the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) **102,** 493–506 doi: 10.1111/tpj.14641

Genome-wide sequence information reveals recurrent hybridization among diploid wheat wild relatives

Nadine Bernhardt[1,*,†,‡], Jonathan Brassac[1,†], Xue Dong[2,3], Eva-Maria Willing[2], C. Hart Poskar[1], Benjamin Kilian[1,4] and Frank R. Blattner[1,5,*]

[1]*Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), 06466 Gatersleben, Germany,*
[2]*Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany,*
[3]*Plant Germplasm and Genomics Centre, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, 650201 Kunming, Yunnan, China,*
[4]*Global Crop Diversity Trust, 53113 Bonn, Germany, and*
[5]*German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, 04103 Leipzig, Germany*

Received 1 July 2019; revised 13 November 2019; accepted 28 November 2019; published online 10 December 2019.

*For correspondence (e-mails nadine.bernhardt@julius-kuehn.de; blattner@ipk-gatersleben.de).
†These authors contributed equally to the work.
‡Present address: Julius Kühn-Institut, 06484 Quedlinburg, Germany

SUMMARY

Many conflicting hypotheses regarding the relationships among crops and wild species closely related to wheat (the genera *Aegilops*, *Amblyopyrum*, and *Triticum*) have been postulated. The contribution of hybridization to the evolution of these taxa is intensely discussed. To determine possible causes for this, and provide a phylogeny of the diploid taxa based on genome-wide sequence information, independent data were obtained from genotyping-by-sequencing and a target-enrichment experiment that returned 244 low-copy nuclear loci. The data were analyzed using Bayesian, likelihood and coalescent-based methods. *D* statistics were used to test if incomplete lineage sorting alone or together with hybridization is the source for incongruent gene trees. Here we present the phylogeny of all diploid species of the wheat wild relatives. We hypothesize that most of the wheat-group species were shaped by a primordial homoploid hybrid speciation event involving the ancestral *Triticum* and *Am. muticum* lineages to form all other species except *Ae. speltoides*. This hybridization event was followed by multiple introgressions affecting all taxa except *Triticum*. Mostly progenitors of the extant species were involved in these processes, while recent interspecific gene flow seems insignificant. The composite nature of many genomes of wheat-group taxa results in complicated patterns of diploid contributions when these lineages are involved in polyploid formation, which is, for example, the case for tetraploid and hexaploid wheats. Our analysis provides phylogenetic relationships and a testable hypothesis for the genome compositions in the basic evolutionary units within the wheat group of Triticeae.

Keywords: phylogenomics, hybridization, introgression, *Triticum*, *Aegilops*, *Amblyopyrum*, crop wild relatives, genotyping-by-sequencing, nuclear single-copy genes, target enrichment.

INTRODUCTION

Different molecular marker types resulted in widely incongruent hypotheses of relationships for the species belonging to the wheat wild relatives (WWR) of the grass tribe Triticeae (Mason-Gamer and Kellogg, 1996; Escobar *et al.*, 2011; Bernhardt, 2015; Glémin *et al.*, 2019), that is the genera *Aegilops*, *Amblyopyrum*, and *Triticum* (van Slageren, 1994; Kilian *et al.*, 2011). Thus, despite their economic importance both as crops and as wild species contributing to the continued improvement of wheat, no comprehensive and generally agreed phylogeny for these species is currently available. This hampers the understanding of the evolution of morphological, physiological, and genetic traits, the biogeography of the species and their environmental adaptation, polyploid formation, speciation, and ultimately the search for useful alleles for plant breeding.

Hybridization is an important evolutionary process (Mallet *et al.*, 2016). It describes the crossing of individuals belonging to different species. On the homoploid level, that is if no whole-genome duplication is involved, hybridization results in first generation (F_1) offspring that possesses half of the genome of each of its parents. If this F_1 generation becomes reproductively isolated from its parents and evolves into a new species the process is termed homoploid hybrid speciation. If over time repeated backcrossing with one parent dilutes the contribution of the second parent this process is called introgression and means that genomic material (nuclear, chloroplast or mitochondrial DNA) can cross species borders. In contrast, incomplete lineage sorting (ILS) describes the process in which, during speciation, DNA polymorphisms that occur in an ancestral taxon are stochastically passed on to daughter taxa. Depending on the allele composition in individuals at certain genomic loci, phylogenetic analyses can arrive at different species relationships when different individuals and/or loci are analyzed (Maddison, 1997). As ILS mostly depends on population sizes together with mutation rates, the process of lineage sorting can be modelled in a coalescent framework (Kingman, 1982). Although it is not always possible to discern hybridization from ILS, multilocus coalescent analyses including multiple individuals per species can, in part, overcome this problem (Green *et al.*, 2010a; Durand *et al.*, 2011; Pease and Hahn, 2015; Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016; Wen and Nakhleh, 2018; Chao *et al.*, 2018).

The recent advent of genomic data for *T. aestivum* (International Wheat Genome Sequencing Consortium, 2014, 2018), an allohexaploid with three subgenomes (termed **A**, **B**, and **D**), and the related diploid species *Ae. tauschii* (Jia *et al.*, 2013; Luo *et al.*, 2013, 2017) and *T. urartu* (Ling *et al.*, 2013), allows for the comparative analyses of genome structure and gene content. Marcussen *et al.* (2014), when analyzing relationships among the three subgenomes of wheat, postulated that the **D**-genome lineage occurring in *Ae. tauschii* is of homoploid hybrid origin involving the ancestors of the **A** (occurring in *T. urartu*) and **B**-genomes (similar to *Ae. speltoides*). This finding spurred a discussion regarding the hybrid origin of *Ae. tauschii* (Sandve *et al.*, 2015; Li *et al.*, 2015a,b). El Baidouri *et al.* (2017) analyzed sequences of homeologous genes and transposable elements derived from *T. aestivum* (**ABD**), tetraploid *T. durum* (**AB**), *T. urartu* (**A**), *Ae. speltoides* (**B**), and *Ae. tauschii* (**D**). They deduced that, about six million years ago (Mya), an ancestral **D**-genome introgressed into a homoploid hybrid of the ancestral **A**- and **B**-genomes. The ancestral **D**-genome became extinct sometime later. Today's **D**-genome, occurring in diploid *Ae. tauschii* and as one subgenome in *T. aestivum* and other polyploid species of *Aegilops*, is, therefore, a hybrid genome combining three genomes (El Baidouri *et al.*, 2017). As the **B**-genome of polyploid wheat is different from its closest extant relative *Ae. speltoides*, they assumed that the **B**-genome itself might also have been introgressed by species of the **S**-genome group of *Aegilops* sect. *Sitopsis*. Recently, Glémin *et al.* (2019) developed a new framework to investigate hybridizations. Based on transcriptome data for all species, they proposed a complex scenario of hybridizations identifying *Am. muticum* (**T**), instead of *Ae. speltoides* (**B**), as an ancestor of the **D**-genome lineage and at least two more hybridization events.

In Triticeae it is generally agreed that the diploid taxa and cytotypes form the basic units of evolution and are involved in different combinations in the formation of polyploid taxa (Kellogg, 2015). Polyploids occur mostly as allopolyploid taxa, combining the genomes of different parental species after hybridization and whole-genome duplication (WGD). Except for Glémin *et al.* (2019) and Huynh *et al.* (2019), the recent studies of the evolution of wheat included only a few species and mostly single individuals (although with huge amount of genome data) of WWR. Here we describe the analyses of two genome-wide datasets obtained for all diploid species of *Aegilops*, *Amblyopyrum*, and *Triticum*, and always multiple individuals per taxon to improve the understanding of evolutionary relationships in the wheat group. This work employs DNA sequences of 244 nuclear low-copy genes uniformly distributed among all seven chromosomes of the taxa. These were obtained through a set of gene-specific hybridization probes used to enrich the target loci before next-generation sequencing (Hyb-seq; Weitemier *et al.*, 2014). Based on this set of genes, species relationships were calculated using diverse phylogenetic algorithms. In addition, genome-wide single-nucleotide polymorphism (SNP) data were obtained through genotyping-by-sequencing (GBS; Elshire *et al.*, 2011). Both datasets were compared for signals of directed introgression and hybridization. Our results provide species relationships within the wheat-group taxa, and lead to new hypotheses on far-reaching hybridization and introgression influencing the evolutionary origins and composition of all extant basic diploid genomes in this species group.

RESULTS AND DISCUSSION

Sequence assembly of the target-enriched loci

Loci for target enrichment were selected via the comparison of available genome information from different Poaceae like *Brachypodium distachyon*, rice, and sorghum, barley and wheat (Vogel *et al.*, 2010; Matsumoto *et al.*, 2011; Mayer *et al.*, 2011), aiming for orthologous loci with an even distribution on the genome (Materials and Methods S1). Our design of capture probes was finally based on 451 loci evenly distributed over the **A-**, **B-**, and **D**-genomes of *T. aestivum* (Supporting Information Table S1 and Figure S1).

