Pervasive hybridizations in the history of wheat relatives

Sylvain Glémin, Celine Scornavacca, Jacques Dainat, Concetta Burgarella, Veronique Viader, Morgane Ardisson, Gautier Sarah, Sylvain Santoni, Jacques David, Vincent Ranwez

AGRICULTURE

Pervasive hybridizations in the history of wheat relatives

Sylvain Glémin[1,2]*†, Celine Scornavacca[3†], Jacques Dainat[4,5], Concetta Burgarella[6,7], Véronique Viader[6], Morgane Ardisson[6], Gautier Sarah[6,8], Sylvain Santoni[6], Jacques David[6], Vincent Ranwez[6]

Cultivated wheats are derived from an intricate history of three genomes, A, B, and D, present in both diploid and polyploid species. It was recently proposed that the D genome originated from an ancient hybridization between the A and B lineages. However, this result has been questioned, and a robust phylogeny of wheat relatives is still lacking. Using transcriptome data from all diploid species and a new methodological approach, our comprehensive phylogenomic analysis revealed that more than half of the species descend from an ancient hybridization event but with a more complex scenario involving a different parent than previously thought—*Aegilops mutica*, an overlooked wild species—instead of the B genome. We also detected other extensive gene flow events that could explain long-standing controversies in the classification of wheat relatives.

INTRODUCTION

Reconstructing phylogenetic relationships between domesticated plant species and their wild relatives is of central interest for agriculture and breeding. Gene flow and hybridization between related species are relatively common in plants and make phylogeny reconstruction difficult because of numerous conflicts among individual gene genealogies (*1*). During rapid species divergence, incomplete lineage sorting (ILS), which occurs when ancestral polymorphisms are still shared by two species or more, is another source of phylogenetic conflicts (*2*). The *Aegilops*/*Triticum* genus, which includes cultivated wheat species, combines these challenging problems, and despite its high practical and economical importance, the phylogenetic relationships among species are still poorly resolved. These species form a group of annual Mediterranean and Middle East grasses comprising 13 diploid and 18 polyploidy species (including durum wheat and bread wheat). This genus belongs to the Triticeae tribe that is already known for its complex reticulated history (*3*, *4*), and the occurrence of many alloploid species (*5*) shows that hybridization is possible and has promoted species formation. Moreover, species diversification likely occurred rather rapidly [around 4 to 7 million years (Ma) (*6*, *7*)], and some species are highly polymorphic, with a large effective population size (*8*), generating a potentially high level of ILS. Both hybridization and ILS could explain why many conflicting results have been obtained for single-gene phylogenies so far (*9*, *10*). In particular, it has proven difficult to resolve the relationships among the diploid parental donors of the polyploid domesticated wheats, *Triticum urartu* (A genome), *Aegilops speltoides* (S genome, considered to be the closest current genome of the B genome), and *Aegilops tauschii* (D genome): A and B genomes constitute the tetraploid durum wheat, and A, B, and D genomes comprise the hexaploid bread wheat.

Recently, Marcussen *et al.* (*7*) proposed the hypothesis that the D genome lineage arose 5 to 6 Ma ago through a homoploid hybrid speciation between the A genome and B genome lineages (A, B, and D lineages hereafter), explaining the difficult resolution of a consensual tree-like history among these three groups. This result has been questioned, and more complex scenarios with several rounds of hybridization have been proposed since then (*5*, *11*, *12*). However, none of the previous large-scale studies included all diploid species. For example, Marcussen *et al.* (*7*) built their large multi-gene analysis only on the three diploid progenitors (plus one outgroup species) and the three corresponding genomes of the hexaploid wheat, whereas the 13 diploid species were analyzed using only six genes [see fig. S6 in (*7*)]. A genome-wide analysis including all diploid species is still lacking. We propose to re-evaluate the scenario of the homoploid speciation of the D genome and to position it in the complex phylogeny of the diploid relatives of cultivated wheats. To do so, we obtained and analyzed a comprehensive genomic dataset including all extant diploid species and developed a new framework to test intricate hybridization scenarios. Our results shed a new light on the history of wheat relatives, and we proposed a complex but robust scenario that resolves long-standing controversies on the history of these species.

RESULTS

A phylogenomic view of the history of wheat relatives

We produced a transcriptome-based dataset of orthologous coding sequences (CDSs) including at least two (and up to four) individuals for each of all the 13 diploid *Aegilops*/*Triticum* species plus one individual of three close outgroups belonging to the Triticeae tribe: *Taeniatherum caput-medusae*, *Secale vavilovii*, and *Eremopyrum bonaepartis* (table S1). In addition, we used the published sequence of the *Hordeum vulgare* genome (Genome Assembly ASM32608v1) as the most distant outgroup. We separately assembled the transcriptome of each individual and stringently clustered and aligned the annotated CDSs. After cleaning and processing (see Materials and Methods), we retained 11,033 alignments for the supertree analysis. Among them, we used the 8739 genes containing at most one sequence per individual for the supermatrix analysis and hybrid detection. The 11,033 individual gene

[1]CNRS, Univ Rennes, ECOBIO (Ecosystèmes, biodiversité, évolution)–UMR 6553, F-35042 Rennes, France. [2]Department of Ecology and Genetics, Evolutionary Biology Center, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden. [3]Institut des Sciences de l'Evolution Université de Montpellier, CNRS, IRD, EPHE CC 064, Place Eugène Bataillon, 34095 Montpellier, cedex 05, France. [4]National Bioinformatics Infrastructure Sweden (NBIS), SciLifeLab, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, S-751 23 Uppsala, Sweden. [5]IMBIM–Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala Biomedicinska Centrum (BMC), Husargatan 3, Box 582, S-751 23 Uppsala, Sweden. [6]AGAP, Univ Montpellier, CIRAD, INRA, Montpellier SupAgro, Montpellier, France. [7]CIRAD, UMR AGAP, F-34398 Montpellier, France. [8]South Green Bioinformatics Platform, BIOVERSITY, CIRAD, INRA, IRD, Montpellier SupAgro, Montpellier, France.
*Corresponding author. Email: sylvain.glemin@univ-rennes1.fr
†These authors contributed equally to this work.

