0441500
	Cinnamoyl coenzyme A reductase (Fragment)		



(**A**)

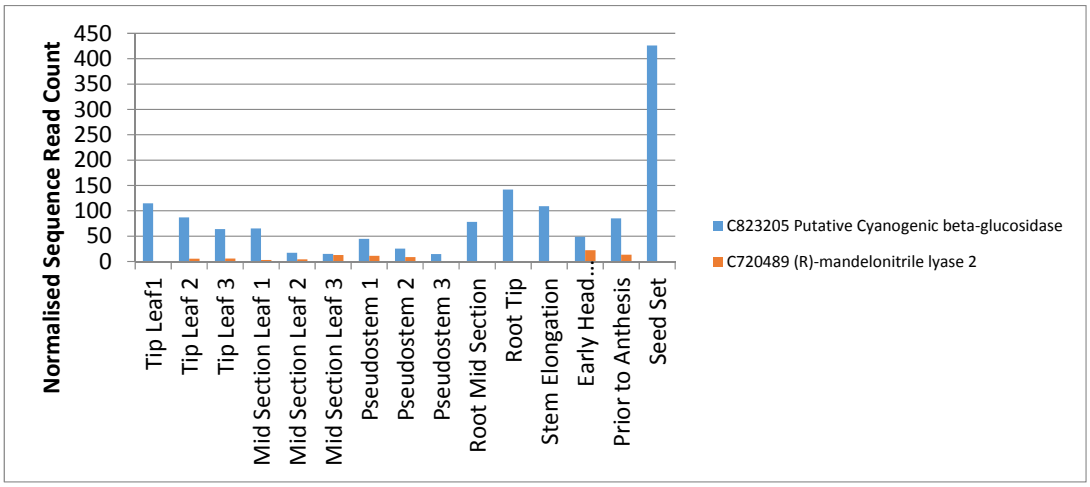


(**B**)



(**C**)

Figure 3. *Cont.*



(**D**)

Figure 3. Expression profiles of selected candidate genes for various trait-specific categories: (**A**) aluminium tolerance; (**B**) reproductive development; (**C**) lignin biosynthesis; (**D**) toxin production. The identities of sampled tissue types are shown on the y = 0 axis, while the x = 0 axis represents transcript abundance in terms of normalized sequence read count.

2.5. Identification of Long Transcripts and Molecular Phylogenetic Analysis

Although the candidate genes for specific agronomic traits were identified on the basis of highly significant BLASTX E-values, none of the gene classes was represented in their entirety by full-length transcripts, and as a consequence, detailed phylogenetic analysis through comprehensive sequence comparison to the predicted orthologues in other Poaceae genomes was not possible. However, proof-of-concept for such analysis was obtained by filtering the unigene set for transcripts of significant length, leading to the identification of 9584 protein sequences (Table S2). Of this sub-set, 9782 generated matches to *B. distachyon* and 9527 generated matches to rice. Only 35 (0.35%) of the predicted proteins failed to generate a match to either of the model species, while 9490 (96.3%) generated matches to both. The sequence set was further refined by the identification of proteins containing predicted conserved start codon locations for all three species. Of these 500 proteins, phylogenetic affinity between phalaris, rice, and *B. distachyon* was determined for a number of representative genes with plausible functional annotations: a cellulose synthase, involved in cell wall biosynthesis; a chitinase-like protein, involved in defence against fungal pathogens; and a cytokinin-*N*-glucosyltransferase, involved in plant growth regulation. In each instance, the relative affinities of the corresponding genes were consistent with taxonomic distance, being lower between phalaris and rice than between phalaris and *B. distachyon* (Figure S3). This finding demonstrates the capacity to identify and exploit comparative genomics for trait improvement in phalaris, given the ability to isolate full-length genes, either directly from transcriptome assemblies, by PCR, or from whole genome sequences.