Target enrichment and Illumina sequencing resulted in 140 million raw reads and 116 million reads after quality filtering. On average, 6% of the reads mapped to the chloroplast genome. Of the 451 loci, 25 (5%) were not sufficiently captured (i.e. not captured in most taxa) and were excluded from further analyses. The capture efficiency was usually taxon/accession independent, indicating no (strong) influence of probe design on the capture efficiency (Tables S1 and S2). The sequences retrieved for the 426 well captured nuclear loci were combined into multiple sequence alignments. Visual inspection of these alignments often showed genus- or species-specific patterns of ambiguous positions indicating the occurrence of different alleles or paralogues. Allelic diversity is assumed to be much lower than 1%. This threshold was set based on a comparison with Jakob *et al.* (2014) that reported an allelic diversity clearly lower than 1% for the analysis of six single-copy loci of large populations of *Hordeum vulgare* subsp. *spontaneum*. Thus, single-copy loci of heterozygous individuals can be expected to show noticeably less than 1% of ambiguous positions in assembled sequences, while a higher percentage indicates the presence of paralogous copies. Moreover, since sequenced accessions within a species mainly share the same combinations of polymorphic positions, this too points to the existence of paralogous gene copies for a locus, either functional or as pseudogenes, rather than to heterozygous loci. The proportion of ambiguous positions per accession and locus was estimated (Table S3). An average of more than 1% of ambiguous sites in more than five species was detected for 62 (~15%) captured loci. These loci were considered as mainly multicopy and excluded from further analyses. Moreover, very short or not variable loci were excluded. The median of the mean coverage for the 244 remaining loci (Dataset S1) was 25X. Large deviations in the mean coverage resulted from the actually achieved sequencing depth (Table S4a). The loci used for phylogenetic inference had, on average, a length of 2278 bp, 43% of non-variable sites and a pairwise identity of 88% (Table S4b). Concatenation of the 244 nuclear loci in a supermatrix resulted in an alignment with a total length of 555 543 bp.

Phylogenies based on target-enrichment data

Supermatrix approach. The first step of our analysis procedure was to use DNA sequences of nuclear genes enriched through hybridization probes for Illumina sequencing to infer phylogenetic relationships from quality filtered alignments. In addition to the wheat group taxa, we included four diploid species as outgroups representing the barley genus *Hordeum* (Table S5). Maximum likelihood (ML) and Bayesian phylogenetic inference (BI) of the concatenated DNA sequences of all loci (i.e. creating a supermatrix with 555 543 alignment positions) resulted in the phylogenetic relationships provided in Figure S2. In this tree *Ae. speltoides* and *Am. muticum* form a clade that is sister to all other ingroup taxa analyzed. Within the latter, *Triticum* is a sister group of the remainder of *Aegilops* species. When analyzing the same dataset with maximum parsimony (MP), *Triticum* and *Ae. speltoides*/*Am. muticum* exchange their respective positions in the phylogenetic tree (Figure S3).

Coalescent-based phylogenetic inference. As data concatenation could potentially result in strong support for wrong species relationships (Xi *et al.*, 2015), gene trees were used to infer a coalescent-based species tree. Individual ML gene trees were used as input for ASTRAL (Mirarab *et al.*, 2014; Chao *et al.*, 2018), which models ILS under the multispecies coalescent (MSC) model (Degnan and Rosenberg, 2009) to deduce species relationships. The resulting phylogeny places *Triticum* as sister to *Amblyopyrum* and all *Aegilops* species (Figures 1a and S4), a topology similar to the one found by MP analysis of the supermatrix (Figure S3). *Aegilops markgrafii*/*Ae. umbellulata* form a clade with *Ae. comosa*/*Ae. uniaristata* (clade **CUMN**), although with very low statistical support (Figure 1a).

While all 244 individual ML gene trees were in conflict with each other and accessions of the same species may be widely scattered in single topologies (Dataset S2), all supermatrix phylogenetic approaches (Figures S2 and S3), the ASTRAL analysis (Figure S4), and the unrooted network obtained via SPLITSTREE (Figure S5) revealed species to be monophyletic. We, therefore, concluded that ongoing gene flow between species is not significantly impacting the data and extant species can be considered as units.

Low support values in the ASTRAL tree (Figures 1a and S4) correspond to branches with topological differences when compared with the supermatrix phylogenies, indicating conflicting phylogenetic signal. The degree of gene tree/species tree conflict was investigated in detail using PHYPARTS (Smith *et al.*, 2015), as it could also stem from hybridization/introgression instead of ILS. For most clades comprising several species, no major alternative to the ASTRAL topology could be identified (Figure S6). However, the clades of **CUMN** and **DS** present in the ASTRAL tree were supported by only seven and 20 out of 244 gene trees, respectively. For the former clade, there were five alternative topologies found to be more frequent and involving members of the **CUMN** clade together with either *Ae. tauschii* (**D**) or the *Triticum* species (**A**): **UD** with 14 supporting topologies, **CD** 12, **MND** 10, **AU** 9, and **ND** 8. For **DS**, there were 20 alternative topologies that grouped *Ae. speltoides* (**B**) instead of *Ae. tauschii* (**D**) together with sect. *Sitopsis* (**S**).

In multilocus analyses, *Ae. speltoides* always forms a moderately supported clade with *Am. muticum* (**T**), and, as in previous studies (e.g. Petersen *et al.*, 2006; Li *et al.*, 2015a), it was always clearly separated from the other species of *Aegilops* sect. *Sitopsis* (**S**), as well as from the remaining *Aegilops* species. In the following, we will use sect. *Sitopsis** to indicate that we refer to the **S**-genome group of sect. *Sitopsis* excluding *Ae. speltoides* (**B**) that

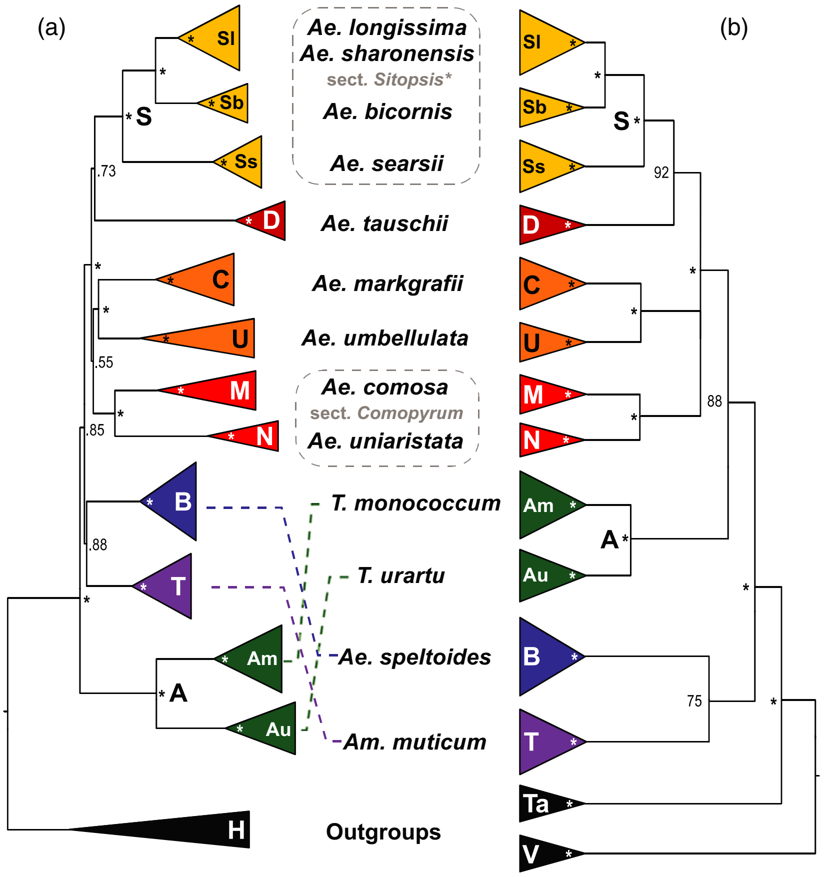

Figure 1. Comparison of coalescent-based phylogenetic trees for the diploid wheat wild relatives. Triticeae-specific genome designations are provided for the respective clades. Fully supported nodes are indicated by asterisks. (a) Schematic representation of the multispecies coalescent tree calculated from separate maximum likelihood gene trees of 244 target-enriched low-copy loci using Astral. Numbers at nodes depict local posterior probabilities. (b) Consensus cladogram derived from a Tetrad analysis of GBS data. Numbers along branches are bootstrap support values (%).

was earlier placed within this group (van Slageren, 1994). *Aegilops tauschii* (**D**), although assumed to be either a homoploid hybrid between the **A**- and **B**-genome lineages (Marcussen *et al*., 2014; Sandve *et al*., 2015; Huynh *et al*., 2019) or the **A**-, **B**-, and **D**-genome ancestors (El Baidouri *et al*., 2017), results in all our analyses as sisters of sect. *Sitopsis**. This indicates that an **S**-genome progenitor may have played a role in its formation. This close relationship has not been previously postulated, although Marcussen *et al.* (2014) used sequences of the **S**-genome species *Ae. sharonensis* (International Wheat Genome Sequencing Consortium, 2014). However, they excluded these from additional analyses, as they assumed *Ae. sharonensis* itself to be a hybrid involving the **B**-genome lineage. Our data show that not only *Ae. sharonensis* is closely related to *Ae. tauschii* but that shared genome parts most probably involved the entire sect. *Sitopsis**. Although the relationship to the **B**-genome was not found in this initial analysis, this clearly indicates a more complex evolutionary history of the *Ae. tauschii* genome and perhaps also that of sect. *Sitopsis** in comparison with previous hypotheses.

Although the discordant topologies revealed by PhyParts are potentially better resolved by modelling ILS, they may also result from past hybridizations or gene flow among species. Both processes would violate the assumption of the coalescent analysis that only ILS contributes to deviation of gene tree topologies. Therefore, our sequence data were further analyzed to uncover past hybridization and introgression events.