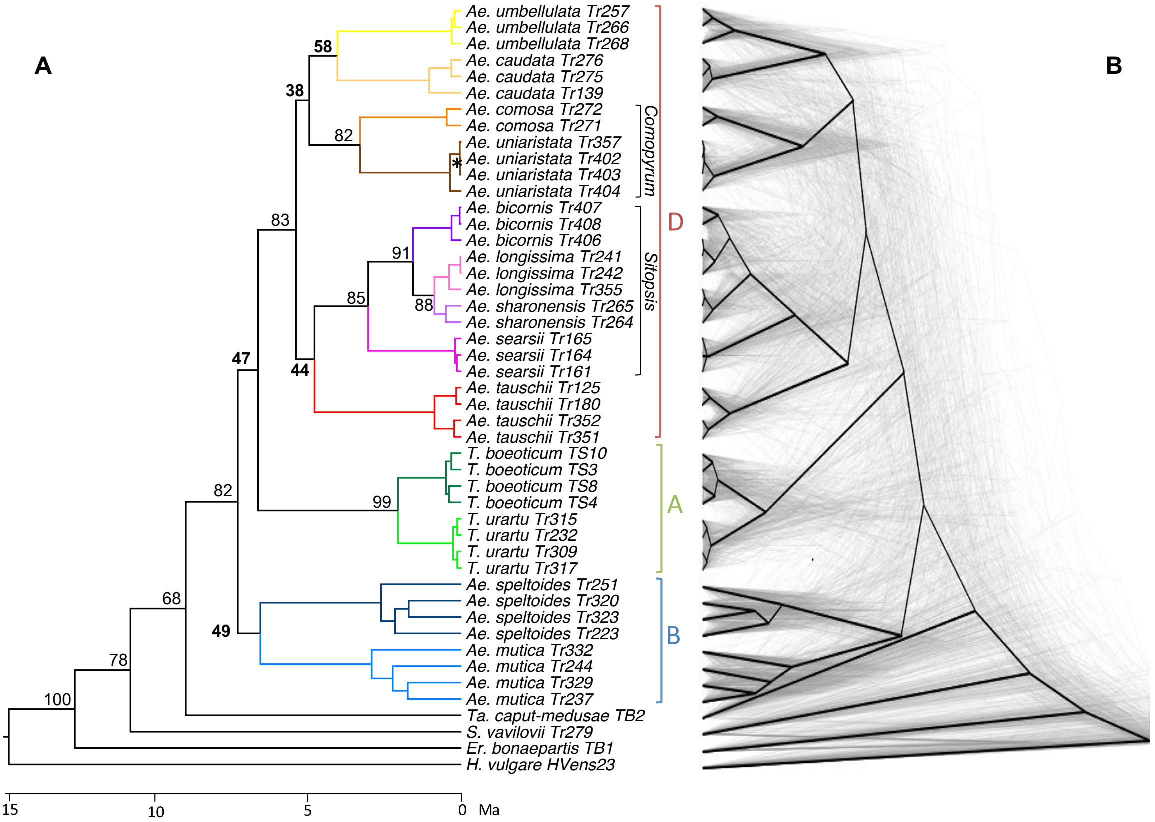


Fig. 1. Reconstructed phylogeny of the *Aegilops*/*Triticum* genus. (**A**) Phylogenetic tree of the *Aegilops*/*Triticum* genus. This same topology was obtained by both the ML analysis of 8739 gene alignments concatenation (supermatrix) and the supertree combination of 11,033 individual gene trees. All bootstrap values of the supermatrix analysis are 100 except those designated by an asterisk (* = 98). Support values for the supertree analysis are given for each interspecies node [percentage of triplets supporting a given node (*13*)]. Time scale was obtained by making the ML tree ultrametric and assuming a divergence of 15 Ma with Hordeum (*7*). (**B**) "Cloudogram" of 248 trees (in gray) inferred from non-overlapping 10-Mb genomic window concatenations. The global phylogeny is superposed in black.

trees used to construct the supertree with SuperTriplets (*13*) were obtained by maximum likelihood (ML) using RAxML v8 (*14*). The total-evidence species tree was also obtained by ML from the concatenation of all 8739 one-copy gene alignments. Both the supertree and the supermatrix approaches gave the same topology (Fig. 1A), distinguishing three main clades that we called the A lineage (the two *Triticum* species, *T. urartu* and *Triticum boeoticum*, the wild ancestor of the domesticated einkorn wheat *T. b. ssp. monococcum*), the B lineage (*Ae. speltoides* + *Aegilops mutica*), and the D lineage (all other species), following the simplified terminology of Marcussen *et al.* (*7*). This topology reveals new insights that partly contradict the traditional view of wheat relative evolution. First, while the *Sitopsis* clade is retrieved (including *Aegilops bicornis*, *Aegilops longissima*, *Aegilops searsii*, and *Aegilops sharonensis*), *Ae. speltoides* is excluded from this clade and appears to be the sister species of *Ae. mutica*. While this latter species has been excluded from the *Aegilops* genus for a long time and its phylogenetic position is debated (*10*), our results show that it is central in the history of wheats. Second, this topology clarifies what the D lineage corresponds to by showing that all nine other diploid *Aegilops* species belong to this clade (*Aegilops caudata*, *Aegilops comosa*, *Ae. tauschii*, *Aegilops umbelullata*, *Aegilops uniaristata*, and the four *Sitopsis* species). This contradicts the result of Marcussen *et al.* based on six genes only [see fig. S6 in (*7*)]; indeed, they claimed that the D lineage only included *Ae. tauschii* and the *Sitopsis* species, whereas the four other species were grouped within the B lineage. Third, it makes the relationships among species within the D lineage clearer, where no consensus had emerged so far. The species clustering is in agreement with their geographic proximity, roughly following an east-west distribution (fig. S1).

A new approach for analyzing multispecies coalescent with hybridization

However, while the two phylogenomic approaches were fully congruent and the supermatrix tree was strongly supported (bootstrap = 100 for all but one node), the supertree support values were low (<60) for 5 of 11 intragenus nodes (Fig. 1). This could be due to both ILS and hybridization. Scenarios with one or more hybridization events have already been proposed, but it was difficult to directly test them because they assumed ancestral events without considering all extant species. In addition, current methods to infer reticulated evolution with ILS are not yet able to deal with these large datasets (43 ingroup individuals here), especially with potential nested rounds of hybridization (text S1). For example, PhyloNetworks does not consider nested hybridizations (more formally, only level 1 networks are considered) (*15*). As an alternative strategy to disentangle hybridization events from ILS, we proceeded in three steps. First, we searched for all potential hybrids among triplets of species. Under pure ILS, one major topology and two equivalent minor topologies are expected, while two topologies can predominate over the third one under hybridization (*16*, *17*). This was the rationale used to propose the hybrid origin of the D genome (*7*). We thus counted the number of sites supporting the three possible topologies, from which we computed a hybridization index and its associated *P* value (*7*, *16*, *17*). Second, using the phylogeny we obtained as a reference (Fig. 1), we identified the possible hybridization scenarios compatible with triplets of species showing a significant departure from the null model with pure ILS. To do so, we analyzed the hybridization indices of all triplets of species in a systematic way (text S2): We hierarchically parsed the indices from the tips to the root of the phylogeny. We started from triplets including two individuals from the same species and a third individual from a sister species (recent hybridizations) to triplets of species belonging to the three main A, B, and D clades (ancient hybridizations). Third, we developed a new composite-likelihood method based on quartets of species to discriminate among complex scenarios: Only one hybridization can be detected with three taxa, whereas with four taxa, nine degrees of freedom are available, allowing us to infer scenarios with two hybridizations (Fig. 2 and texts S2 and S3). We applied the quartet method successively to the groups of species where we identified possible hybridizations.

Reconstruction of hybridization events

Hybridization appeared widespread, with 40% of triplets showing an index higher than 10%. However, the analysis of triplets composed of two individuals of a same species and a third individual from a second species revealed very low indices, suggesting nearly an absence of recent hybridization (text S2).

In previous studies, *Ae. mutica* was not considered as a member of the "B lineage," and the definition of the "D lineage" remained elusive (*5*, *7*, *18*). Thus, we searched to determine the parental species of the D lineage and whether all species of the D clade descended from the same hybridization event proposed by Marcussen *et al.* (*7*). To do so, we considered the indices for which an individual from the D clade could be a hybrid between parents from the A and B clades. The nine species of the D clade showed a clear signature of hybridization with a proportion of B species varying from 30 to 70% (Fig. 3A), suggesting that all D species are issued from hybridizations between the A and B clades. However, the distribution of indices was highly heterogeneous, both across potential hybrids and for potential parents, indicating a complex scenario for the formation of the D clade. The distribution of hybridization indices is similar regarding the A parents. In contrast, *Ae. speltoides* contributed much less than *Ae. mutica* to non-*Sitopsis* species of the D clade, while its contribution appears similar for *Sitopsis* species. In addition, *Ae. mutica* showed the unexpected and contradictory pattern of being both a potential parent of the D clade and a hybrid between *Ae. speltoides* and A or D species (fig. S2.2E). From a simple graphical reasoning and applying more formally our new quartet method, we showed that this could be explained by at least two interwoven hybridization events (text S2). In the most likely scenario, *Ae. mutica*, but not *Ae. speltoides*, hybridized with the ancestor of the A clade to give rise to the ancestor of the whole D clade, with a proportion of the A clade ranging from 0.35 to 0.58, suggesting a rather symmetrical hybridization (Fig. 4 and text S4). Before this event, the *Ae. mutica* ancestor was partly introgressed by the ancestor of the A clade, with proportions ranging from 0.11 to 0.18 (Fig. 4 and text S4). We also computed