3. Discussion

3.1. Quality of Transcriptome Assembly

The degree of completeness and quality of transcriptome assembly has important consequences for the effectiveness and accuracy of subsequent activities in gene identification, differential gene expression analysis, and molecular genetic marker development. The software tool BUSCO was recently developed for assessment of sequence assemblies, by comparison of gene sets or transcriptomes to a core set of single-copy orthologs, allowing the calculation of levels of completeness, duplication, and fragmentation. BUSCO-derived results have been demonstrated to be more consistent than those from previously used software packages such as CEGMA, and more accurate than the

commonly used N50 statistic [30], and so BUSCO has been promoted in several recent studies [28,31]. The de novo phalaris assembly in the present study was assessed as relatively complete and of high quality. Comparison to the properties of other plant transcriptomes such as those of red clover (*Trifolium pratense* L.) (C:76%[D:19%], F:13%, M:9.9%, n:956 [32]), perennial ryegrass (*Lolium perenne* L.) (C:89%[D:82%], F:5.7%, M:4.2%, n:956 [33]), *Triticum turgidum* L. (C:92%[D:54%], F:3.9%, M:3.8%, n:956; [34]), and spinach (*Spinacia oleracea* L.(C:77%[D:35%], F:12%, M:10%, n:956 [30]) demonstrated a relatively low level of duplication in the phalaris reference set, and comparable levels of missing orthologues. When compared to an additional set of de novo-assembled transcriptomes from multiple organisms, mainly of non-plant origin [30], the level of completeness and duplication is generally superior in the phalaris dataset, while fragmentation is comparable.

Transcriptome analysis has previously been performed for reed canary grass as a basis for analysis of gene expression in response to salt stress in both *P. arundinacea* and *P. aquatica* [29]. Due to the close phylogenetic relationship between these taxa, a substantial overlap of gene content between the two studies may be anticipated. However, RNA-Seq analysis of *P. arundinacea* was based on a more restricted set of tissue samples (leaves, stems, and roots under several drought and waterlogging regimes), and generated a much lower number of primary reads (494,477) and predicted genes (18,682) than in the present study. The reference set from the present study may hence prove useful for future gene expression analysis of the kind that has been performed for *P. arundinacea*.

3.2. Identification and Interpretation of Candidate Genes for Agronomic Traits

In order to demonstrate value of the transcriptome for candidate gene identification, a number of key agronomic traits for phalaris improvement were selected, and unigenes for physiological processes related to these traits were identified and patterns of expression were evaluated. Although the gene classes that were identified were not represented in their entirety by full-length sequences, the E values for BLASTX matches to known template sequences were high, suggesting that similar sequences have been identified in phalaris, and so could be isolated as complete genes in subsequent analysis of the transcriptome or genome.

Toxicity due to ionic aluminium (Al^{3+}) in soil, especially due to root growth inhibition [35], is a major environmental stress for land plants including phalaris [36]. Primary physiological mechanisms of tolerance include exclusion of Al^{3+} from access to the cytoplasm of root cells by prevention of motion across the cell membrane, and secretion of organic acids (OAs) across cell membrane into the rhizosphere, leading to chelation of Al^{3+} [37]. In grass species, several genes associated with aluminium tolerance due to OA secretion have been characterized at the sequence level, including aluminium-activated malate transporters (e.g., wheat *TaALMT*1 [38]) and members of the MATE (multidrug and toxic compound extrusion) gene family [39]. Sequences matching both gene classes were identified in the unigene set, including an ALMT1-like gene showing elevated expression in root tissue.

Seed retention is an important trait for phalaris and has also been improved through breeding selection for traits such as reduced panicle shattering character, which is due to breakage of the rachis to which multiple flowers are attached. The wheat *Q* 'super' domestication gene [40] has a primary effect on the 'free-threshing' character, but also pleiotropic effects on rachis fragility, and so may provide a candidate function for the panicle shattering trait. A *Q* locus-like sequence was identified, showing highest expression levels in the seed set developmental stage. Variation of flowering responses has also been a subject of study in phalaris [41]. The unigene set contains a number of *CONSTANS*-like genes, which are involved in floral induction in response to photoperiodic cues [42].