Network approach based on gene tree topologies from target-enrichment data. Even though methods to infer phylogenetic networks are under constant development (e.g. Yu *et al*., 2011; Yu and Nakhleh, 2015; Solís-Lemus and Ané, 2016; Wen *et al*., 2016; Wen and Nakhleh, 2018; Chi *et al*., 2018), the analysis of multiple loci, individuals, and species while modelling ILS and reticulations remains computationally expensive (Hejase and Liu, 2016; Wen *et al*., 2018). Thus, resource demanding methods such as full ML or Bayesian inference (Yu *et al*., 2014; Wen and Nakhleh, 2018) failed to infer networks from our entire sequence data. We therefore used different strategies of data partitioning by reducing the number of individuals or loci. However, these approaches gave incoherent results across replicates.

Nevertheless, we were able to obtain phylogenetic networks from the 244 gene tree topologies under the multispecies network coalescent (MSNC) using maximum pseudolikelihood as implemented in PhyloNet (Yu and Nakhleh, 2015). We allowed for zero to five reticulations

(Figure S7a–f). If no hybridization was assumed, the tree with the best log pseudolikelihood (−7 617 218) had a topology similar to the one obtained via Astral (Figures 1a and S4). However, clades poorly supported in Astral were dissolved resulting in a grade with *Triticum* as sister to the rest of the species, *Am. muticum* and *Ae. speltoides* not being monophyletic, and *Ae. comosa*/*Ae. uniaristata* and *Ae. markgrafii*/*Ae. umbellulata* not clustering together. PhyloNet also retrieved the Astral topology among the top five trees with a slightly lower log pseudolikelihood (−7 617 519). The network with four hybridization nodes (Figures 2 and S7e) was selected with the Akaike information criterion as best fit. In this network, hybridizations are nested within each other. This suggests a sequence of hybridization events, the first one involves the ancestors of *Am. muticum* and the *Triticum* clade each contributing approximately equal proportions (0.54 and 0.46, respectively) to the common ancestor of all other *Aegilops* species except *Ae. speltoides*. This confirms the scenario inferred by Glémin *et al.* (2019) identifying *Am. muticum* instead of *Ae. speltoides* (Marcussen *et al.*, 2014; Huynh *et al.*, 2019) as one of the genome donors. Sect. *Sitopsis** appears as sister to both *Ae. tauschii* and *Ae. markgrafii* and to be introgressed by *Ae. speltoides* (0.31). Finally, the *Ae. comosa*/*Ae. uniaristata* clade is sister to *Ae. markgrafii* with an additional introgression of the *Triticum* clade (0.29). However, phylogenetic networks inferred from gene tree topologies under maximum pseudolikelihood are not necessarily uniquely encoded by their system of rooted triples and this analysis may return an equivalent network to the true network (Yu and Nakhleh, 2015). In this case, the authors suggest investigating the obtained network with other methods and/or data. Here we used GBS to generate genome-wide SNP data from all taxa to evaluate this scenario.

DNA polymorphisms obtained through genotyping-by-sequencing

Sequence assembly of the GBS data. To obtain genome-wide SNP data, a two-enzyme GBS analysis (Poland *et al.*, 2012) was performed by cutting the genome with a frequent and a rare-cutting restriction enzyme followed by sequencing 100 bp of the DNA fragments directly adjacent to the rare restriction sites (Wendler *et al.*, 2014). This method was shown to target the coding parts of the genome (Schreiber *et al.*, 2019). Thus, it can be used to compare SNP patterns between species, which might, in their non-coding genome regions, already be too diverse for meaningful comparisons. As *Hordeum* and the wheat-group lineage were already separated 15 Mya (Marcussen *et al.*, 2014), their genomes have diverged substantially. Therefore, we included *Dasypyrum villosum* and *Taeniatherum caput-medusae* as outgroups. These taxa are outside the wheat group genera (Bernhardt *et al.*, 2017) but still close enough to share multiple GBS loci.

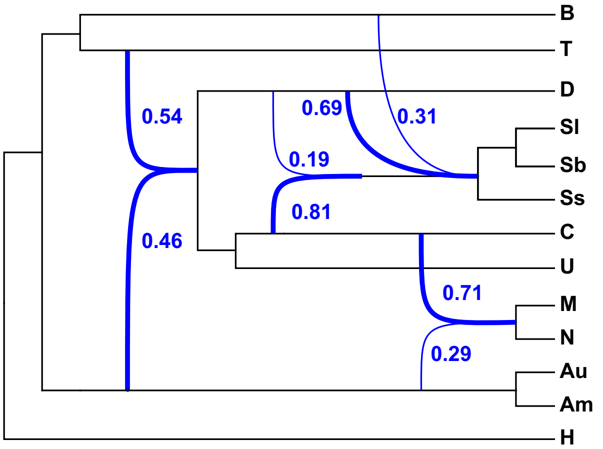

Figure 2. Phylogenetic network inferred under the multispecies network coalescent (MSNC) from the 244 gene tree topologies using maximum pseudolikelihood. The network with four reticulations was selected as best fit among zero to five hybridizations calculated with the routine InferNetwork_MPL of PhyloNet under the Akaike information criterion (see Figure S7). Reticulations are indicated by blue arcs with major contributions from species to hybrid lineages indicated by bold lines. Numbers represent estimated inheritance probabilities.

On average, 1.65 million reads per sample were obtained from Illumina sequencing. After filtering and clustering, on average, 222 185 clusters remained per sample. After consensus calling per cluster the number of loci per individual in the assembly was, on average, 21 000 (with a minimum of 8472 loci for accession AE 739 of *Ae. speltoides* and maximum of 28 469 loci for accession PI 560122 of *Am. muticum*). In total, 140 072 loci having 444 618 phylogenetic informative sites were kept for downstream analysis when it was specified that at least four individuals had to share a locus (Dataset S3 and Table S6).

GBS-based phylogenetic relationships. To analyze phylogenetic relationships based on the GBS data, we conducted an analysis in Tetrad within the ipyrad package (Eaton, 2014; https://github.com/dereneaton/ipyrad). Tetrad uses a single SNP per GBS locus and conducts quartet analyses to infer a species tree that is consistent under the MSC. The phylogenetic tree (Figures 1b and S8) supports the topology of the supermatrix tree of the target-enrichment data (Figure S2) with respect to the relative positions of *Triticum* and *Ae. speltoides*/*Am. muticum* and of the Astral tree regarding the **MN** and **UC** taxa forming together a weakly supported clade (Figure 1a). The unrooted phylogenetic network computed by SplitsTree (Figure S9) is concordant with the one for target-enrichment data (Figure S5) showing that species are monophyletic and can be considered as units for the detection of hybridization.

Even though Zhu and Nakhleh (2018) developed a method (i.e. MLE_BiMarkers) able to deal with more than 50 taxa and four hybridizations using bi-allelic markers under the maximum pseudolikelihood, we could not process our dataset in a reasonable timeframe (i.e. analyses did not finish within 30 days). We assume that the complexity of the relationships, including putative nested hybridization and introgression events (Figure 2), complicates the inference of a network from the GBS data. Nonetheless, we assessed hybrid relationships with Four- and Five-taxon *D* statistics. Those methods, based on the frequency of shared polymorphisms between taxa, are less computing intensive.

GBS-based D *statistics for the detection of hybridization and direction of introgression.* Under a neutral model of sequence evolution, and if speciation events occur in rapid succession, ILS should result in similar amounts of shared polymorphisms among species derived from a common ancestor. However, if hybridization is involved, the amount of shared alleles shifts towards the species connected through gene flow in comparison with the background signal contributed by ILS. *D* statistics, also known as the ABBA–BABA test (Green *et al*., 2010a; Durand *et al*., 2011), is able to discern hybridization from ILS by analyzing allele distribution in three taxa in comparison with an outgroup.

All Four-taxon *D* statistic tests were performed species-wise on unlinked SNPs with the routine *Dtrios* of DSUITE (Malinsky, 2019). First, *D. villosum* was set as an outgroup to test if *Ta. caput-medusae* was involved in hybridizations with any members of the WWR (Figure S10). *Taeniatherum caput-medusae* then was used as an outgroup for all following tests as no hybridization signal was found. In total, 220 tests were performed of which 64 were significant (P-value < 0.05 after Benjamini–Yekutieli correction) with *D* statistics ranging between 0.10 and 0.33 (Figure 3 and Table S7). All species were involved in potential hybridizations. The strongest signal revealed a relationship between both *Triticum* species and *Ae. markgrafii*/*Ae. umbellulata*, and to a lesser extent *Ae. tauschii* and *Ae. comosa/Ae. uniaristata*. The latter relationship is in conflict with the results from the network analysis (Figure 2) that suggested an additional introgression of *Triticum* into the ancestor of *Ae. comosa/Ae. uniaristata*. In addition, *Ae. markgrafii* showed a strong tie with the members of sect. *Sitopsis** (**S**). This analysis confirmed the strong and exclusive relationships between *Ae. speltoides* and the latter.