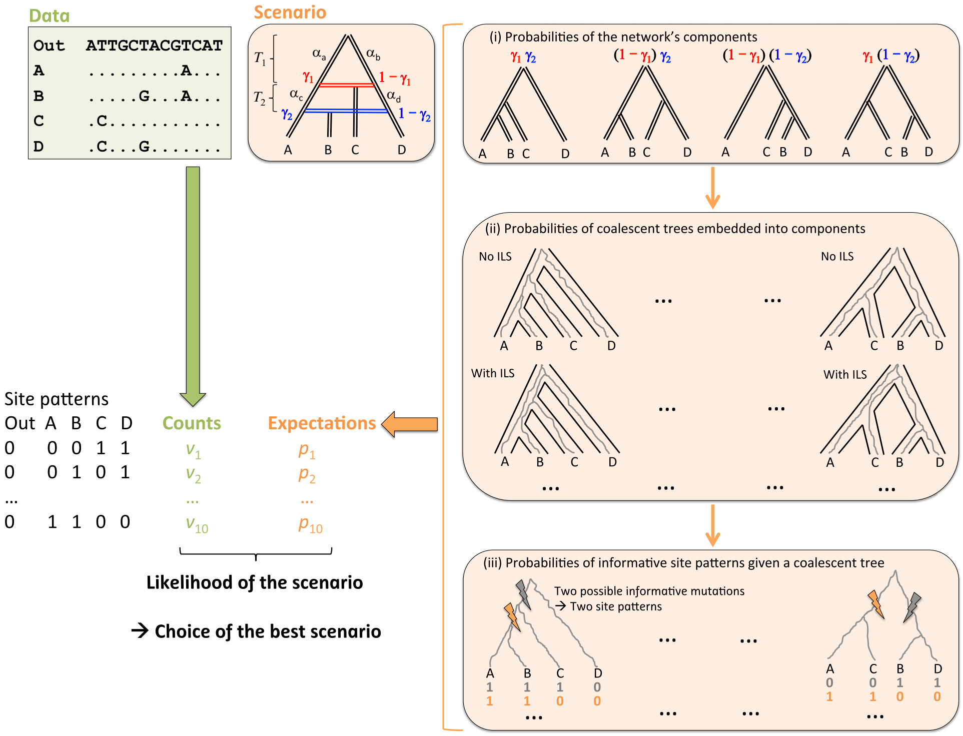


Fig. 2. Rationale of the quartet method. The dataset is composed of the counts of the 10 informative site patterns associated with four taxa (so nine degrees of freedom): 0 and 1 are the ancestral and derived states, respectively (polarization with an outgroup). A scenario corresponds to a network with four taxa and up to two hybridizations. It can be decomposed into components (i) with probabilities given by the hybridization proportions (γ_i). The model also includes the times of hybridization (T_i) and the coalescent rates on each branch (α_i) (eight parameters in total). For each component, (ii) the probabilities of embedded coalescent trees and (iii) the probabilities of site patterns given a coalescent tree are computed. They are function of T_i and α_i. Together, they give the expected frequencies of each site pattern for a given scenario. The likelihood of a scenario is given by the multinomial distribution of observing the count vector $\{v_1,\ldots, v_{10}\}$ given the expected frequencies $\{p_1,\ldots, p_{10}\}$. Likelihood comparison was used to choose the best scenario.

hybridization indices along chromosomes for 10-Mb windows. The distribution of indices did not indicate any large and contiguous introgression block (Fig. 3B and text S5) as would be expected under a single hybridization event followed by rapid speciation (*19*, *20*). Simulations showed that introgression blocks should be smaller than 10 Mb (the window size) to explain the observed patterns, even if chromosome rearrangements are included in the simulations (text S5). This is hardly compatible with a single and simple homoploid hybrid speciation event, so recurrent gene flow may have occurred during species divergence.

Among D species, the *Sitopsis* clade showed a distinctive hybridization signature compared to other species (Fig. 3A and text S2), likely due to a secondary introgression by *Ae. speltoides* (text S2). This scenario reconciles the morphological and cytological classifications of *Ae. speltoides* in the *Sitopsis* clade and some molecular-based phylogenies excluding *Ae. speltoides* from this clade (*10*, *21*). Last, we searched for other possible hybridization events within the D lineage. We found no signature of hybridization after the divergence of the *Sitopsis* and *Comopyrum* clades, in agreement with the strong supertree supports for these clades (Fig. 1) and with their ancient recognition as taxonomic

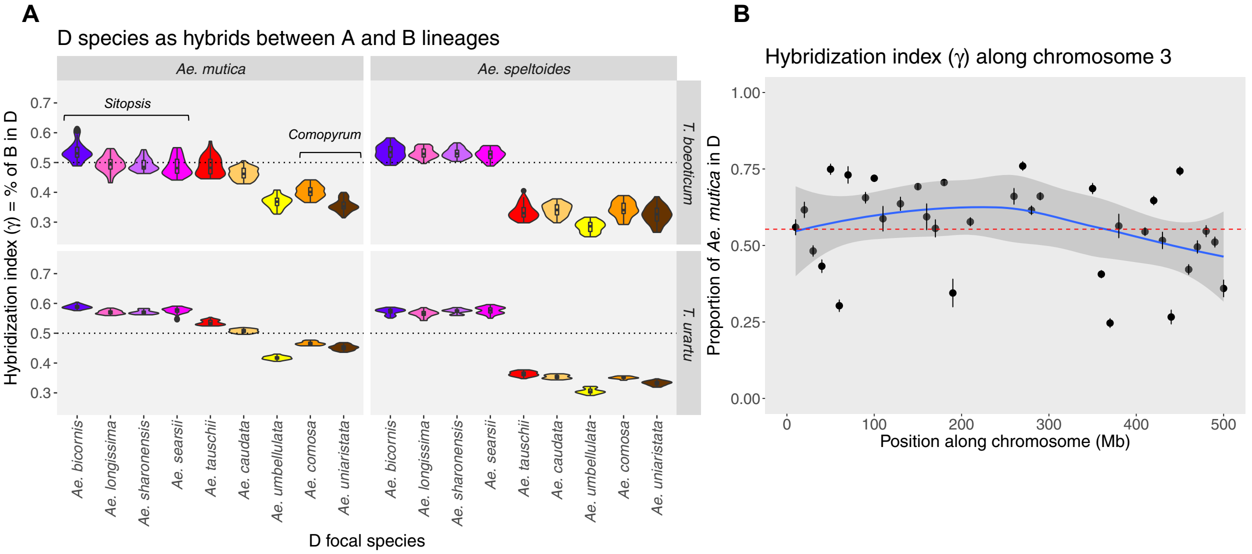
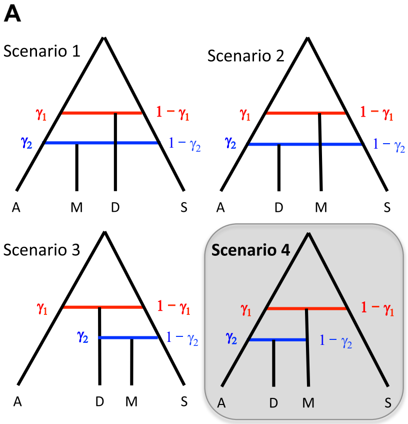


Fig. 3. Distribution of the hybridization index for the origin of the D clade. (**A**) Violin plots of the hybridization index for the nine species of the D lineage as a function of the A (*T. urartu* or *T. boeoticum*) and B (*Ae. speltoides* or *Ae. mutica*) parents. The dotted lines correspond to a perfect 50/50 hybridization. All indices are significantly different from 0 ($P < 10^{-6}$ after Bonferroni correction). (**B**) Distribution of the mean hybridization index [and 95% confidence interval (CI)] calculated on 10-Mb windows, along chromosome 3. Red dashed line, chromosome mean; blue line, loess regression with 95% CI in dark gray. The *Sitopsis* section and *Ae. speltoides* were excluded because of additional introgression (event 3 on Fig. 4).



B

Species A	Species D	Saturated	Scenario 1	Scenario 2	Scenario 3	Scenario 4	γ_1	γ_2
T. boeoticum	*Ae. caudata*	83.7	146.0	141.6	107.7	**93.5**	**0.76**[a]	**0.47**
	Ae. comosa	84.2	125.7	115.6	113.0	**103.1**	**0.76**[b]	**0.37**
	Ae. tauschii	84.0	150.9	145.5	105.7	**92.8**	**0.76**[c]	**0.48**
	Ae. umbellulata	84.3	115.8	109.3	***99.3***	**93.5**	**0.12**	**0.35**
	Ae. uniaristata	84.1	122.5	**104.0**	124.4	***109.2***	**0.13**	**0.42**
T. urartu	*Ae. caudata*	94.8	310.4	296.2	137.9	**116.8**	**0.11**	**0.50**
	Ae. comosa	95.0	221.0	196.2	154.3	**131.4**	**0.13**	**0.54**
	Ae. tauschii	95.2	399.9	347.3	212.6	**133.1**	**0.18**	**0.53**
	Ae. umbellulata	95.5	249.9	224.4	149.8	**120.2**	**0.13**	**0.55**
	Ae. uniaristata	95.3	245.7	211.0	174.9	**154.6**	**0.14**	**0.58**

[a] Second optimum AIC = 95.6, $\gamma_1 = 0.17$ and $\gamma_2 = 0.47$
[b] Second optimum AIC = 103.7, $\gamma_1 = 0.13$ and $\gamma_2 = 0.44$
[c] Second optimum AIC = 94.3, $\gamma_1 = 0.15$ and $\gamma_2 = 0.48$

Fig. 4. The best scenario for the origin of the D clade determined by the quartet method. (**A**) Schematic representation of the two-hybridization tested scenarios (A, species from the A clade; D, species from the D clade; M, *Ae. mutica*; S, *Ae. speltoides*). (**B**) Akaike Information Criterion (AIC) of the saturated model and the four tested scenarios. Models were run with the 10 different combinations of species from the A and D clades. The best AIC are in bold. In two cases, two models have close AIC (the second one is in italics). Scenario 4 is the best model in nine combinations and the second one (with close AIC) in one combination. Point estimates of γ_1 and γ_2 are given for scenario 4: D is the result of two successive hybridizations A + S → M then A + M → D. For the three first combinations, there is a second best model with a very close AIC with a much lower γ_1, in agreement with other values. Scenarios with no or only one reticulation were also tested, and all have much higher AIC (text S4).

entities. However, we found complex patterns for *Ae. umbellulata* and especially for *Ae. caudata*, suggesting pervasive gene flows before and during the divergence of this poorly supported clade (Fig. 1). Although we could not identify the complete scenario, at least three hybridization events are required to explain the results (Fig. 5 and text S3). This sheds light on the recent analysis of the genome of *Ae. caudata* (syn *Aegilops markgrafii*) that showed major structural rearrangements compatible with hybridization events (*22*).