Nutritive quality of summer herbage is an important productivity trait for phalaris, and content and structure of lignin is a determinant of forage digestibility. Candidate genes for several of the key monolignol biosynthetic enzymes, such as cinnanoyl CoA reductase (CCR) and cinnamyl alcohol dehydrogenase (CAD), were identified in the unigene set. The putative CCR gene present in contig

C1048703 shows elevated expression during early head emergence, when herbage quality is known to decline in concert with increased lignification.

Production of phytotoxins by phalaris is a detrimental trait that has hindered broader adoption of the species. Synthesis of *N*,*N*-dimethyltyramine (DMT) is dependent on decarboxylation of L-tryptophan by an aromatic acid decarboxylase, followed by trans-methylation by the indoleethylamine-*N*-methyltransferase (INMT) enzyme. The transcriptome dataset contains a number of genes for aromatic decarboxylases, although specificity for tryptophan is not clear. Cyanogenesis has been studied in a range of plant species, and sorghum (*Sorghum bicolor* L.) contains a cyanogenic glycoside called dhurrin [43] which is degraded by the action of a β-glucosidase to regenerate *p*-hydroxymandelonitrile, followed by release of hydrogen cyanide (HCN) through the activity of a hydroxynitrile lyase [44]. The phalaris transcriptome contains a putative mandelonitrile lyase gene, expressed at low levels throughout the sampled tissues, and a putative cyanogenic glucosidase, with predominant expression during reproductive development. Varietal differences in the level of cyanogenic glycosides have been reported in phalaris [45], supporting the inference of genetic variation for this trait.

3.3. Applications to Genomics-Assisted Breeding of Phalaris

The high level of quality, based on degree of completeness and low incidence of duplication, suggests that the phalaris transcriptome developed in the present study will provide an important resource for further candidate gene identification, as a reference for assembly of transcripts in expression analysis studies and for the process of molecular genetic marker development. In the latter category, individual marker loci are likely to be of limited value, as for most other forage species, due to a relative paucity of agronomic traits under simple genetic control. As genomic selection requires a very large number of sequence polymorphism evenly distributed across the entire genome, genotyping-by-sequencing (GBS) methods provide the method of choice [46]. A GBS methodology based on sampling of the expressed component of the genome by RNA-Seq has previously been reported [47]. The unigene set described in the present study will be a valuable resource for implementation of such a method, which ideally uses alignment to a high-quality reference assembly. Due to the outbreeding breeding nature of phalaris, and the complexity of genomic architecture, predicted single nucleotide polymorphism (SNP) and homoeologous sequence variant (HSV) loci may be categorised on the basis of segregation patterns characteristic of disomic and polysomic modes of inheritance. This analysis is capable of addressing the various models for genomic structure in phalaris, and will be critically assisted by the dataset described in the present study. The resulting GBS system is anticipated to be of high value for genomics-assisted breeding of this relatively under-developed forage crop.

4. Materials and Methods

4.1. Plant Materials

A single plant from the cultivar Landmaster (identified as #D19-17) was selected as the reference genotype for transcriptome assembly. Clonal copies of the plant were produced by vegetative propagation and grown in standard potting mix in 200 mm plastic pots at 22 ± 2 °C with a photoperiod of 16/8-h (light/dark) in a glasshouse.

A total of 11 cDNA libraries were developed from vegetative tissues: tips and mid-sections of individual leaves 1, 2, and 3 from a single tiller; whole pseudostem (designated pseudostem 1); lower pseudostem (pseudostem 2); upper pseudostem (pseudostem 3); root tip and mid-root. In addition, four cDNA libraries were constructed from tissues associated with reproductive development: elongated stems prior to flowering; early emerging flowering head; whole head prior to anthesis; and whole head at seed set (Figure 4).

