Network Analysis Provides Insights Into Evolution of 5S rDNA Arrays in Triticum and Aegilops

Robin G. Allaby and Terence A. Brown

Department of Biomolecular Sciences, University of Manchester Institute of Science and Technology, Manchester M60 1QD, United Kingdom

Manuscript received September 2, 1999
Accepted for publication November 27, 2000

ABSTRACT

We have used network analysis to study gene sequences of the Triticum and Aegilops 5S rDNA arrays, as well as the spacers of the *5S-DNA-A1* and *5S-DNA-2* loci. Network analysis describes relationships between 5S rDNA sequences in a more realistic fashion than conventional tree building because it makes fewer assumptions about the direction of evolution, the extent of sexual isolation, and the pattern of ancestry and descent. The networks show that the 5S rDNA sequences of Triticum and Aegilops species are related in a reticulate manner around principal nodal sequences. The spacer networks have multiple principal nodes of considerable antiquity but the gene network has just one principal node, corresponding to the correct gene sequence. The networks enable orthologous groups of spacer sequences to be identified. When orthologs are compared it is seen that the patterns of intra- and interspecific diversity are similar for both genes and spacers. We propose that 5S rDNA arrays combine sequence conservation with a large store of mutant variations, the number of correct gene copies within an array being the result of neutral processes that act on gene and spacer regions together.

THE 5S rDNA of Triticeae genomes is arranged in tandem arrays of several thousand repeating units of gene and spacer. Variation occurs within a single array, for both genes and spacers (Gerlach and Dyer 1980), and there is a high divergence of spacer sequences among different members of the tribe, with only infrequent homogenization between orthologous and paralogous loci (Scoles *et al.* 1988; Dvorák *et al.* 1989). These observations have prompted attempts to use the phylogenetic information contained in 5S rDNA arrays to resolve close relationships within the Triticeae. Initial studies suggested that sequences of unknown locus origin could be attributed to genomes with only a few cases of ambiguity (Appels *et al.* 1992; Sastri *et al.* 1992), but more detailed analyses revealed a high degree of character conflict, resulting in poor resolution when conventional tree building methodologies are used (Kellogg and Appels 1995). Character conflict probably arises from homogenization processes, such as unequal crossing over and biased gene conversion (Dover 1982, 1986; Dvorák *et al.* 1989; Hillis *et al.* 1991), these processes occurring within arrays and providing most 5S rDNA repeat units with multiple ancestry. As an alternative to conventional tree building, we have applied network analysis to the 5S rDNA sequences of Triticeae, with the results reported in this article.

5S rDNA organization: The 5S rRNA molecule, which is 120 bp long and highly conserved across species (Erdmann and Wolters 1986), complexes with the large rRNA and several proteins and helps maintain translational fidelity (Dinman and Wickner 1995). In most eukaryotes, the 5S rDNA genes are arranged in tandem arrays separate from the arrays specifying the other rRNA molecules. Several 5S rDNA arrays may be present in a single genome (*e.g.*, Goldsborough *et al.* 1981; Dvorák *et al.* 1989; Reddy and Appels 1989), usually with a few hundred to a few thousand repeat units per array, although much longer arrays are known [*e.g.*, ~25,000 units in *Xenopus laevis* (Ford and Southern 1973)]. Copy numbers vary 10–20-fold within a single species [*e.g.*, Lagudah *et al.* (1989) for *Aegilops squarrosa*], and many genomes, including those of the Triticeae, probably also possess multiple minor arrays (Reddy and Appels 1989; Dubcovsky and Dvorák 1995). These length variations suggest that amplification and/or deletion processes such as unequal crossing over act on individual arrays. In plants, the length of the spacer varies between 100 and 700 bp (Sastri *et al.* 1992) and is often characteristic of a particular locus [*e.g.*, Cox *et al.* (1992) for *Triticum aestivum*]. The spacers are thought to play some role in transcription initiation and termination (Scoles *et al.* 1988), but the location of the 5S rDNA promoters within the gene sequences, coupled with the occurrence of some extremely short spacers in some species [*e.g.*, 30 bp for Brachypodium (Cox *et al.* 1992)], suggests that most of the spacer sequence is devoid of function.