DISCUSSION

Examples of nonbifurcating speciation histories are accumulating (*23*, *24*). Reconstructing species trees despite ILS and detecting introgression events are now feasible (*1*), but inferring the detailed history of multiple and successive events with more than a few species remains challenging. We proposed a new methodological framework to tackle this issue. First, we showed that a hierarchical analysis of hybridization indices helps identify the main potential events, hence simplifying further analyses. Where systematic methods that can deal with large datasets and complex scenarios are not available, such a "manual" step can be useful to reduce the analysis to a smaller number of taxa. Then, using quartets of species instead of triplets allows for a higher degree of freedom, hence fitting more complex scenarios [see also (*25*)]. Compared to triplet-based methods such as HyDe (*17*) or ABBA-BABA (*26*), this method does not require to assume a constant effective population size across species divergence to properly estimate hybridization or introgression proportions. In addition, our method goes beyond previous methods using quartet of species. For example, Pease and Hahn (*25*) only used the symmetry in site patterns to detect and polarize departure from pure ILS by defining a statistics with null expectation under the null hypothesis. Here, we obtain a full analytical expression for the expectation of site patterns, allowing writing likelihood functions, hence to test competitive models and to estimate their parameters. The detailed statistics properties of the model remained to be fully explored, and as for other methods based on site patterns (*17*, *25*, *26*), bias can occur because of misspecifications due to, for example, polarization errors or unbalanced missing data. These improvements still need to be developed. Furthermore, extending the method to five species or more would still allow more elaborated scenarios, but the exponentially growing number of parameters prevents any simple development for now. Alternatively, quartet statistics could be used as elementary blocks for iterative methods [e.g., (*15*, *27*, *28*)]. The findings presented here will be instrumental for these developments.

Owing to these new developments, we were able to propose a core reference scenario for the history of diploid *Aegilops*/*Triticum* species that should be pivotal for future research on wheats and their relatives (Fig. 5). We confirmed the occurrence of an ancient hybridization event that gave rise to the D lineage, but we showed (i) that this lineage includes 9, not only 5, of the 13 diploid species of the genus and (ii) that the hybridization scenario is a more complex scenario than previously proposed and involves a different parental species, *Ae. mutica* instead of *Ae. speltoides*. For a long time, *Ae. mutica* has been an overlooked species with a debated phylogenetic position. Our results plead for reconsideration and extensive study of this key species in the history of

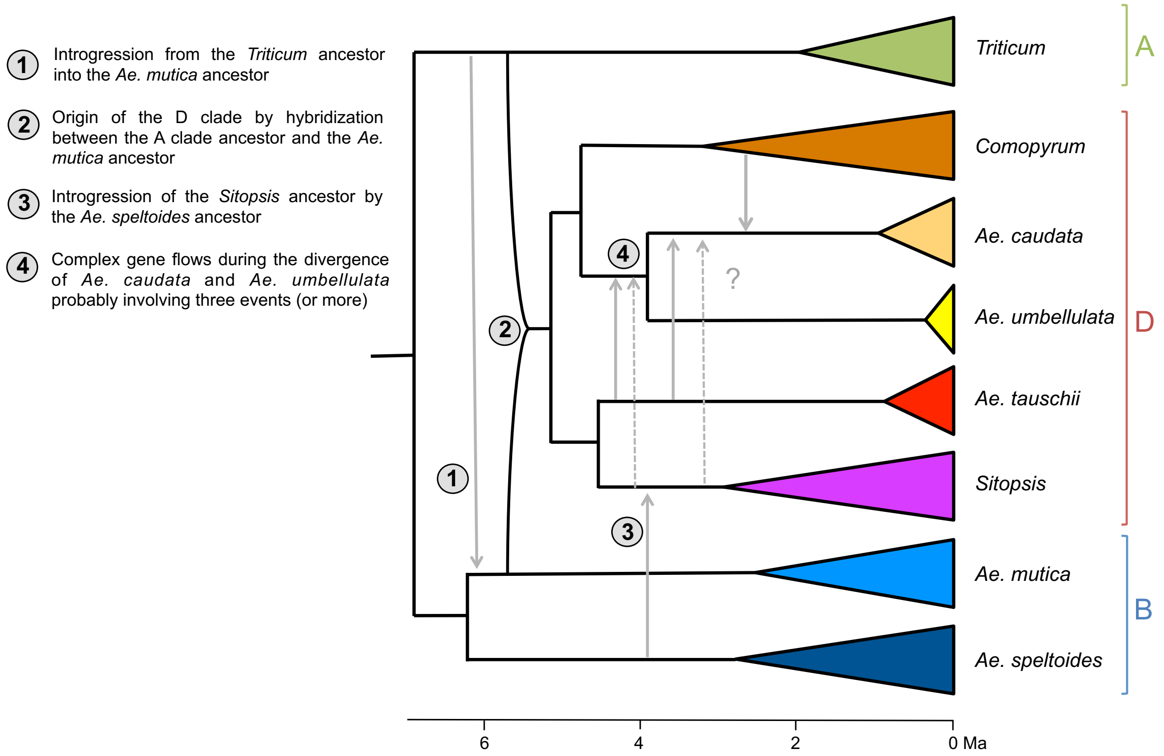

1 Introgression from the *Triticum* ancestor into the *Ae. mutica* ancestor
2 Origin of the D clade by hybridization between the A clade ancestor and the *Ae. mutica* ancestor
3 Introgression of the *Sitopsis* ancestor by the *Ae. speltoides* ancestor
4 Complex gene flows during the divergence of *Ae. caudata* and *Ae. umbellulata* probably involving three events (or more)
Triticum
A
Comopyrum
Ae. caudata
Ae. umbellulata
D
Ae. tauschii
Sitopsis
Ae. mutica
B
Ae. speltoides
?
1
2
3
4
6
4
2
0 Ma

Fig. 5. The proposed scenario for the history of diploid *Aegilops*/*Triticum* species. The proposed events obtained from the analysis of hybridization indices and from the quartet method have been added to the global phylogeny. Well-supported clades have been collapsed. The length of triangles corresponds to the divergence age as in Fig. 1. For event 4, the question mark indicates the uncertainties of this complex event. Solid arrows correspond to the most likely detected events, and dashed arrows correspond to possible additional ones: We could not exclude hybridization from *Sitopsis* as we could not formally test this hypothesis (text S3).

wheat relatives. *Ae. mutica* is a self-incompatible species with potentially large genetic diversity, and its direct implication in the history of the D genome makes it a strong candidate as a new reservoir of genetic diversity for wheat breeding programs. In addition, its potential interest is increased by its proximity with *Ae. speltoides*, the closest extant species of the progenitor of the B genome.

Our results also pointed to other introgression events to various extents, especially the introgression of *Ae. speltoides* in the ancestor of the *Sitopsis* clade, which can explain long-term controversies in the classification of wheat relatives. *Ae. speltoides* has been considered alternatively as a member of the *Sitopsis* section (*29*) or excluded from it (*9*, *10*). Our results explain these contradictory results. The scenario we proposed also suggests that chromosome similarities of repetitive elements between *Ae. speltoides* and the *Sitopsis* clade (*29*) may have resulted from transposable element exchanges following hybridization, a hypothesis that can now be tested within a clear phylogenetic framework. While our analysis pointed to other hybridization events, the signature is much less clear, likely because at least three events seem to be involved whereas the method we developed can only consider a maximum of two events.

Overall, these results suggest that this genus is especially prone to hybridization. A high hybridization potential could contribute to explain that more than half of *Aegilops* species are polyploids (especially allopolyploids). However, despite such a high ability to hybridize, we did not detect any recent hybridization events, in contrast to the many ancient events we found. The reason for this pattern still needs to be understood.