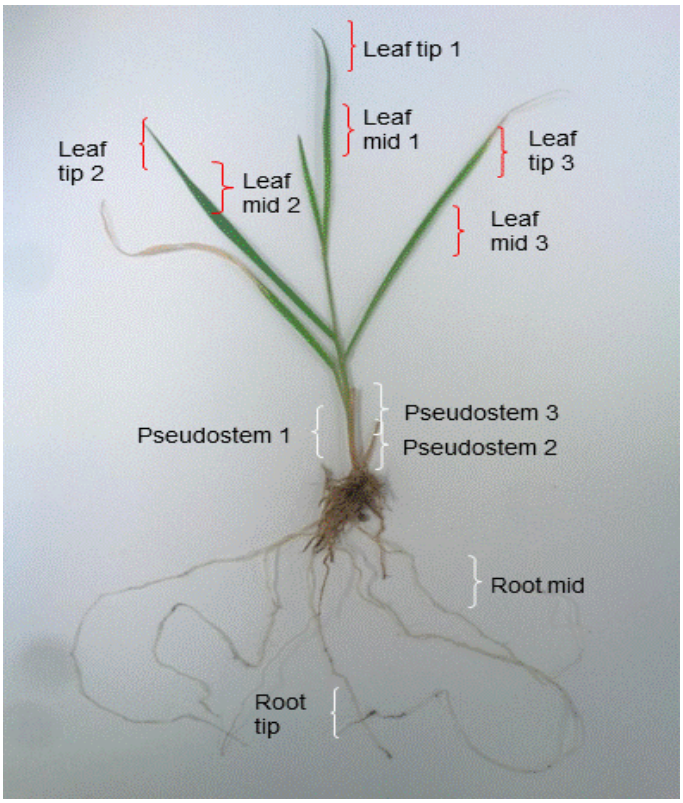


Figure 4. Pictorial representation and identification of the tissues of origin for vegetative sequencing libraries.

4.2. De Novo Transcriptome Sequence Assembly

RNA extraction was performed using the RNeasy (Qiagen, Hilden, Germany) protocol following the manufacturer's instructions. RNA-Seq libraries with an insert size of c. 350 bp were generated using the SureSelect Strand-Specific RNA Library Prep Kit and evaluated using the TapeStation 2200 platform with D1000 ScreenTape System (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions. Each RNA-Seq library was generated with a unique barcode, and an equal mass of each sequencing library was combined to create a single pooled sample for sequencing. The pooled sample was quantified using the KAPA library quantification kit (KAPA Biosystems, Boston, MA, USA). Libraries were pair-end sequenced using the HiSeq 2000 or 3000 or NextSeq 500 systems (Illumina Inc., San Diego, CA, USA).

Following fastq data generation, the raw sequence reads were filtered using a custom PERL script as well as the Cutadapt v1.4.1 software [48]. The filtered reads were then de novo assembled using SOAPdenovo-Trans v. 1.03 [49] using a k-mer of 71. Fork, bubble, and complex loci from the SOAPdenovo-Trans assembly were further combined using CAP3 assembler v 12/21/07 [50] with 95% identity to produce longer, more complete consensus sequences.

4.3. De Novo Transcriptome Sequence Annotation and Tissue-Specific Expression

The assembled transcripts were compared by use of BLASTX to the UniRef90 database [51] and the highest match was recorded. All transcripts that showed a primary significant match to a non-plant species were removed. For further annotation, assembled transcripts were also BLASTN compared to the coding DNA sequences (CDSs) of the Poaceae model species, using the *B. distachyon* (v1) and rice (*Oryza sativa* ssp. *japonica* group) (GCA_000005425.2 Build 4) databases with an E-value threshold of 10^{-10}. Assessment of the degree of completeness of the assembly was assessed by comparison to the early access Plantae reference set of orthologs using BUSCOv1.1b1 [29]. In addition,

the assembled transcripts were assessed using gene ontology (GO) terms, via the BLAST2GO PRO software program [52] with an E-value threshold of 10^{-10}, using Nr annotations.

Once the reference assembly was generated, trimmed, unassembled sequence reads from each of the individual tissue samples were aligned against the reference using the BWA software package and the mem algorithm [53]. The data was normalised on the 75th percentile, as described previously [54].