An extension of *D* statistics is the D_{FOIL} test (Pease and Hahn, 2015) that allows not only the detection of hybridization in the presence of ILS but also infers the direction of introgression in a five-taxon phylogeny. This analysis only accepts an alignment of five sequences, therefore we created consensus sequences for each species. D_{FOIL} tests were performed with *Ta. caput-medusae* used as the outgroup to polarize the comparisons of all species. Altogether 216 unique combinations of five taxa were tested, but only 143 tests were considered after removing tests that did not fulfil the requirements of estimated divergence times (see Experimental procedures; Pease and Hahn, 2015). On average 292 602 alignment positions (233 791–379 867) were used resulting in 6738 (952–10 354) SNP patterns that could be compared (Table S7 and Figure 4). Overall, the relationships inferred are similar to the ones identified by the ABBA–BABA test (Figure 3 and Table S6), however directions of gene flow could be inferred for nine relationships (11 tests). A large proportion of tests (42) revealed undirected patterns involving three taxa indicative of complex or ancient introgressions, or reciprocal gene flow. Evidence of introgression/hybridization was found for all species (Figure 4a–k), with a low number of significant tests involving *Ae. uniaristata* and *Ae. umbellulata* (Figure 4e–f) and a high number involving *Ae. markgrafii* and *Ae. longissima* (Figure 4g, k). This analysis confirms the close relationships between the members of sect. *Sitopsis** (**S**) and *Ae. speltoides* (**B**), but, in contrast with the network inferred with PHYLONET (Figure 2), D_{FOIL} identifies gene flow from **S** to **B** (Figure 4b). Among the members of sect. *Sitopsis**, *Ae. longissima* (**Sl**) appeared as a major introgressor of **B** but also of *Ae. comosa* (**M**), *Ae. markgrafii* (**C**), and *Ae. tauschii* (**D**) (Figure 4k). This may explain the high number of tests returning undirected signals involving those four species. The close relationship between *Triticum* species and the **CUMND** clade was confirmed, although no direction could be inferred (Figure 4c). This analysis also suggests that *Am. muticum* was affected by gene flow from *Ae. comosa* and *Ae. tauschii* (Figure 4a).

Homoploid hybrid speciation and major introgressions

In the following, we describe our hypothesis for the evolution of WWR (Figure 5). Overall, the scenario inferred is similar to the one identified by Glémin *et al.* (2019). Nonetheless, as we did not focus on identifying the progenitors of the '**D**-genome lineage', we are able to propose a more complete picture. However, as the relationships we identified are highly reticulate, there are partly alternative scenarios possible. We limit our interpretation to the most strongly supported relationships to avoid false positives (Eaton *et al*., 2015).

As our phylogenetic analyses revealed the monophyly of all species, we are certain that hybridizations and introgressions involved mainly ancestral taxa and not the extant species. Our results suggest that there are different classes of taxa, that is lineages that introgressed others, lineages that are recipients of introgressions from one or several taxa, and/or lineages that originated via homoploid hybrid speciation.

We hypothesize that most of the wheat-group species were shaped by a primordial homoploid hybrid speciation

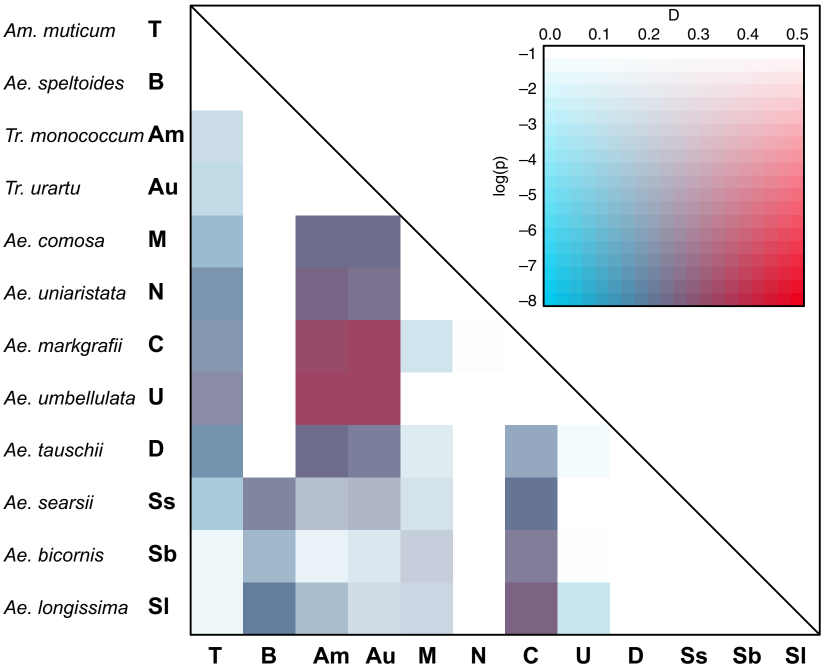

Figure 3. Heatmap summarizing Four-taxon *D* statistic tests using *Taeniatherum caput-medusae* as the outgroup. The plot is based on 220 tests. It shows the *D* statistic results and their significance for each pair of species. Red and blue indicate high and low *D* statistic values, respectively. The intensity of the colour corresponds to the *P*-value (in log scale) assessed using the block jackknife procedure and corrected with Benjamini–Yekutieli for multiple testing. All *D* statistic results are summarized in Table S7.

event, that is the *Triticum* lineage merged with the ancestor of *Am. muticum* to form all other species except *Ae. speltoides* (Figure 5, event **1**). These results highlight the pivotal role of *Am. muticum*, instead of *Ae. speltoides*, in the formation of the WWR. This hybridization event was followed by multiple introgressions affecting all taxa except *Triticum*. In contrast with Glémin *et al.* (2019), we do not find introgression of *Triticum* into *Am. muticum*, instead our results indicated that *Am. muticum* may have been introgressed by *Ae. umbellulata* or the common ancestor of the **CU**(**MND**) clade (Figures 3, 4a, S7d and Figure 5, event **2**). Previously published chloroplast phylogenies (Yamane and Kawahara, 2005; Bordbar *et al.*, 2011; Bernhardt *et al.*, 2017) support a chloroplast capture event, as the maternal lineage of *Am. muticum* does not group with *Ae. speltoides*, although both are sister taxa in nuclear phylogenies, but it shares a common ancestor with *Ae. umbellulata*.

For *Ae. speltoides* (**B**) conflicting results were obtained with either sect. *Sitopsis** (**S**) being introgressed by **B** (Figure 2) or the other way around (Figure 4b). This suggests that either reciprocal gene flow occurred between those species (Figure 5, event **3**) or that at least one of the applied methods revealed false positives. Both methods have drawbacks: phylogenetic networks obtained under maximum pseudolikelihood may not be true but rather equivalent to the true network (Yu and Nakhleh, 2015), and *D* statistics are only analyzing three or four taxa simultaneously. Nevertheless, sect. *Sitopsis**, and especially *Ae. longissima* that has been described as an outcrossing taxon (Escobar *et al.*, 2010), was repeatedly identified as an introgressor, as it exhibits relationships with all taxa except the *Triticum* lineage (Figure 4k).

Signals for involvement of the sect. *Sitopsis** genomes can be found in *Ae. comosa* (**M**) and *Ae. markgrafii* (**C**), for which a hybrid origin has been recently proposed (Danilova *et al.*, 2017). Both taxa presented patterns of introgressions different from their respective sister species *Ae. umbellulata* and *Ae. uniaristata*. These two species were involved in the least number of hybridizations. This seems to indicate that the **C** and **M** lineages diverged from their respective sister species due to minor introgressions from *Ae. longissima* or other species of the sect. *Sitopsis** (Figure 5, event **4**).

It is further suspected that *Ae. longissima* or sect. *Sitopsis**, possibly together with members of **CU**(**MN**), were involved in the formation of *Ae. tauschii*, as the complexity of the observed pattern does not resemble a simple sister-species relationship (Figures 3, 4h and S6). Indeed, chloroplast phylogenies (Yamane and Kawahara, 2005; Bernhardt *et al.*, 2017) place the maternal lineage of *Ae. tauschii* sister to the **CUMNS** clade, suggesting that one of its ancestors is an ancient, perhaps extinct (El Baidouri *et al.*, 2017), lineage. This idea is in contrast with its placement in nuclear phylogenies in which *Ae. tauschii* shows a moderately supported sister relationship to sect. *Sitopsis** or members of **CU**(**MN**) (see PhyParts analysis). Therefore, we hypothesize that the introgressions from one or possibly both of those clades resulted in the position of *Ae. tauschii* inside the clade partially depicted by the event **5** in Figure 5. Finally, due to the primordial homoploid hybrid speciation, *Ae. tauschii* displays similarities with *Triticum* (**A**) and *Am.*