MATERIALS AND METHODS

Data acquisition

Data were obtained following the same procedure as in Sarah *et al.* (*30*) and redescribed here for comprehensiveness. Sequences of *T. boeoticum*, *T. caput-medusae*, and *E. bonaepartis* were already obtained by Clément *et al.* (*31*). Sequences of all other species were newly obtained.

All samples were constituted by a combination of leaves (20%) and inflorescence tissues (80%). RNAs were extracted and prepared separately for each organ and then mixed according to the given proportions. Samples were ground in liquid nitrogen, and total cellular RNA was extracted using a Spectrum Plant Total RNA kit (Sigma-Aldrich, USA) with a deoxyribonuclease treatment. RNA concentration was first measured using the NanoDrop ND-1000 Spectrophotometer and then with the Quant-iT RiboGreen (Invitrogen, USA) protocol on a Tecan Genios spectrofluorometer (Tecan Ltd., Switzerland). RNA quality was assessed by running 1 μl of each RNA sample on an RNA 6000 Pico chip on a Bioanalyzer 2100 (Agilent Technologies, USA). Samples with an RNA Integrity Number value greater than eight were deemed acceptable according to the Illumina TruSeq mRNA protocol.

The TruSeq Stranded mRNA Library Prep Kit (Illumina, USA) was used according to the manufacturer's protocol with the following modifications. Polyadenylate–containing mRNA molecules were purified from 2 μg of total RNA using poly-T oligo-attached magnetic beads. The purified mRNA was fragmented by the addition of the fragmentation buffer and was heated at 94°C in a thermocycler for 4 min. A fragmentation time of 4 min was used to yield library fragments of 250 to 300 base pairs (bp). First-strand complementary DNA (cDNA) was synthesized using random primers to eliminate the general bias toward the 3′ end of the transcript. Second-strand cDNA synthesis, end repair, A-tailing, and adapter ligation were performed in accordance with the protocols supplied by the manufacturer. Purified cDNA templates were enriched by 15 cycles of polymerase chain reaction (PCR) for 10 s at 98°C, 30 s at 65°C, and 30 s at 72°C using PE1.0 and PE2.0 primers and the Phusion High-Fidelity PCR Master Mix (Thermo Fisher Scientific, USA). Each indexed cDNA library was verified and quantified using a DNA 100 chip on a Bioanalyzer 2100 to build pooled libraries made of 12, equally represented, genotypes.

The final pooled library was quantified by quantitative PCR with the KAPA Library Quantification Kit (KAPA Biosystems, USA) and provided to the Get-PlaGe core facility (GenoToul platform, INRA Toulouse, France; www.genotoul.fr) for sequencing. Each final pooled library (12 genotypes) was sequenced using the Illumina paired-end protocol on a single lane of a HiSeq 3000 sequencer for 2 × 150 cycles.

Transcriptome assembly and annotation

Reads were cleaned and assembled following the pipeline described in Sarah *et al.* (*30*) and recalled here for comprehensiveness. Reads were preprocessed with cutadapt (*32*) using the TruSeq index sequence corresponding to the sample, searching within the whole sequence. The end of the reads with low-quality scores (parameter, −q 20) was trimmed, and we only kept trimmed reads with a minimum length of 35 bp and a mean quality higher than 30. Orphan reads were then discarded using a homemade script. Remaining paired reads were assembled using ABySS (*33*) followed by one step of Cap3 (*34*). Reads returned as singletons by the first assembly run were discarded. ABySS was launched using the paired-end option with a kmer value of 60. Cap3 was launched with the default parameters, including 40 bases of overlap, and the percentage of identity was set at 90%.

We slightly modified the RAPSearch program (*35*) to make its blast formatted output compatible with the expected input format of prot4est. We used this modified version of RAPSearch to identify protein sequences similar to our contigs either in plant species of UniProt SwissProt (www.uniprot.org) or in the Monocotyledon species of GreenPhyl (www.greenphyl.org/cgi-bin/index.cgi). We then used the prot4est program (*36*) to predict the CDS embedded in our contigs based on the following input: RAPSearch similarity output, Oryza matrix model for de novo–based predictions, and the codon usage bias observed in *T. boeoticum*.

Short sequences are often difficult to cluster into reliable orthologous groups and are not very informative for phylogeny inference; hence, we discarded predicted CDS with less than 250 bp as done in a similar context to populate the OrthoMaM database (*37*). The total numbers of contigs per species are given in table S2.

Orthologous search

We relied on USEARCH v7 (*38*) to cluster the predicted CDSs. We designed a four-step approach that limits the impact of taxon sampling and sequence ordering during cluster creation, avoids assigning sequences to an arbitrary cluster in case of tile, and can easily handle our large dataset (both in terms of required memory and computation time). First, for each species of the ingroup, we selected the accession with the highest number of CDSs to represent this species during the first step of cluster creation. Second, we used UCLUST to cluster these sequences and to output the median sequence of each cluster, which will be used as cluster bait. Third, we used USEARCH to identify, for each predicted CDS, the set of clusters for which the considered CDS and the cluster bait had a similarity above 85% along at least 50% of their length. Last, all predicted CDSs having such a similarity with one single cluster bait were assigned to this cluster; all others were discarded.

Alignment and cleaning

Following the strategy used to populate the OrthoMaM database (*37*), CDSs were aligned at the nucleotide level based on their amino acid translation, combining the speed of MUSCLE (*39*) and the ability of MACSE (*40*) to handle sequence errors in predicted CDSs resulting in apparent frameshift and erroneous amino acid translations. In more detail, for each cluster, we did the following: First, CDSs were translated into amino acids, these amino acid sequences were then aligned using MUSCLE, and the obtained protein alignment was used for deriving the nucleotide one using MACSE reportGapsAA2NT routine. Second, this nucleotide alignment was refined using MACSE refineAlignment routine. Last, the resulting amino acid alignment was cleaned with HMMcleaner (*41*), and a homemade script (that will be part of the next MACSE release) was used to report the obtained amino acid masking at the nucleotide level.

Phylogeny reconstructions

Gene trees were inferred with RAxML v8 (*14*) using the General Time Reversible (GTR) model with a four-category gamma distribution (GTR+Γ4) to accommodate for evolution rate heterogeneity among sites and using RAxML fast-bootstrap option (–f).

BppReroot of the BppSuite (*42*, *43*) was used to reroot the 13,288 gene trees, using as outgroups the following ordered list of species: *H_vulgare*, *Er_bonaepartis*, *S_vavilovii*, and *Ta_caputMedusae*. In more detail, for each of the gene tree, we considered each species of the outgroup list one after the other until finding the first one present in the current gene tree (if none was found, we discarded the tree). Having identified the most relevant outgroup species for this gene tree, we then checked whether all the individuals of this outgroup species formed a monophyletic clade; if yes, we rooted the tree on this clade, otherwise we discarded the tree. This resulted in a forest of 12,959 rooted gene trees, which we denoted by $\boldsymbol{F_i}$.

Since our aim here was to build a phylogeny of species and not of individuals, we focused on the identification of reliable species clades from the information contained in the gene trees. Therefore, we derived from $\boldsymbol{F_i}$ two forests of multilabel trees by renaming each sequence by the species to which it belongs to (forest $\boldsymbol{F_m}$) and keeping only clades with a bootstrap value greater than 95 (forest $\boldsymbol{F^{95}{}_m}$). Almost all trees (99.99%) in these forests are multilabeled as alignments include several individuals for at least some species. We thus used SSIMUL (*44*) to process the multilabel of trees of $\boldsymbol{F_m}$ and $\boldsymbol{F^{95}{}_m}$ by turning—without losing phylogenetic signal when possible—its multilabeled trees into single-labeled trees. This was done by removing a copy of each pair of isomorphic sibling subtrees (*44*). We denoted by $\boldsymbol{F_s}$ and $\boldsymbol{F^{95}{}_s}$ the new forests obtained by pruning isomorphic trees of $\boldsymbol{F_m}$ and $\boldsymbol{F^{95}{}_m}$, respectively. We used SuperTriplets (*13*) to construct a supertree from the 11,033 trees in $\boldsymbol{F^{95}{}_s}$. The resulting supertree is depicted in Fig. 2. The support values given by SuperTriplets to the clades are very low (only three clades have a support greater than 90); this shows that, even if we only keep clades with a support greater than 95, $\boldsymbol{F^{95}{}_s}$ contains a high level of contradiction.