4.4. Identification of Long Transcripts and Molecular Phylogenetic Analysis

All transcripts were processed through the emboss getorf software package [55], generating sets of protein sequences with a minimum of 300 amino acids. The protein sequences that were generated were then compared to the genome assemblies of *B. distachyon* and rice using BLAST with a protein query (BLASTP) with an E value threshold of 10^{-20}. The output of the BLASTP analysis was recorded as a table with the single highest match recorded in each instance. A sub-set of 500 proteins was identified in which the BLASTP alignment started at the first amino acid in the phalaris reference set as well as both of the model species, and which were therefore amenable to comparative sequence alignment. All of the protein sequences from the three species were then aligned using Clustalw2 [56] and neighbour joining trees based on percentage identity were generated for a selected sub-set of aligned protein sequences, selected on the basis of sequence annotation.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4395/7/1/14/s1, Figure S1: Pie-chart of the distribution of the species for highest matches to phalaris unigenes, as identified from BLASTX analysis of the UniRef90 database, Figure S2: Distribution of annotated phalaris transcripts under gene ontology categories, Figure S3: Neighbour-joining trees based on percentage identity for alignment of putative phalaris, rice, and *Brachypodium* orthologues of cellulose synthasem cytokinin-*N*-glucosyltransferase and chitinase-like protein, Table S1: Complete summary information on BLAST analysis, UniRef90 classification, and gene expression analysis for phalaris transcripts. Table S2: Summary of BLASTP statistics comparing the predicted protein sequences from the phalaris dataset to reference proteins from rice and *B. distachyon*. All the sequences described in the present study have been submitted to NCBI under the Bioproject accession number PRJNA373821.

Acknowledgments: This work was supported by funding from the Victorian Department of Economic Development Jobs, Transport and Resources, the Commonwealth Scientific and Industrial Research Organisation (CSIRO), the University of Melbourne, and Meat and Livestock Australia. The authors thank Ben Cocks for helpful critical comments.

Author Contributions: N.O.I.C., J.W.F., K.F.S., R.A.C., and G.C.S. conceived and designed the experiments; R.C.B. and M.C.D. prepared plant materials, and performed RNA extraction and sequencing library preparation; N.O.I.C., S.K., and L.W.P. performed the data analysis; N.O.I.C., S.K., L.W.P., R.A.C., K.F.S., and J.W.F. assisted in drafting the manuscript; all authors read and approved the final manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The funding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Oram, R.N.; Ferreira, V.; Culvenor, R.A.; Hopkins, A.A.; Stewart, A. The first century of *Phalaris aquatica* L. cultivation and genetic improvement: A review. *Crop Pasture Sci.* **2009**, *60*, 1–15. [CrossRef]
2. Anderson, M.W.; Cunningham, P.J.; Reed, K.F.M.; Byron, A. Perennial grasses of Mediterranean origin offer advantages for central western Victorian sheep pasture. *Aust. J. Exp. Agric.* **1999**, *39*, 275–284. [CrossRef]
3. Culvenor, R.A. Breeding and use of summer-dormant grasses in Australia, with special reference to phalaris. *Crop Sci.* **2009**, *49*, 2335–2346. [CrossRef]
4. Helyar, K.R.; Anderson, A.J. Effects of lime on the growth of five species, on aluminium toxicity, and on phosphorus availability. *Aust. J. Agric. Res.* **1971**, *22*, 707–721. [CrossRef]
5. McWilliam, J.R. Selection for seed retention in *Phalaris tuberosa* L. *Aust. J. Agric. Res.* **1963**, *14*, 755–764. [CrossRef]
6. Alden, R.; Hackney, B.; Weston, L.A.; Quinn, J.C. Phalaris toxicoses in Australian livestock production systems: Prevalence, aetiology and toxicology. *J. Toxins* **2014**, *1*, 1–7.
7. Gallagher, C.H.; Koch, J.H.; Moore, R.M.; Hoffmann, H. Toxicity of *Phalaris tuberosa* for sheep. *Nature* **1964**, *204*, 542–545. [CrossRef] [PubMed]