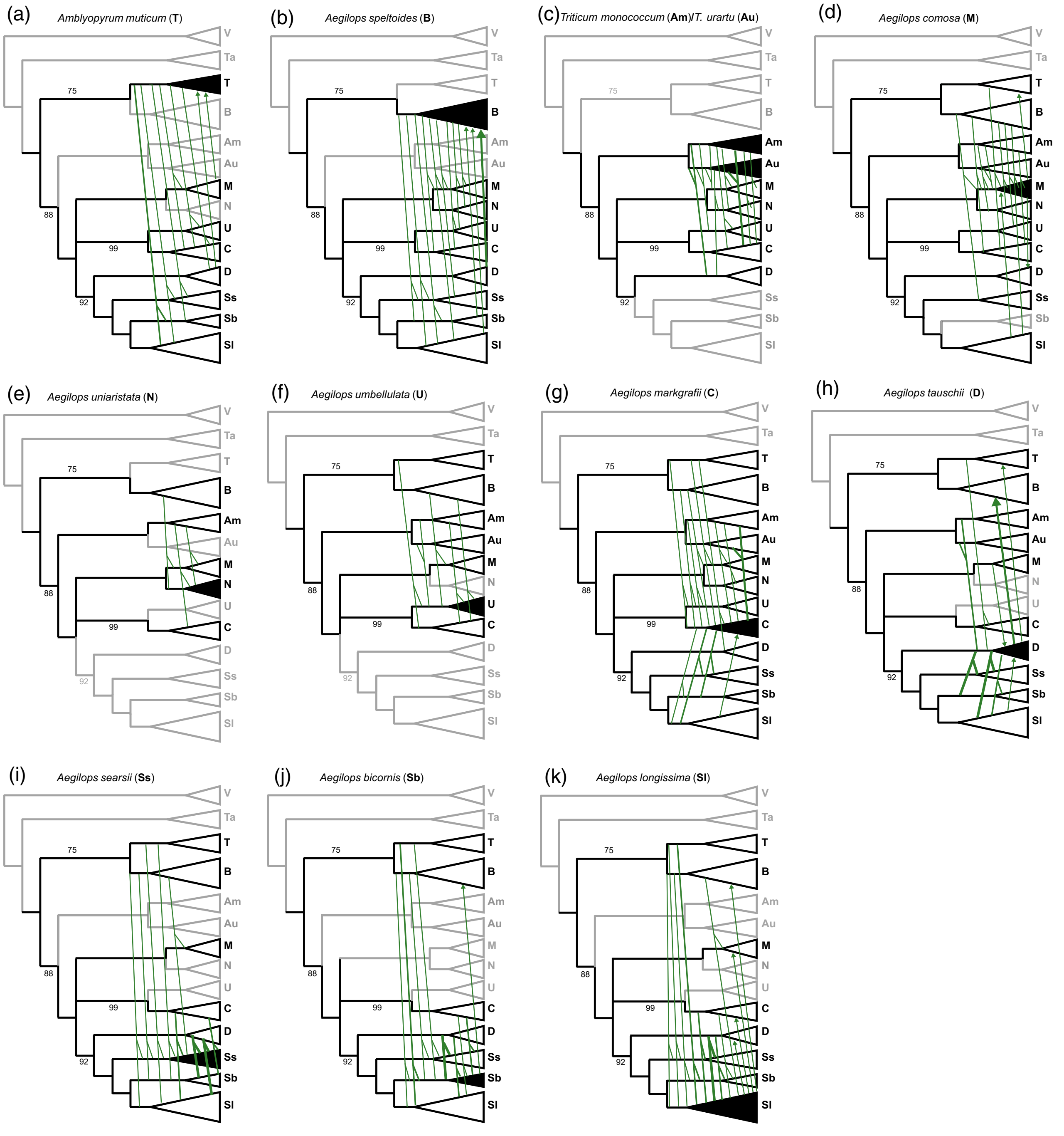

Figure 4. Representation of D_{FOIL} results for genotyping-by-sequencing data. All significant relationships after Benjamini–Yekutieli correction are shown on a modified version of the TETRAD species tree. Each tree shows all significant relationships for a focal taxon. An arrowhead indicates the direction of hybridization/introgressions between two taxa. Undirected relationships involving three taxa are shown using a branched line. Taxa not contributing to hybridization signal for the focal taxon are shown in grey for easier visualization. All D_{FOIL} results are summarized in Table S8.

muticum (**T**; Figures 2 and 5, event **1**) and to a lower extent with *Ae. comosa* (**M**) and *Ae. markgrafii* (**C**; Figures 3 and 4h). Moreover, it is connected to *Ae. speltoides* through its ancestor belonging to sect. *Sitopsis** (Figure 4h).

In addition to the major evolutionary scenario developed in this work, past or present gene flow among the different lineages of WWR cannot be ruled out entirely, whenever species come into contact with each other (Arrigo *et al.*, 2011; Bernhardt *et al.*, 2017). The existence of extinct ancestral lineages (Brassac and Blattner, 2015) that could not be sampled may, in general, mislead the results of *D* statistics (Beerli, 2004; Slatkin, 2005). However, in that case, *D* statistics are expected to return mostly false-negative test results (Pease and Hahn, 2015) instead of arriving

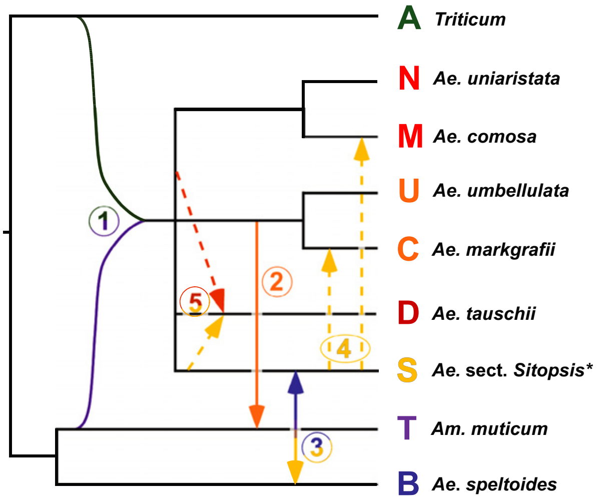

Figure 5. Total evidence evolutionary scenario for the wheat wild relatives. All diploid *Aegilops* species except *Ae. speltoides* are derived from an initial homoploid hybridization event involving the ancient A (*Triticum*) and T (*Am. muticum*) lineages (1). Strong signals of introgression were found for *Am. muticum* (from the U/C group; 2) and between *Ae. speltoides* and sect. *Sitopsis** (3). For the latter, introgression seems to have happened in both directions. Weaker signals of introgression (dashed arrows) were found by GBS-based *D* statistics from (4) sect. *Sitopsis** into *Ae. markgrafii* (C) and *Ae. comosa* (M), as well as (5) for *Ae. tauschii* (D).

at wrong species connections. Conversely, although we took a conservative approach, ancestral population structure, non-random mating, and small effective population sizes, characteristic of inbreeding species like most WWR species, could lead to high *D* statistic values (Eriksson and Manica, 2012; Martin *et al.*, 2015). New methods accounting for demographic processes at the scale of a genus are necessary to overcome this limitation.

CONCLUSIONS

We obtained DNA sequences of 244 nuclear low-copy genes evenly distributed among the Triticeae chromosomes and genome-wide SNP for all diploid species of the WWR. A combination of different phylogenetic and network approaches together with *D* statistics revealed ancient complex reticulated processes partly involving multiple rounds of introgression as well as at least one homoploid hybrid speciation during the formation of the extant taxa.

Based on our comprehensive taxon sampling, we are able to propose a detailed scheme of events that shaped the close relatives of wheat, which is much more complex than previously suggested (Marcussen *et al.*, 2014; Sandve *et al.*, 2015; Li *et al.*, 2015a,b; El Baidouri *et al.*, 2017; Huynh *et al.*, 2019). With two independent datasets, we were not only able to confirm the scenario developed by Glémin *et al.* (2019), that up to now seems not only to best reflect the evolution of WWR but also to uncover more complex patterns of interspecific gene flow. Our hypothesis is congruent with the proposed formation of the '**D**-genome lineage' through homoploid hybrid speciation (Marcussen *et al.*, 2014; Huynh *et al.*, 2019) but proposes, in agreement with Glémin *et al.* (2019), *Am. muticum* together with the *Triticum* lineage as progenitors. Furthermore, we suggest that *Ae. longissima* or members of sect. *Sitopsis** played an important role in the formation of *Ae. comosa* (**M**), *Ae. markgrafii* (**C**), and *Ae. tauschii* (**D**). We propose that an ancient, now extinct, lineage was introgressed by *Ae. longissima* or sect. *Sitopsis** and possibly also by an ancestor of the **CUMN** clade to form *Ae. tauschii*. Moreover, our data provide evidence of gene flow between sect. *Sitopsis** and the **B**-genome lineage, a hypothesis raised by El Baidouri *et al.* (2017) and Glémin *et al.* (2019). We also show that *Am. muticum* cannot be separated from *Aegilops*, as it is sister taxon to *Ae. speltoides* for nuclear data and is both a progenitor of, and introgressed by, other *Aegilops* species as shown from *D* statistics and plastid phylogenies (Yamane and Kawahara, 2005; Bordbar *et al.*, 2011; Bernhardt *et al.*, 2017). As the scenario proposed here is highly reticulate, it is necessary to obtain extensive genome information for all diploid species of this group to test predictions regarding composite genomes. Hybrid speciation and introgression should influence genome organization, the presence of syntenic blocks, and the occurrence of different transposable elements within the basic and hybrid lineages of the wheat-group taxa. In more general terms, the question remains if the important role of hybrid speciation and introgression that we found in the wheat group is a peculiarity of these taxa or if it plays an important role in most grasses, or generally in plant evolution but has not yet been detected, as studies using an approach similar to ours are still mainly in their infancy.

EXPERIMENTAL PROCEDURES

Plant materials

We analyzed 97 individuals representing all diploid species of the WWR with multiple individuals plus three outgroup taxa (i.e. *Dasypyrum*, *Hordeum*, *Taeniatherum*) of the grass tribe Triticeae (Table S5). All materials were grown from seed and identified based on morphological characteristics if an inflorescence was produced. Vouchers of the morphologically identified materials were deposited in the herbarium of IPK (GAT). Genome size and ploidy level of 83 individuals were initially verified by flow cytometry and genomic DNA was extracted as in Bernhardt *et al.* (2017).

Design of capture probes and library preparation for target enrichment

We used the assembly of *H. vulgare* cv 'Morex' (Mayer *et al.*, 2012), the only Triticeae draft genome that was available at the time of bait design, to select loci for which orthology could be confirmed when comparing these to the fully sequenced grass genomes of *Brachypodium distachyon*, rice, and sorghum (Vogel *et al.*, 2010; Matsumoto *et al.*, 2011; Mayer *et al.*, 2011).