Supermatrix analysis

In the forest $\boldsymbol{F_i}$, some trees are also multilabeled at the individual level either because of paralogy or because the two allelic copies were split. From $\boldsymbol{F_i}$, we extracted the set of 8739 trees containing at most one sequence per individual. We built the concatenation of all the 8739 alignments corresponding to these trees, giving a supermatrix with one sequence for all individuals. We inferred the phylogeny from this supermatrix with RAxML v8 (*14*) using the GTR+Γ4 model and the fast-bootstrap option. The resulting phylogeny has the same topology than the supertree shown in Fig. 1, and all nodes but one have bootstrap values equal to 100. Using the Hordeum genome as a reference, we also concatenated genes in 10-Mb windows along chromosomes, obtaining 298 alignments with at least three genes per window. (For this analysis, 5976 genes were kept since the others either could not be assigned to a position on the Hordeum genome or were isolated—one or two sequences—in their 10-Mb window.) We reconstructed the phylogeny of each alignment using the same method. The global tree and the 298 10-Mb trees were made ultrametric using the chronos function of the ape R package (*45*). Among the 298 10-Mb trees, only 248 contained all individuals. We used them to draw the cloudogram presented in Fig. 2B using Densitree (*46*).

Detection of hybridization events

We used the same supermatrix alignments to detect possible hybridization events by applying the rationale developed by Meng and Kubatko (*16*) and Kubatko and Chiffman (*47*). Note that this was also the rationale used to propose the hybrid origin of the D genome (*7*). In broad strokes, if we consider a triplet of lineages A, B, and C, with B being a hybrid between A (in proportion 1 − γ) and C (in proportion γ), then the probabilities of the three rooted topologies are given by

$$\begin{aligned} P[\mathrm{A},(\mathrm{B},\mathrm{C})] &= \gamma(1-2\exp(-t)/3)+(1-\gamma)\ \exp(-t)/3 \\ P[\mathrm{C},(\mathrm{A},\mathrm{B})] &= (1-\gamma)(1-2\exp(-t)/3)+\gamma\exp(-t)/3 \\ P[\mathrm{B},(\mathrm{A},\mathrm{C})] &= \ \exp(-t)/3 \end{aligned}$$

where t represents the time between speciation events on the parental trees measured in $2N_e$ generations. It can be easily shown that

$$\frac{P[\mathrm{A},(\mathrm{B},\mathrm{C})]-P[\mathrm{B},(\mathrm{A},\mathrm{C})]}{P[\mathrm{A},(\mathrm{B},\mathrm{C})]+P[\mathrm{C},(\mathrm{A},\mathrm{B})]-2P[\mathrm{B},(\mathrm{A},\mathrm{C})]}=\gamma$$

In addition, note that $2P[\mathrm{B},(\mathrm{A},\mathrm{C})] = 2\exp(-t)/3$ directly gives the probability of incongruence due to ILS (*2*). Thus, γ can be estimated by counting the number of two-state (*i* and *j*) positions supporting each topology using an outgroup (O) to polarize mutations (*47*). Considering the order O/A/B/C, we have

$$\begin{aligned} x &= \#i,i,j,j \rightarrow \mathrm{A},(\mathrm{B},\mathrm{C}) \\ y &= \#i,j,j,i \rightarrow \mathrm{C},(\mathrm{A},\mathrm{B}) \\ z &= \#i,j,i,j \rightarrow \mathrm{B},(\mathrm{A},\mathrm{C}) \end{aligned}$$

So we can define a hybridization index that is an estimator of γ as

$$\hat{\gamma}=\frac{x-z}{x+y-2z}$$

To test the significance of this estimator, that is, to identify γ values not due to random sampling under pure ILS, Kubatko and Chiffman (*17*, *47*) proposed a statistics (called the "Hils statistics") that is normally distributed with mean zero and variance one. It allows rapidly detecting significant potential hybrids among all possible triplets in a large phylogeny. We used this test to filter out the γ estimates and only consider significant ones. Because of the high rate of false positive of this test (*17*, *47*) and of the large number of sites in the alignment, we used the very stringent threshold of 10^{-6} (instead of 0.05) after Bonferroni correction for multiple testing. In addition, we focused on major events for which $\gamma >$ 10%. Notably, the above rationale implicitly assumes that the effective size, N_e, remained the same in the two diverging A and C lineages. Relaxing this assumption biases the estimation of γ, but $\hat{\gamma}$ is still expected to be null only without hybridization, so that the detection of hybridization is conservative. However, a single $\hat{\gamma}$ value can be difficult to interpret when multiple hybridization events occurred. Thus, we first computed the statistics for all triplets to list all possible hybridization events. Then, we formally tested the proposed scenarios within an ML framework (see below).

To compute the values of $\hat{\gamma}$ for each triplet of individuals, we applied a modified version of the HyDe program (*17*, *47*) to allow retrieving not only the Hils statistics but also the counts of each patterns (x, y, and z). As outgroup, we used the consensus sequence of the four outgroup species to limit homoplasy, which can bias statistics. For each triplet, we ordered topologies and species such that $x > y > z$ and computed $\hat{\gamma}$. We applied it to the full alignment and to the 298 10-Mb window alignments.

With 43 ingroup individuals, 74,046 triplets are possible, making the analysis of individual triplets useless. Instead, we parsed the results hierarchically based on the clades previously obtained with phylogenetic analyses: We started from triplets of species belonging to the same species and sister species to triplets of species belonging to the three main clades (A, B, and D). From this analysis (detailed in text S2), we proposed a series of hybridization scenarios. To detect possible heterogeneity of ILS and hybridization events across the genome, we also analyzed the variation of the two statistics along chromosomes and performed simulations to evaluate the size of hybridization blocks across the genome (see text S5).

Test of multiple hybridization scenarios

With three taxa, only three rooted topologies are possible, leaving only two degrees of freedom to estimate scenario parameters, which is not sufficient if multiple hybridization events occurred. Using four taxa, 10 informative biallelic site patterns are possible, leaving nine degrees of freedom to infer scenarios (text S3). Noting 0 the ancestral and 1 the derived allele, the 10 informative site patterns are 0|0111, 0|1011, 0|1101, 0|1110, 0|0011, 0|0101, 0|0110, 0|1001, 0|1010, and 0|1100. Scenarios with four taxa and up to two hybridization events can be described with eight parameters (see below and fig. S2.3). In text S3, we show how to write the probabilities of the 10 site patterns under a four-taxon multispecies coalescent model with up to two hybridization events. To do so, we need to compute both the probabilities of the compatible gene tree topologies and the expected length of branches for which the occurrence of a mutation leads to the given pattern. Then, to obtain the probabilities for a full scenario, we need to take the weighted sum of all possible gene trees embedded in the four-taxon hybridization network. Formally, the probability of site pattern i within a scenario $\mathbb{S}$ can be written as

$$p_i(\mathbb{S}) = \frac{\sum_{\mathbb{C}\in\mathbb{S}} P(\mathbb{C}|\mathbb{S}) \sum_{\mathbb{T}\in\mathbb{C}} P(\mathbb{T}|\mathbb{C}) d_i|\mathbb{T}}{\sum_{\mathbb{C}\in\mathbb{S}} P(\mathbb{C}|\mathbb{S}) \sum_{\mathbb{T}\in\mathbb{C}} P(\mathbb{T}|\mathbb{C}) \sum_{i=1}^{10} d_i|\mathbb{T}}$$

where $\mathbb{C}$ is a component of the decomposition of the scenario $\mathbb{S}$ (for the species tree or one-reticulation network, see below), $\mathbb{T}$ is a gene tree embedded in component $\mathbb{C}$, and $d_i|\mathbb{T}$ is the expected length of the branch where a mutation leads to site pattern i for a given gene tree, $\mathbb{T}$.

Scenarios with two non-nested reticulations can be decomposed into the four trees displayed by the corresponding phylogenetic network (*48*, *49*). We first obtained the vectors of expected branch lengths leading to the 10 site patterns for these four trees—denoted by $\boldsymbol{l_i}$, with i ranging from 1 to 4. Note that the longer a branch, the higher the probability for a mutation to occur, so that branch lengths directly affect observed pattern frequencies. We hence enumerated all possible gene trees embedded in a given four-taxon species tree and computed both the probabilities of the compatible topologies and the mean length of the branches where the occurrence of a mutation leads to a given site pattern. Probabilities and branch lengths are function of divergent times and coalescent rates (text S3). Then, a full scenario with hybridization can be obtained by combining the corresponding trees with their respective weights. Consider two non-nested hybridization events with proportions of the parental lineages being γ_1 and $1 - \gamma_1$ for the first event and γ_2 and $1 - \gamma_2$ for the second one. The vector of probabilities for the full network is thus

$$\boldsymbol{p} = \frac{1}{K}(\gamma_1\gamma_2\boldsymbol{l}_1 + \gamma_1(1-\gamma_2)\boldsymbol{l}_2 + (1-\gamma_1)\gamma_2\boldsymbol{l}_3 + (1-\gamma_1)(1-\gamma_2)\boldsymbol{l}_4)$$

where K is a normalization constant such that $\sum_{i=1}^{10} p_i = 1$.