8. Bourke, C.A. Toxins in pasture plants—Phalaris toxicity. *Anim. Prod. Aust.* **1992**, *19*, 399–402.
9. Bourke, C.A.; Carrigan, M.J. Mechanisms underlying *Phalaris aquatica* "sudden death" syndrome in sheep. *Aust. Vet. J.* **1992**, *69*, 165–167. [CrossRef] [PubMed]
10. Quintanar, A.; Castroviejo, S.; Catalán, P. Phylogeny of the trobe Aveneae (Pooideae, Poaceae) inferred from plastid *trnT-F* and nuclear ITS sequences. *Am. J. Bot.* **2007**, *94*, 1554–1569. [CrossRef] [PubMed]
11. Voshell, S.M.; Baldini, R.M.; Kumar, R.; Tatalovich, N.; Hilu, K. Canary grasses (*Phalaris*, Poaceae): Molecular phylogenetics, polyploidy and floret evolution. *Taxon* **2011**, *60*, 1306–1316.
12. Voshell, S.M.; Hilu, K.W. Canary grasses (*Phalaris*, Poaceae): Biogeography, molecular dating and the role of floret structure in dispersal. *Mol. Ecol.* **2014**, *23*, 212–214. [CrossRef] [PubMed]
13. Starling, J.I. Cytogenetic study of interspecific hybrids between *Phalaris arundinacea* and *P. tuberosa*. *Crop Sci.* **1961**, *1*, 107–111. [CrossRef]
14. Putievsky, E.; Oram, R.N.; Malafant, K. Chromosomal differentiation among ecotypes of *Phalaris aquatica* L. *Aust. J. Bot.* **1980**, *28*, 645–657. [CrossRef]
15. Stebbins, G.L. *Variation and Evolution in Plants*; Columbia University Press: New York, NY, USA; London, UK, 1950.
16. Ferreira, V.; Reynoso, L.; Szpiniak, B.; Grass, E. Cytological analysis of the *Phalaris arundinacea* (L.) x *Phalaris aquatica* (L.) amphidiploid. *Caryologia* **2002**, *55*, 151–160. [CrossRef]
17. Lavergne, S.; Muenke, N.J.; Molofsky, J. Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Ann. Bot.* **2010**, *105*, 109–116. [CrossRef] [PubMed]
18. Doležel, J.; Bartoš, J.; Voglmayr, H.; Greilhuber, J. Nuclear DNA content and genome size of trout and human. *Cytometry A* **2003**, *51*, 127–128. [CrossRef] [PubMed]
19. Akiyama, Y.; Kimura, K.; Kubota, A.; Fujimori, M.; Yamada-Akiyama, H.; Takahara, Y.; Ueyama, Y. Comparison of genome size in reed canarygrass (*Phalaris arundinacea* L.) exotic and putative native Japanese genotypes by flow cytometry. *Jpn. Agric. Res. Q.* **2015**, *49*, 345–350. [CrossRef]
20. Gutierrez-Gonzalez, J.J.; Tu, Z.J.; Garvin, D.F. Analysis and annotation of the hexaploid oat seed transcriptome. *BMC Genom.* **2013**, *14*, 471. [CrossRef] [PubMed]
21. Kaur, S.; Francki, M.G.; Forster, J.W. Identification, characterisation and interpretation of single nucleotide sequence variation in allopolyploid crop plant species. *Plant Biotechnol. J.* **2012**, *10*, 125–138. [CrossRef] [PubMed]
22. Mian, M.A.R.; Zwonitzer, J.C.; Chen, Y.; Saha, M.C.; Hopkins, A.A. AFLP diversity within and among hardinggrass populations. *Crop Sci.* **2005**, *45*, 2591–2597. [CrossRef]
23. Hayes, B.J.; Cogan, N.O.I.; Pembleton, L.W.; Goddard, M.E.; Wang, J.; Spangenberg, G.C.; Forster, J.W. Prospects for genomic selection in forage plant species. *Plant Breed.* **2013**, *132*, 133–143. [CrossRef]
24. Forster, J.W.; Hand, M.L.; Cogan, N.O.I.; Hayes, B.; Spangenberg, G.C.; Smith, K.F. Resources and strategies for implementation of genomic selection in breeding of forage species. *Crop Pasture Sci.* **2014**, *65*, 1238–1247. [CrossRef]
25. Garg, R.; Jain, M. RNA-Seq for transcriptome analysis in non-model plants. *Methods Mol. Biol.* **2013**, *1069*, 43–58.
26. Zonneveld, B.J.M.; Leitch, I.J.; Bennett, M.D. First nuclear DNA amounts in more than 300 angiosperms. *Ann. Bot.* **2005**, *96*, 229–244. [CrossRef] [PubMed]
27. Vogel, J.P.; Garvin, D.F.; Mockler, T.C.; Schmutz, J.; Rokhsar, D.; Bevan, M.W.; Barry, K.; Lucas, S.; Harmon-Smith, M.; Lail, K. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **2010**, *463*, 763–768. [CrossRef] [PubMed]
28. Moreton, J.; Izquierdo, A.; Emes, R.D. Assembly, assessment and availability of de novo generated eukaryotic transcriptomes. *Front. Genet.* **2016**, *6*, 361. [CrossRef] [PubMed]
29. Haiminen, N.; Klaas, M.; Zhou, Z.; Utro, F.; Cormican, P.; Didion, T.; Sig Jensen, C.; Mason, C.E.; Barth, S.; Parida, L. Comparative exomics of *Phalaris* cultivars under salt stress. *BMC Genom.* **2014**, *15* (Suppl. 6), S18. [CrossRef] [PubMed]
30. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef] [PubMed]