Subsequently, one locus was selected every 0.5 cM on all *H. vulgare* chromosomes. These loci were used for BLAST comparisons (Altschul *et al.*, 1990) against available data of *Brachypodium*, rice, sorghum, barley, and wheat. Multiple sequence alignments were built including full-length cDNA (fl-cDNA) and genomic DNA sequences. Finally, 451 loci were chosen for the design of hybridization probes, if they showed: (i) a conserved exon-intron structure, (ii) a total length of exonic region larger than 1000 bp, with (iii) a minimum size of single exons being 120 bp, and (iv) introns separating adjacent short exons being smaller than 400 bp. The design of capture probes for the selected loci was finally based only on fl-cDNAs from *H. vulgare* and *T. aestivum*, two distantly related Triticeae taxa, and *Brachypodium distachyon,* which was used to broaden the taxonomic spectrum. Capture probes for each of the loci were designed on exon sequences of all three species. The loci used for bait design are evenly distributed over the **A-**, **B-**, and **D**-genomes of *T. aestivum* (Table S1 and Figure S1). The total exonic sequence information considered in bait design amounts to 690 kb. Custom Perl scripts were used to design bait sequences that were submitted to the web-based application eArray (Agilent Technologies). A detailed description of the bait design can be found in the Experimental procedures S1.

For each of the selected 69 samples (Table S5) 3 μg genomic DNA was sheared into fragments having an average length of 400 bp. The sheared DNA was used in a sequence-capture approach (SureSelect^{XT} Target Enrichment for Illumina Paired-End Sequencing, Agilent Technologies). All samples were barcoded, pooled, and sequenced on the Illumina HiSeq 2000 or MiSeq. For further details see Experimental procedures SI.

Library construction and sequencing for genotyping-by-sequencing

GBS and Illumina sequencing were performed for 57 individuals (Table S5) following Wendler *et al.* (2014). *Dasypyrum villosum* and *Taeniatherum caput-medusae* were included as outgroup taxa. For each individual, 200 ng genomic DNA were digested by two restriction enzymes *Pst*I-HF (CTGCAG, NEB Inc.) and *Msp*I (CCGG, NEB Inc.). Sequencing was carried out on an Illumina HiSeq 2500 obtaining 100 bp single-end reads.

Target-enrichment data assembly and analyses

Assembly. The loci were assembled in a two steps procedure. First, all 451 loci were assembled in a fast and non-stringent approach to evaluate if the capture worked sufficiently and if the loci are truly single-copy in most of the taxa. For each sample, the sequence reads were mapped to the barley genome assembly (Mayer *et al.*, 2012) using the Burrows–Wheeler Alignment (BWA) Tool v.0.7.8 (Li and Durbin, 2009). Consensus sequences were called using SAMtools v.1.1. (Li *et al.*, 2009; Li, 2011) and converted into FASTA sequences using vcfutils and Seqtk v.1.0 (Heng Li, https://github.com/lh3/seqtk). The percentage of ambiguous sites was determined for each sequence in locus-wise multiple sequence alignments. Allelic diversity is assumed to be much lower than 1% for single- and low-copy-number loci (for comparison see Jakob *et al.*, 2014). Therefore, a high percentage of ambiguous positions for sequences of the same species are assumed to reflect the presence of paralogous gene copies. Finally, loci with an average number of ambiguous sites >1% in six or more species of *Aegilops* and *Triticum* were considered as multicopy loci (Table S3). Then, the loci found to be mainly low-copy-number loci were kept and selected for a refined assembly procedure if they had a length of at least 1000 bp, contained less than 25% of missing data and had at least 15% of parsimony-informative positions, as identified with Paup v.*4.0a146 (Swofford, 2002). The refined assembly was performed in Geneious v.10.0.5 (Kearse *et al.*, 2012), as it can reliably assemble short insertions and deletions (Smith, 2015). For further details see Experimental procedures SI.

Phylogenetic analyses. To infer the phylogeny of the wheat relatives, we adopted an analysis approach consisting of the following steps. After aligning the sequences for all loci separately: (i) models of sequence evolution were determined for each locus; (ii) gene trees were inferred for each locus by M; (iii) the degree of gene tree/species tree conflict was investigated in detail using PhyParts; (iv) concatenated sequences from all loci (supermatrix) were used for BI, ML, MP, and NeighborNet analyses; (v) multispecies coalescent-based analyses were conducted to infer species trees from the ML gene trees; and (vi) phylogenetic networks were calculated based on the ML gene tree topologies. These analysis steps are detailed below.

Gene tree inference. Individual gene trees were inferred using RAxML v.8.1 (Stamatakis, 2014) under the GTRCAT model, rapid bootstrapping of 100 replicates and search for the best-scoring ML tree. To reduce noise from the data, the ML trees were further processed by contracting low support branches (bootstrap-values < 10) as suggested by (Chao *et al.*, 2018) with the Newick utilities function nw_ed and rerooted using the MRCA of *Hordeum* as the outgroup with the function nw_reroot (Junier and Zdobnov, 2010).

Supermatrix phylogeny. Multiple sequence alignments of all 244 loci were concatenated. Bayesian inference was performed in MrBayes v.3.2.6 (Ronquist *et al.*, 2012) on Cipres, Cyberinfrastructure for Phylogenetic Research Science Gateway 3.3 (Miller *et al.*, 2010). The best-fitting models of sequence evolution were estimated by making the MCMC sampling to be across all substitution models, as described in Bernhardt *et al.* (2017). *Hordeum vulgare* was set as the outgroup. An alternative approach to visualize the variation in the data was conducted by computing an unrooted phylogenetic network via SplitsTree v.4.14.8 (Huson and Bryant, 2006). The tool was run using the algorithms Uncorrected P, NeighborNet, and EqualeAngle for the matrix of the 244 concatenated target-enrichment loci.

An MP analysis of the supermatrix was conducted in Paup* v.4.0a146 (Swofford, 2002) to see if the phylogeny obtained by BI were sufficiently robust with regards to different analysis algorithms. The MP analysis was run using a heuristic search with 100 random-addition sequences and tree bisection and reconnection (TBR) branch swapping, saving all shortest trees. Node support was evaluated by 500 bootstrap re-samples with the same settings but without random-addition sequences.

Coalescent-based species tree estimation. The effect of gene tree conflicts due to ILS was addressed using the short-cut coalescence method Astral (Mirarab *et al.*, 2014; Chao *et al.*, 2018), which is able to estimate the true species tree with high probability, given a sufficiently large number of correct gene trees under the MSC model. Astral-III v.5.6.3 was run using 244 the ML edited and rerooted gene trees pre-estimated in RAxML.

Differences among gene trees. PhyParts (Smith *et al.*, 2015) was used to summarize the amount of concordant and conflicting phylogenetic signals from the 244 ML gene trees with the Astral topology as species tree. Visualization of the output was done as in Kates *et al.* (2018) and Villaverde *et al.* (2018), and using the phypartspiecharts.py script of M. Johnson available at www.github.com/mossmatters/phyloscripts.

Maximum pseudolikelihood gene tree-based phylogenetic networks estimation. Throughout all analyses *Ae. sharonensis* grouped within *Ae. longissima* and *T. monococcum* within *T. boeoticum.* This result confirmed previously known findings for these species, that is *Ae. sharonensis* and *Ae. longissima* are closely related taxa, and the unified or separate treatment of the two *Triticum* taxa is debated (van Slageren, 1994; Bernhardt, 2015). Here we used *Ae. sharonensis* and *T. boeoticum* if accessions were assigned to this taxon in the donor seed bank. However, due to their strong genetic similarity we treated *Ae. sharonensis* and *Ae. longissima* as well as *T. boeoticum* and *T. monococcum* as con-specific.

The effect of gene tree conflicts due to hybridizations was investigated with the maximum pseudolikelihood method InferNetwork_MPL (Yu and Nakhleh, 2015) included in the package PhyloNet (Than *et al.*, 2008; Wen *et al.*, 2018). The set of ML gene trees analyzed with astral was used as input for PhyloNet, allowing for zero to five hybridizations, other options were left to default. For each analysis, the best network was recorded and these were compared using the Akaike information criterion (AIC; Akaike, 1974). As suggested by Yu *et al.* (2012) and Morales-Briones *et al.* (2018), the number of parameters was set to the number of branches plus the number of hybridization probabilities being estimated. The network with the lowest AIC score was selected as the best-fit multispecies network. The network was visualized with Dendroscope (Huson and Scornavacca, 2012).

Assembly and analysis of GBS data

The assembly of the GBS data was performed *de novo* using ipyrad v.0.7.17 (Eaton, 2014; https://github.com/dereneaton/ipyrad), with strict filtering for adapters and restricting the maximum number of heterozygous sites per locus to 25%. Default settings were used for the remaining parameters.

A species tree based on SVDquartets (Chifman and Kubatko, 2014) under multispecies coalescence was estimated using Tetrad, as implemented in ipyrad v.0.7.17 with 100 bootstrap replicates. For comparison with the target-enrichment data, Splitstree v.4.14.8 (Huson and Bryant, 2006) was run using the methods Uncorrected P, NeighborNet and EqualeAngle to compute unrooted phylogenetic networks for 807 909 SNPs of the GBS analysis.

Identification of hybrid taxa

We used Four-taxon *D* statistics (Green *et al.*, 2010a; Durand *et al.*, 2011; Eaton and Ree, 2013) for the GBS data to identify candidate lineages involved in the introgressive hybridization within a fixed phylogeny (((P1, P2) P3), O). Under ILS alone, the number of shared SNPs resulting in an incongruent topology (i.e. ABBA and BABA) are expected to be equivalent. If P3 was involved in an introgressive event with P1, it will share more SNPs with P1 (i.e. BABA patterns) than with P2 (i.e. ABBA patterns).