For scenarios with two nested reticulations, hybridization and coalescent processes cannot be fully decoupled (*49*), and some embedded coalescent trees must be computed directly on a network component instead on a tree component. If only one species is issued from two nested hybridization events (the only case considered here), then the initial network can be decomposed into two trees in proportions $\gamma_1\gamma_2$, $(1-\gamma_1)\gamma_2$ and one one-reticulation network in proportion $(1-\gamma_2)$. Noting $\boldsymbol{l_1}$ and $\boldsymbol{l_2}$ the vectors of branch lengths for the two trees and $\boldsymbol{\lambda}$ for the one-reticulation network, the vector of probabilities for the full network is thus

$$\boldsymbol{p} = \frac{1}{K}(\gamma_1\gamma_2\boldsymbol{l}_1 + (1-\gamma_1)\gamma_2\boldsymbol{l}_2 + (1-\gamma_2)\lambda)$$

where K is the normalization constant.

Noting $\boldsymbol{v}$ the vector of the number of positions corresponding to the 10 biallelic patterns, the likelihood of a network is given by the multinomial sampling

$$L = \left(\sum_{i=1}^{10} v_i\right)! \prod_{i=1}^{10} \frac{p_i^{v_i}}{v_i!}$$

By fixing either γ_1 or γ_2 to 0 or 1, we obtained a scenario with only one reticulation, and by fixing both parameters to 0 or 1, we achieved a tree-like scenario without any reticulation. A scenario with one reticulation has six free parameters and that without any reticulation has only four. As all scenarios cannot be nested in each other, we used Akaike Information Criterion (AIC) to compare them, where AIC = $2k - 2\ln(L)$. Below, we show how to compute the $\boldsymbol{p}$ vectors. Likelihood maximization was made with a Mathematica script provided in the Supplementary Materials. The FindMaximum function was used with 100 random starting points.

In the following, we excluded the *Sitopsis* clade from the analyses because of the additional hybridization with *Ae. speltoides*. We first applied the model to the four taxa: A clade, D clade, *Ae. mutica*, and *Ae. speltoides* to elucidate the origin of the D clade. Because the triplet analysis showed heterogeneity among species, we successively run the model for the 10 combinations of the two species from the A clade (*T. boeoticum* and *T. urartu*) and the five species of the D clade

(*Ae. caudata*, *Ae. comosa*, *Ae. tauschii*, *Ae. umbellulata*, and *Ae. uniaristata*). As only four sequences are required for this analysis, we used the strict consensus of the different sequences of the same species. As for the triplet analysis, we used the consensus sequence of the four outgroup to polarized mutations. We only tested scenarios where the D clade and *Ae. mutica* could be potential hybrids as there was no signature that neither *Ae. speltoides* nor the two *Triticum* species could be potential hybrids according to the distribution of hybridization indices. We then applied the method to *Ae. caudata*, *Ae. tauschii*, *Ae. umbellulata*, and either *Ae. comosa* or *Ae. uniaristata* from the *Comopyrum* clade.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/5/5/eaav9188/DC1
Text S1. Jointly inferring phylogenetic relationships and hybridization events
Text S2. Determination of possible hybridization scenarios
Text S3. Inferring multiple hybridizations in four-taxon scenarios (quartet method)
Text S4. Application of the quartet method
Text S5. Pattern of hybridization indices along chromosomes
Fig. S1. Geographic distribution of the 13 diploid *Aegilops*/*Triticum* species.
Fig. S2.1. Distribution of hybridization indices for triplets composed of two individuals of the same focal species and a third individual from another species.
Fig. S2.2. Distribution of hybridization indices for triplets composed of two individuals from two sister species and a third one.
Fig. S2.3. Distribution of hybridization indices for triplets composed of two species belonging to sister clades within the D lineage and a third species.
Fig. S2.4. *Sitopsis* species as potential hybrids between species of the D clade and either *Ae. mutica* or *Ae. speltoides*.
Fig. S2.5. An example where successive hybridization events lead to apparent contradiction in hybrid triplets.
Fig. S3.1. The 10 informative site patterns with four species and mutations polarized with an outgroup.
Fig. S3.2. Notations for symmetric and asymmetric coalescent gene trees with four sequences.
Fig. S3.3. Decomposition of a hybridization network with two reticulations.
Fig. S3.4. Example of scenario parameterization.
Fig. S3.5. Possible gene trees embedded in a four-taxon species tree.
Fig. S3.6. Possible gene trees embedded in a four-taxon network tree as drawn on fig. S3.3C.
Fig. S4.1. Tested scenarios for the origin of the D clade.
Fig. S4.2. Tested scenarios for hybridization within the D clade.
Fig. S5.1. Distribution of the mean hybridization index for (A, *Ae. mutica*, and D) triplets (proportion of *Ae. mutica* in D) across 10-Mb concatenations.
Fig. S5.2. Distribution of the mean hybridization index and CI for (A, *Ae. mutica*, and D) triplets (proportion of *Ae. mutica* in D) along chromosomes.
Fig. S5.3. Simulation of the distribution of the hybridization index with mean size of genomic blocs from 5 to 50 Mb.
Fig. S5.4. Simulation of the distribution of the hybridization index with mean size of genomic blocs of 10 Mb and with 10 to 200 rearrangements per chromosome (500 Mb).
Table S1. List of accessions used in the study.
Table S2. Number of contigs (size, >250 bp) per individual trancriptome.
Table S3.1. Expected numbers of site patterns for symmetric taxon trees.
Table S3.2. Expected numbers of site patterns for asymmetric taxon trees.
Table S3.3. Expected numbers of site patterns for one-reticulation networks.
Table S4.1. Count numbers for the 10 informative site patterns with the different combinations of species from the A and D clade.
Table S4.2. AIC of the different scenarios with zero, one, or two reticulations.
Table S4.3. ML estimates of the parameters for the best model (M 4) for the 10 combinations of A and D species.
Table S4.4. Count numbers for the 10 informative site patterns with either *Ae. comosa* or *Ae. uniaristata* as a potential parent.
Table S4.5. AIC of the different scenarios with zero, one, or two reticulations.
Data file S1. This Mathematica notebook file contains two main parts: (i) the detailed derivation of the different equations presented in text S3 and (ii) the implementation of the likelihood method presented in text S3 (quartet method) and its application to the dataset.