31. Xu, C.; Jiao, C.; Zheng, Y.; Sun, H.; Liu, W.; Cai, X.; Wang, X.; Liu, S.; Xu, Y.; Mou, B.; et al. De novo and comparative transcriptome analysis of cultivated and wild spinach. *Sci. Rep.* **2015**, *5*, 17706. [CrossRef] [PubMed]
32. Yates, S.A.; Swain, M.T.; Hegarty, M.J.; Chernukin, I.; Lowe, M.; Allison, G.G.; Ruttink, T.; Abberton, M.T.; Jenkins, G.; Skøt, L. De novo assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. *BMC Genom.* **2014**, *15*, 453. [CrossRef] [PubMed]
33. Farrell, J.D.; Byrne, S.; Paina, C.; Asp, T. De novo assembly of the perennial ryegrass transcriptome using RNA-Seq strategy. *PLoS ONE* **2014**, *9*, e103567. [CrossRef] [PubMed]
34. Krasileva, K.V.; Buffalo, V.; Bailey, P.; Pearce, S.; Ayling, S.; Tabbita, F.; Soria, M.; Wang, S.; Akhunov, E.; Uauy, C.; et al. Separating homeologs by phasing in the tetraploid wheat transcriptome. *Genome Biol.* **2013**, *14*, R66. [CrossRef] [PubMed]
35. Delhaize, E.; Ryan, P.R. Aluminium toxicity and tolerance in plants. *Plant Phys.* **1995**, *107*, 315–321. [CrossRef]
36. Culvenor, R.A.; Oram, R.N.; Wood, J.T. Inheritance of aluminium tolerance in *Phalaris aquatica* L. *Aust. J. Agric. Res.* **1986**, *37*, 397–408. [CrossRef]
37. Inostroza-Blancheteau, C.; Soto, B.; Ibáñez, C.; Ulloa, P.; Aquea, F.; Arce-Johnson, P.; Reyes-Diaz, M. Mapping aluminium tolerance in cereals: A tool available for crop breeding. *Eur. J. Biotechnol.* **2010**, *13*, 4.
38. Sasaki, T.; Yamamoto, Y.; Ezaki, B.; Katsuhara, M.; Sung, J.U.; Ryan, P.R.; Delhaize, E.; Matsumoto, H. A wheat gene encoding an aluminium-activated malate transporter. *Plant J.* **2004**, *37*, 645–653. [CrossRef] [PubMed]
39. Furukawa, J.; Yamaji, N.; Wang, H.; Mitani, N.; Murata, Y.; Sato, K.; Katsuhara, M.; Takeda, K.; Ma, J.-F. An aluminium-activated citrate transporter of barley. *Plant Cell Phys.* **2007**, *48*, 1081–1091. [CrossRef] [PubMed]
40. Simons, K.J.; Fellers, J.P.; Trick, H.N.; Zhang, Z.; Tai, Y.-S.; Gill, B.S.; Faris, J.D. Molecular characterisation of the major wheat domestication gene *Q*. *Genetics* **2006**, *172*, 547–555. [CrossRef] [PubMed]
41. Cooper, J.P.; McWilliam, J.R. Climatic variation in forage grasses. II. Germination, flowering and leaf development in Mediterranean populations of *Phalaris tuberosa*. *J. Appl. Ecol.* **1966**, *3*, 191–212. [CrossRef]