The VCF file generated by ipyrad was first filtered with SAMtools/BCFtools (Li, 2011) retaining only unlinked SNPs. Four-Taxon *D* statistic tests were performed using the routine *Dtrios* of Dsuite (Malinsky, 2019; https://github.com/millanek/Dsuite). We first tested if *Taeniatherum caput-medusae* was involved in any introgressions. As no hybridization signal was found (Figure S10) and, because it is sharing more loci with the WWR than *D. villosum*, *Ta. caput-medusae* was used as outgroup taxon for all following tests. The VCF file was further processed to exclude all *D. villosum* individuals and *Dtrios* was used to perform 220 tests. Astral topology (Figure 1a) was used to specify species relationships. *D* statistics significance was assessed using jackknife (Green *et al.*, 2010a) on blocks of 100 SNPs. The function *p.adjust* in R v.3.5.3 (R Core Team, 2019) was used to apply a Benjamini–Yekutieli correction (Benjamini and Yekutieli, 2001). All 220 tests are summarized in Table S7. The results were visualized with the Ruby script 'plot_d.rb' available from M. Matschiner (https://github.com/mmatschiner).

The D_{FOIL} test (Pease and Hahn, 2015; https://github.com/jbpease/dfoil/) was used on the GBS data. It relies on a symmetric five-taxon phylogeny (((P1, P2), (P3, P4)), O) to identify the direction of introgressions among the candidate taxa identified using the Four-taxon *D* statistic. All tests were performed on species-specific consensus sequences. For each species, the alignment of all loci was used to call a consensus sequence that represented all diversity within the species. Therefore, we used the '0% identical' threshold in Geneious that minimizes the number of ambiguities. A custom workflow in Geneious was used to create datasets of five species including *Ta. caput-medusae* as outgroup. For all tests, we made sure that the estimated divergence times fitted the assumptions of the program, that is that P1 and P2 diverged after P3 and P4 in forward time, by excluding all tests that raised the warning 'b' (Table S8). We also used a feature of D_{FOIL}, that is D_{FOIL}alt, that excluded single derived-allele count for tests with an error warning 'c' (Table S8) following Leduc-Robert and Maddison (2018). As 216 tests were conducted, a Benjamini–Yekutieli correction (Benjamini and Yekutieli, 2001) was applied to all four statistics for each test with the function *p.adjust* in R 3.5.3 (R Core Team, 2019). A significance level of 0.01 was then used on the adjusted *P*-values to identify patterns of introgression.

ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) through grant BL462/10 to FRB and BK, and by basic funds from IPK Gatersleben. We thank M. Pfeiffer for help during the initial steps of bait design and K. Schneeberger for helpful discussions regarding locus choice and data assembly for the target-enrichment experiment. We are grateful to R. Brandt and A. Himmelbach for performing the Illumina sequencing, C. Koch, S. Koenig, B. Kraenzlin and B. Wedemeier for technical assistance, and S. Beier for bioinformatic support with the genetic map of *T. aestivum*. We thank ICARDA, IPK, USDA, the Czech Crop Research Institute, and the Kyoto University Laboratory of Plant Genetics for providing seed materials. We also thank the anonymous reviewers for helpful comments on earlier versions of the manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR CONTRIBUTIONS

Designed study: FRB, NB, BK. Coordinated study: NB. Provided data or materials: EMW, BK. Performed experiments: NB. Analyzed data: NB, JB, XD, FRB, and CHP. NB and FRB wrote the initial manuscript. All authors contributed to and approved the final version.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Representation of the physical locations of loci considered in bait design (left side) and low-copy loci used for analyses (right side) in the **A**-, **B**-, and **D**-genomes of *T. aestivum*.

Figure S2. Phylogenetic tree derived from Bayesian phylogenetic inference of the concatenated matrix of 244 loci.

Figure S3. Single maximum parsimony tree derived from the concatenated matrix of 244 loci.

Figure S4. Multispecies coalescent tree calculated with ASTRAL from separate ML trees of 244 loci.

Figure S5. Phylogenetic network inferred via the NEIGHBORNET method in SPLITSTREE4 from the concatenated matrix of 244 target-enriched loci.

Figure S6. Visualization of the degree of gene tree/species tree conflict.

Figure S7. Phylogenetic networks inferred under the multispecies network coalescent (MSNC) from the 244 gene tree topologies with maximum pseudolikelihood.

Figure S8. Consensus cladogram derived from a TETRAD analysis of the GBS data.

Figure S9. Phylogenetic network inferred via the NEIGHBORNET method in SPLITSTREE4 from 807 909 SNPs using the full GBS matrix.

Figure S10. Heatmap summarizing Four-taxon *D* statistic tests using *Dasypyrum villosum* as outgroup.

Table S1. Overview of loci selected for the sequence capture.

Table S2. Read statistics of the target-enrichment experiment.

Table S3. The proportion of ambiguous positions per accession, locus, and among species/genera.

Table S4. Read mapping per accession and information on the multiple sequence alignments and model of evolution for each of the 244 loci used for phylogenetic inference.

Table S5. Overview of the material considered in this study.

Table S6. Read and assembly statistics for the GBS data.

Table S7. Four-taxon *D* statistics.

Table S8. D_{FOIL} test for introgression in wheat wild relatives from GBS data.

Dataset S1. Assemblies of the 244 target enriched nuclear loci.

Dataset S2. Maximum likelihood gene trees of the 244 target enriched nuclear loci.

Dataset S3. Raw GBS data and concatenated GBS loci data matrix.

Experimental procedures S1. Extended descriptions of bait design and conducted analyses.

OPEN RESEARCH BADGE

This article has earned an Open Data Badge for making publicly available the digitally shareable data necessary to reproduce the reported results. The data are available at http://dx.doi.org/10.5447/IPK/2019/18

DATA AVAILABILITY STATEMENT

The assemblies of the 244 enriched nuclear loci (Dataset S1), the 244 gene trees (Dataset S2), the demultiplexed fasta-file of the barcoded reads for each accession used for GBS, and the matrix for the filtered loci (Dataset S3) are published via e!DAL-PGP (Arend *et al*., 2014, 2016) at http://dx.doi.org/10.5447/IPK/2019/18.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Arend, D., Lange, M., Chen, J., Colmsee, C., Flemming, S., Hecht, D. and Scholz, U. (2014) e!DAL - A framework to store, share and publish research data. *BMC Bioinformatics*, **15**, 214.

Arend, D., Junker, A., Scholz, U., Schüler, D., Wylie, J. and Lange, M. (2016) PGP repository: a plant phenomics and genomics data publication infrastructure. *Database*, **2016**, baw033.

Arrigo, N., Guadagnuolo, R., Lappe, S., Pasche, S., Parisod, C. and Felber, F. (2011) Gene flow between wheat and wild relatives: empirical evidence from *Aegilops geniculata*, *Ae. neglecta* and *Ae. triuncialis*. *Evol. Appl.* **4**, 685–695.

Beerli, P. (2004) Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations. *Mol. Ecol.* **13**, 827–836.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188.

Bernhardt, N. (2015) Taxonomic treatments of Triticeae and the wheat genus *Triticum*. In *Alien Introgression in Wheat* (Molnár-Láng, M., Ceoloni, C. and Doležel, J., eds). Cham, Switzerland: Springer, pp. 1–19.

Bernhardt, N., Brassac, J., Kilian, B. and Blattner, F.R. (2017) Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evol. Biol.* **17**, 141.

Bordbar, F., Rahiminejad, M.R., Saeidi, H. and Blattner, F.R. (2011) Phylogeny and genetic diversity of D-genome species of *Aegilops* and *Triticum* (Triticeae, Poaceae) from Iran based on microsatellites, ITS, and *trn*L-F. *Plant Syst. Evol.* **291**, 117–131.

Brassac, J. and Blattner, F.R. (2015) Species-level phylogeny and polyploid relationships in *Hordeum* (Poaceae) inferred by next-generation sequencing and *in silico* cloning of multiple nuclear loci. *Syst. Biol.* **64**, 792–808.

Chao, Zhang, Rabiee, M., Sayyari, E. and Mirarab, S. (2018) ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 153.

Chi, Zhang, Ogilvie, H.A., Drummond, A.J. and Stadler, T. (2018) Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* **35**, 504–517.

Chifman, J. and Kubatko, L. (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics*, **30**, 3317–3324.

Danilova, T.V., Akhunova, A.R., Akhunov, E.D., Friebe, B. and Gill, B.S. (2017) Major structural genomic alterations can be associated with hybrid speciation in *Aegilops markgrafii* (Triticeae). *Plant J.* **92**, 317–330.

Degnan, J.H. and Rosenberg, N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340.

Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011) Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252.

Eaton, D.A.R. (2014) PyRAD: Assembly of *de novo* RADseq loci for phylogenetic analyses. *Bioinformatics*, **30**, 1844–1849.

Eaton, D.A.R. and Ree, R.H. (2013) Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis*: Orobanchaceae). *Syst. Biol.* **62**, 689–706.

Eaton, D.A.R., Hipp, A.L., González-Rodríguez, A. and Cavender-Bares, J. (2015) Historical introgression among the American live oaks and the comparative nature of tests for introgression. *Evolution*, **69**, 2587–2601.

El Baidouri, M., Murat, F., Veyssiere, M., Molinier, M., Flores, R., Burlot, L., Alaux, M., Quesneville, H., Pont, C. and Salse, J. (2017) Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytol.* **213**, 1477–1486.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S. and Mitchell, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Eriksson, A. and Manica, A. (2012) Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl Acad. Sci. USA*, **109**, 13956–13960.

Escobar, J.S., Cenci, A., Bolognini, J., Haudry, A., Laurent, S., David, J. and Glémin, S. (2010) An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae): Selfing evolution in grasses. *Evolution*, **64**, 2855–2872.