REFERENCES AND NOTES

1. L. Nakhleh, Computational approaches to species phylogeny inference and gene tree reconciliation. *Trends Ecol. Evol.* **28**, 719–728 (2013).
2. P. Pamilo, M. Nei, Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583 (1988).
3. N. Bernhardt, J. Brassac, B. Kilian, F. R. Blattner, Dated tribe-wide whole chloroplast genome phylogeny indicates recurrent hybridizations within Triticeae. *BMC Evol. Biol.* **17**, 141 (2017).
4. J. S. Escobar, C. Scornavacca, A. Cenci, C. Guilhaumon, S. Santoni, E. J. P. Douzery, V. Ranwez, S. Glémin, J. David, Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol. Biol.* **11**, 181 (2011).
5. L.-F. Li, B. Liu, K. M. Olsen, J. F. Wendel, A re-evaluation of the homoploid hybrid origin of *Aegilops tauschii*, the donor of the wheat D-subgenome. *New Phytol.* **208**, 4–8 (2015).
6. S. Huang, A. Sirikhachornkit, X. Su, J. Faris, B. Gill, R. Haselkorn, P. Gornicki, Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum*/*Aegilops* complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8133–8138 (2002).
7. T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer; The International Wheat Genome Sequencing Consortium, K. S. Jakobsen, B. B. H. Wulff, B. Steuernagel, K. F. X. Mayer, O.-A. Olsen, Ancient hybridizations among the ancestral genomes of bread wheat. *Science* **345**, 1250092 (2014).
8. J. Dvořák, M.-C. Luo, Z.-L. Yang, Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing aegilops species. *Genetics* **148**, 423–434 (1998).
9. K. Yamane, T. Kawahara, 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 (2005).
10. G. Petersen, O. Seberg, M. Yde, K. Berthelsen, 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 (2006).
11. L.-F. Li, B. Liu, K. M. Olsen, J. F. Wendel, Multiple rounds of ancient and recent hybridizations have occurred within the *Aegilops-Triticum* complex. *New Phytol.* **208**, 11–12 (2015).
12. M. El Baidouri, F. Murat, M. Veyssiere, M. Molinier, R. Flores, L. Burlot, M. Alaux, H. Quesneville, C. Pont, J. Salse, Reconciling the evolutionary origin of bread wheat (*Triticum aestivum*). *New Phytol.* **213**, 1477–1486 (2017).
13. V. Ranwez, A. Criscuolo, E. J. P. Douzery, SuperTriplets: A triplet-based supertree approach to phylogenomics. *Bioinformatics* **26**, i115–i123 (2010).
14. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
15. C. Solís-Lemus, P. Bastide, C. Ané, PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.* **34**, 3292–3298 (2017).
16. C. Meng, L. S. Kubatko, Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: A model. *Theor. Popul. Biol.* **75**, 35–45 (2009).
17. P. D. Blischak, J. Chifman, A. D. Wolfe, L. S. Kubatko, HyDe: A Python package for genome-scale hybridization detection. *Syst. Biol.* **67**, 821–829 (2018).
18. S. R. Sandve, T. Marcussen, K. Mayer, K. S. Jakobsen, L. Heier, B. Steuernagel, B. B. H. Wulff, O. A. Olsen, 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 (2015).
19. M. C. Ungerer, S. J. E. Baird, J. Pan, L. H. Rieseberg, Rapid hybrid speciation in wild sunflowers. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 11757–11762 (1998).
20. C. A. Buerkle, L. H. Rieseberg, The rate of genome stabilization in homoploid hybrid species. *Evolution* **62**, 266–275 (2008).
21. P. Gornicki, H. Zhu, J. Wang, G. S. Challa, Z. Zhang, B. S. Gill, W. Li, The chloroplast view of the evolution of polyploid wheat. *New Phytol.* **204**, 704–714 (2014).
22. T. V. Danilova, A. R. Akhunova, E. D. Akhunov, B. Friebe, B. S. Gill, Major structural genomic alterations can be associated with hybrid speciation in *Aegilops markgrafii* (Triticeae). *Plant J.* **92**, 317–330 (2017).
23. P. Y. Novikova, N. Hohmann, V. Nizhynska, T. Tsuchimatsu, J. Ali, G. Muir, A. Guggisberg, T. Paape, K. Schmid, O. M. Fedorenko, S. Holm, T. Säll, C. Schlötterer, K. Marhold, A. Widmer, J. Sese, K. K. Shimizu, D. Weigel, U. Krämer, M. A. Koch, M. Nordborg, Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).
24. J. B. Pease, D. C. Haak, M. W. Hahn, L. C. Moyle, Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLOS Biol.* **14**, e1002379 (2016).
25. J. B. Pease, M. W. Hahn, Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64**, 651–662 (2015).
26. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
27. V. Ranwez, O. Gascuel, Quartet-based phylogenetic inference: Improvements and limits. *Mol. Biol. Evol.* **18**, 1103–1116 (2001).
28. E. Sayyari, S. Mirarab, Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).

29. E. D. Badaeva, B. Friebe, B. S. Gill, Genome differentiation in *Aegilops*. 1. Distribution of highly repetitive DNA sequences on chromosomes of diploid species. *Genome* **39**, 293–306 (1996).
30. G. Sarah, F. Homa, S. Pointet, S. Contreras, F. Sabot, B. Nabholz, S. Santoni, L. Sauné, M. Ardisson, N. Chantret, C. Sauvage, J. Tregear, C. Jourda, D. Pot, Y. Vigouroux, H. Chair, N. Scarcelli, C. Billot, N. Yahiaoui, R. Bacilieri, B. Khadari, M. Boccara, A. Barnaud, J.-P. Péros, J.-P. Labouisse, J.-L. Pham, J. David, S. Glémin, M. Ruiz, A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Mol. Ecol. Resour.* **17**, 565–580 (2017).
31. Y. Clément, G. Sarah, Y. Holtz, F. Homa, S. Pointet, S. Contreras, B. Nabholz, F. Sabot, L. Sauné, M. Ardisson, R. Bacilieri, G. Besnard, A. Berger, C. Cardi, F. De Bellis, O. Fouet, C. Jourda, B. Khadari, C. Lanaud, T. Leroy, D. Pot, C. Sauvage, N. Scarcelli, J. Tregear, Y. Vigouroux, N. Yahiaoui, M. Ruiz, S. Santoni, J.-P. Labouisse, J.-L. Pham, J. David, S. Glémin, Evolutionary forces affecting synonymous variations in plant genomes. *PLOS Genet.* **13**, e1006799 (2017).
32. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).
33. J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, İ. Birol, ABySS: A parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
34. X. Huang, A. Madan, CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
35. Y. Ye, J.-H. Choi, H. Tang, RAPSearch: A fast protein similarity search tool for short reads. *BMC Bioinformatics* **12**, 159 (2011).
36. J. D. Wasmuth, M. L. Blaxter, prot4EST: Translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* **5**, 187 (2004).
37. E. J. P. Douzery, C. Scornavacca, J. Romiguier, K. Belkhir, N. Galtier, F. Delsuc, V. Ranwez, OrthoMaM v8: A database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol. Biol. Evol.* **31**, 1923–1928 (2014).
38. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
39. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
40. V. Ranwez, S. Harispe, F. Delsuc, E. J. P. Douzery, MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PLOS ONE* **6**, e22594 (2011).
41. H. Philippe, D. M. de Vienne, V. Ranwez, B. Roure, D. Baurain, F. Delsuc, Pitfalls in supermatrix phylogenomics. *Eur. J. Taxon.* **283**, 1–25 (2017).
42. L. Gueguen, S. Gaillard, B. Boussau, M. Gouy, M. Groussin, N. C. Rochette, T. Bigot, D. Fournier, F. Pouyet, V. Cahais, A. Bernard, C. Scornavacca, B. Nabholz, A. Haudry, L. Dachary, N. Galtier, K. Belkhir, J. Y. Dutheil, Bio++: Efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.* **30**, 1745–1750 (2013).
43. J. Dutheil, S. Gaillard, E. Bazin, S. Glémin, V. Ranwez, N. Galtier, K. Belkhir, Bio++: A set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics* **7**, 188 (2006).
44. C. Scornavacca, V. Berry, V. Ranwez, Building species trees from larger parts of phylogenomic databases. *Inf. Comput.* **209**, 590–605 (2011).
45. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
46. R. R. Bouckaert, J. Heled, DensiTree 2: Seeing trees through the forest. *bioRxiv* 10.1101/012401 (2014).
47. L. S. Kubatko, J. Chifman, An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *bioRxiv* 10.1101/034348 (2015).
48. D. H. Huson, R. Rupp, C. Scornavacca, *Phylogenetic Networks: Concepts, Algorithms and Applications* (Cambridge Univ. Press, 2010).
49. S. Zhu, J. H. Degnan, Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst. Biol.* **66**, 283–298 (2017).

Acknowledgments: Analyses were performed on the computing cluster of the Montpellier Bioinformatics Biodiversity (MBB) platform, the South Green Bioinformatics platform, and the UPPMAX platform in Uppsala. **Funding:** This work was supported by the French National Research Agency (ANR-11-BSV7-013-03). **Author contributions:** Conception: J. Dav., S.G., and V.R.; funding acquisition: J. Dav. and S.G.; biological data acquisition and management: C.B. and V.V.; sequence data acquisition: G.S., M.A., and S.S.; method development: S.G.; data analysis: C.S., J. Dai., S.G., and V.R.; interpretation of the results: C.B., C.S., J. Dav., S.G., and V.R.; drafting of the manuscript: S.G.; reviewing and editing of the manuscript: C.B., C.S., J. Dav., S.G., S.S., and V.R. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Sequence alignments and phylogenetic trees generated and used to produce the results are available at https://bioweb.supagro.inra.fr/WheatRelativeHistory/. Additional data related to this paper may be requested from the authors.

Submitted 31 October 2018
Accepted 20 March 2019
Published 1 May 2019
10.1126/sciadv.aav9188

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

Science Advances

Pervasive hybridizations in the history of wheat relatives

Sylvain Glémin, Celine Scornavacca, Jacques Dainat, Concetta Burgarella, Véronique Viader, Morgane Ardisson, Gautier Sarah, Sylvain Santoni, Jacques David and Vincent Ranwez

Sci Adv **5** (5), eaav9188.
DOI: 10.1126/sciadv.aav9188