42. Suárez-López, P.; Wheatley, K.; Robson, F.; Onouichi, H.; Valverde, F.; Coupland, G. *CONSTANS* mediates between the circadian clock and the control of flowering in *Arabidopsis*. *Nature* **2001**, *410*, 1116–1120. [CrossRef] [PubMed]
43. Busk, P.K.; Møller, B.L. Dhurrin synthesis in sorghum is regulated at the transcriptional level and induced by nitrogen fertilisation in older plants. *Plant Phys.* **2002**, *129*, 1222–1231. [CrossRef] [PubMed]
44. Cheeke, P.R. Endogenous toxins and mycotoxins in forage grasses and their effects on livestock. *J. Anim. Sci.* **1995**, *73*, 909–918. [CrossRef] [PubMed]
45. Oram, R.N.; Edlington, J.P. Breeding non-toxic phalaris (*Phalaris aquatica* L.). In Proceedings of the 8th Australian Agronomy Conference, University of Southern Queensland, Toowoomba, Queensland, Australia, 30 January–2 February 1996; Asghar, M., Ed.; Australian Society of Agronomy Carlton: Victoria, Australia, 1996. Available online: http://www.regional.org.au/au/asa/1996/contributed/450oram.htm (accessed on 13 February 2017).
46. He, J.; Zhao, X.; Laroche, A.; Lu, Z.-X.; Liu, H.; Li, Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* **2014**, *5*, 484. [CrossRef] [PubMed]
47. Harper, A.L.; Trick, M.; Higgins, J.; Fraser, F.; Clissold, L.; Wells, R.; Hattori, C.; Werner, P.; Bancroft, I. Associative transcriptomics of traits in the polyploidy crop species *Brassica napus*. *Nat. Biotechnol.* **2012**, *30*, 798–802. [CrossRef] [PubMed]
48. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **2011**, *17*, 10–12. [CrossRef]
49. Xie, Y.; Wu, G.; Tang, J.; Luo, R.; Patterson, J.; Liu, S.; Wang, J. SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **2014**, *30*, 1660–1666. [CrossRef] [PubMed]
50. Huang, X.; Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **1999**, *9*, 868–877. [CrossRef] [PubMed]

51. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
52. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef] [PubMed]
53. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.
54. Sudheesh, S.; Sawbridge, T.I.; Cogan, N.O.I.; Kennedy, P.; Forster, J.W.; Kaur, S. De novo assembly and characterisation of the field pea transcriptome using RNA-Seq. *BMC Genom.* **2015**, *16*, 611. [CrossRef] [PubMed]
55. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
56. Larkin, M.A.; Blackshields, G.; Brown, N.P.; Chenna, R.; McGettigan, P.A.; McWilliam, H.; Valentin, F.; Wallace, I.M.; Wilm, A.; Lopez, R.; et al. ClustalW and ClustalX Version 2. *Bioinformatics* **2007**, *23*, 2947–2948. [CrossRef] [PubMed]