Escobar, J.S., Scornavacca, C., Cenci, A., Guilhaumon, C., Santoni, S., Douzery, E.J., Ranwez, V., Glémin, S. and David, J. (2011) Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol. Biol.* **11**, 181.

Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J. and Ranwez, V. (2019) Pervasive hybridizations in the history of wheat relatives. *Sci. Adv.* **5**, eaav9188.

Green, R.E., Krause, J., Briggs, A.W. *et al.* (2010a) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Hejase, H.A. and Liu, K.J. (2016) A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, **17**, 422.

Huson, D.H. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267.

Huson, D.H. and Scornavacca, C. (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**, 1061–1067.

Huynh, S., Marcussen, T., Felber, F. and Parisod, C. (2019) Hybridization preceded radiation in diploid wheats. *Mol. Phylogenet. Evol.* **139**, 106554.

International Wheat Genome Sequencing Consortium (2014) A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, **345**, 1251788–1251788.

International Wheat Genome Sequencing Consortium (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, **361**, eaar7191.

Jakob, S.S., Rödder, D., Engler, J.O., Shaaf, S., Özkan, H., Blattner, F.R. and Kilian, B. (2014) Evolutionary history of wild barley (*Hordeum vulgare* subsp. *spontaneum*) analyzed using multilocus sequence data and paleodistribution modeling. *Genome Biol. Evol.* **6**, 685–702.

Jia, J., Zhao, S., Kong, X. *et al.* (2013) *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, **496**, 91–95.

Junier, T. and Zdobnov, E.M. (2010) The Newick utilities: high-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*, **26**, 1669–1670.

Kates, H.R., Johnson, M.G., Gardner, E.M., Zerega, N.J.C. and Wickett, N.J. (2018) Allele phasing has minimal impact on phylogenetic reconstruction from targeted nuclear gene sequences in a case study of *Artocarpus*. *Am. J. Bot.* **105**, 404–416.

Kearse, M., Moir, R., Wilson, A. *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.

Kellogg, E.A. (2015) *Flowering Plants. Monocots: Poaceae*. Berlin: Springer.

Kilian, B., Mammen, K., Millet, E., Sharma, R., Graner, A., Salamini, F., Hammer, K. and Özkan, H. (2011) Aegilops. In *Wild Crop Relatives: Genomic and Breeding Resources* (Kole, C., ed). Berlin, Heidelberg: Springer, pp. 1–76.

Kingman, J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.* **13**, 235–248.

Leduc-Robert, G. and Maddison, W.P. (2018) Phylogeny with introgression in *Habronattus* jumping spiders (Araneae: Salticidae). *BMC Evol. Biol.* **18**, 24.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**(16), 2078–2079.

Li, L.-F., Liu, B., Olsen, K.M. and Wendel, J.F. (2015a) A re-evaluation of the homoploid hybrid origin of *Aegilops tauschii*, the donor of the wheat D-subgenome. *New Phytol.* **208**, 4–8.

Li, L.-F., Liu, B., Olsen, K.M. and Wendel, J.F. (2015b) Multiple rounds of ancient and recent hybridizations have occurred within the *Aegilops-Triticum* complex. *New Phytol.* **208**, 11–12.

Ling, H.-Q., Zhao, S., Liu, D. *et al.* (2013) Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*, **496**, 87–90.

Luo, M.-C., Gu, Y.Q., You, F.M. *et al.* (2013) A 4-gigabase physical map unlocks the structure and evolution of the complex genome of *Aegilops tauschii*, the wheat D-genome progenitor. *Proc. Natl Acad. Sci. USA*, **110**, 7940–7945.

Luo, M.-C., Gu, Y.Q., Puiu, D. *et al.* (2017) Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*, **551**, 498–502.

Maddison, W.P. (1997) Gene trees in species trees. *Syst. Biol.* **46**, 523–536.

Malinsky, M. (2019) Dsuite - fast D-statistics and related admixture evidence from VCF files. *bioRxiv:634477*.

Mallet, J., Besansky, N. and Hahn, M.W. (2016) How reticulated are species? *BioEssays*, **38**, 140–149.

Marcussen, T., Sandve, S.R., Heier, L. *et al.* (2014) Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, **345**, 1250092.

Martin, S.H., Davey, J.W. and Jiggins, C.D. (2015) Evaluating the use of ABBA–BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257.

Mason-Gamer, R.J. and Kellogg, E.A. (1996) Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). *Syst. Biol.* **45**, 524–545.

Matsumoto, T., Tanaka, T., Sakai, H. *et al.* (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* **156**, 20–28.

Mayer, K.F.X., Martis, M., Hedley, P.E. *et al.* (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, **23**, 1249–1263.

Mayer, K.F.X., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K. and Close, T.J. (2012) A physical, genetic and functional sequence assembly of the barley genome. *Nature*, **491**, 711–716.

Miller, M.A., Pfeiffer, W. and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Gateway Comput Environ Workshop*, **2010**, 1–8.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S. and Warnow, T. (2014) ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.

Morales-Briones, D.F., Liston, A. and Tank, D.C. (2018) Phylogenomic analyses reveal a deep history of hybridization and polyploidy in the Neotropical genus *Lachemilla* (Rosaceae). *New Phytol.* **218**, 1668–1684.

Pease, J.B. and Hahn, M.W. (2015) Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64**, 651–662.

Petersen, G., Seberg, O., Yde, M. and Berthelsen, K. (2006) Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Mol. Phylogenet. Evol.* **39**, 70–82.

Poland, J.A., Brown, P.J., Sorrells, M.E. and Jannink, J.-L. (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS ONE*, **7**, e32253.

R Core Team (2019) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A. and Huelsenbeck, J.P. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542.

Sandve, S.R., Marcussen, T., Mayer, K., Jakobsen, K.S., Heier, L., Steuernagel, B., Wulff, B.B.H. and Olsen, O.A. (2015) Chloroplast phylogeny of *Triticum*/*Aegilops* species is not incongruent with an ancient homoploid hybrid origin of the ancestor of the bread wheat D-genome. *New Phytol.* **208**, 9–10.

Schreiber, M., Himmelbach, A., Börner, A. and Mascher, M. (2019) Genetic diversity and relationship between domesticated rye and its wild relatives as revealed through genotyping-by-sequencing. *Evol. Appl.* **12**, 66–77.

van Slageren, M.W. (1994) *Wild wheats: A monograph of Aegilops* L. and *Amblyopyrum* (Jaub. & Spach) Eig (Poaceae). Wageningen, the Netherlands: Agriculture University Papers.

Slatkin, M. (2005) Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Mol. Ecol.* **14**, 67–73.

Smith, D.R. (2015) Buying in to bioinformatics: An introduction to commercial sequence analysis software. *Brief. Bioinform.* **16**, 700–709.

Smith, S.A., Moore, M.J., Brown, J.W. and Yang, Y. (2015) Analysis of phylogenomic datasets reveals conflict, concordance, and gene duplications with examples from animals and plants. *BMC Evol. Biol.* **15**, 150.

Solís-Lemus, C. and Ané, C. (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* **12**, e1005896.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Swofford, D.L. (2002) *PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4.b10*. Sunderland, MA: Sinauer Associates.

Than, C., Ruths, D. and Nakhleh, L. (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, **9**, 322.

Villaverde, T., Pokorny, L., Olsson, S., Rincón-Barrado, M., Johnson, M.G., Gardner, E.M., Wickett, N.J., Molero, J., Riina, R. and Sanmartín, I. (2018) Bridging the micro- and macroevolutionary levels in phylogenomics: Hyb-Seq solves relationships from populations to species and above. *New Phytol.* **220**, 636–650.

Vogel, J.P., Garvin, D.F., Mockler, T.C. ***et al.*** (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

Weitemier, K., Straub, S.C.K., Cronn, R.C., Fishbein, M., Schmickl, R., McDonnell, A. and Liston, A. (2014) Hyb-Seq: combining target enrichment and genome skimming for plant phylogenomics. *Appl. Plant Sci.* **2**, 1400042.

Wen, D. and Nakhleh, L. (2018) Coestimating reticulate phylogenies and gene trees from multilocus sequence data. *Syst. Biol.* **67**, 439–457.

Wen, D., Yu, Y. and Nakhleh, L. (2016) Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet.* **12**, e1006006.

Wen, D., Yu, Y., Zhu, J. and Nakhleh, L. (2018) Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* **67**, 735–740.

Wendler, N., Mascher, M., Nöh, C., Himmelbach, A., Scholz, U., Ruge-Wehling, B. and Stein, N. (2014) Unlocking the secondary gene-pool of barley with next-generation sequencing. *Plant Biotechnol. J.* **12**, 1122–1131.

Xi, Z., Liu, L. and Davis, C.C. (2015) Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. *Mol. Phylogenet. Evol.* **92**, 63–71.

Yamane, K. and Kawahara, T. (2005) Intra- and interspecific phylogenetic relationships among diploid *Triticum-Aegilops* species (Poaceae) based on base-pair substitutions, indels, and microsatellites in chloroplast noncoding sequences. *Am. J. Bot.* **92**, 1887–1898.

Yu, Y. and Nakhleh, L. (2015) A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genom.* **16**, S10.

Yu, Y., Than, C., Degnan, J.H. and Nakhleh, L. (2011) Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* **60**, 138–149.

Yu, Y., Degnan, J.H. and Nakhleh, L. (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* **8**, e1002660.

Yu, Y., Dong, J., Liu, K.J. and Nakhleh, L. (2014) Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl Acad. Sci. USA*, **111**, 16448–16453.

Zhu, J. and Nakhleh, L. (2018) Inference of species phylogenies from bi-allelic markers using pseudolikelihood. *Bioinformatics*, **34**, i376–i385.