There are two lineages of 5S rDNA loci in the Triticeae, called *5S-DNA-1* and *5S-DNA-2* (Gerlach and Dyer 1980). The two loci are paralogous (Scoles *et al.*

Corresponding author: Terry Brown, Department of Biomolecular Sciences, UMIST, Manchester M60 1QD, United Kingdom.
E-mail: terry.brown@umist.ac.uk

1988) and are on homeologous chromosomes 1 and 5, respectively, in Triticum and Aegilops species (Dvořák *et al.* 1989) and on chromosomes 2 and 3 in Hordeum (Kolchinsky *et al.* 1990; Kanazin *et al.* 1993). The two lineages can be distinguished by the lengths and sequences of their spacers (Gerlach and Dyer 1980; Dvořák *et al.* 1989; Appels *et al.* 1992; Baum and Appels 1992). In Triticum the *5S-DNA-1* and *5S-DNA-2* spacers are 200–349 and 350–380 bp, respectively, the size difference being due to an insertion/deletion in the mid-spacer region (Appels *et al.* 1992). The *5S-DNA-1* loci can be further subdivided because the spacers on chromosome 1A are shorter (240 bp) than those occurring on chromosomes 1B and 1D (290–349 bp; Allaby and Brown 2000). Sequence alignment between spacers of different size classes is difficult (*e.g.*, Kellogg and Appels 1995) due to the high frequency of indels. For this reason, we analyze the *5S-DNA-A1* and *5S-DNA-2* spacers separately in this article.

Gene and spacer variation within and between species: It has been known for some time that the spacer sequences of Triticum species vary by 2–13% in a single 5S array, with an average heterogeneity of 5% (Gerlach and Dyer 1980; Appels *et al.* 1992). Kellogg and Appels (1995) have shown that a similar amount of variability is displayed by the genes. In *T. monococcum*, for example, the nucleotide diversity (π) values (Nei 1987) are 0.028 ± 0.021 for the gene and 0.031 ± 0.023 for the spacer. These comparisons suggest that the majority (>70%) of the gene copies within an array do not code for functional 5S rRNA molecules and that the selection pressure on any single gene copy is weak. Curiously, when arrays are compared between species, the nucleotide diversity of the gene sequences is about equal to or slightly less than that found within a single species, and no differences are fixed (Tables 1 and 2 in Kellogg and Appels 1995). These authors have highlighted the apparent paradox that the high variation within an array suggests that little selection is acting on individual genes, but the conservation of genes between species implies that variation is periodically removed at a rate approximately equal to that of speciation. Spacers, on the other hand, display higher nucleotide diversity between rather than within species, with many interspecific differences apparently fixed, although there are examples of homoplasy and some closely related taxa show higher intra- compared with interspecific divergence (Tables 1 and 4 in Kellogg and Appels 1995). This pattern of sequence variation could be explained by a selection pressure acting on the array as a whole, which must retain a critical number of functional genes to remain viable (Kellogg and Appels 1995), similar to events thought to occur in Drosophila (Schlötterer and Tautz 1994). According to this model, selection pressure on any one gene is weak, the homogenization mechanism for the array as a whole is also weak, and there are periods of rapid array removal that may correspond to the speciation rate.

Phylogenetic analysis of 5S rDNA repeat units: Conventional tree building methods are based on comparison of homologous characters in different taxa that are assumed to be reproductively isolated and whose relationship can be described by an essentially dichotomous branching pattern. The direction of evolution, although not always known if the tree is unrooted, is assumed to be linear. These assumptions are carried over to molecular phylogenetics, when gene trees are used to infer species trees, but are violated when comparisons are made of 5S rDNA sequences in a single array or in closely related species. This is partly due to the high incidence of character conflict when 5S rDNA trees are constructed, but is also a result of the recent evolutionary timescale, which means that ancestral states still exist and multiple apomorphisms are being fixed or lost. The overall result is that 5S rDNA sequences are related by a multifurcating rather than dichotomous branching pattern. Similar problems in the analysis of molecular variance in mitochondrial DNA (mtDNA) were addressed by Excoffier *et al.* (1992) through the use of minimum spanning networks, as previously described for nongenetical applications by Prim (1957). Bandelt *et al.* (1995, 1999) have also used a system of network construction for phylogenetic analysis of mtDNA sequences that display frequent homoplasy, all the most parsimonious trees being described in a network into which ancestral states, if still extant, are easily incorporated, and this method has been applied with success to comparisons of human populations (*e.g.*, Richards *et al.* 1995; Sykes *et al.* 1995; Forster *et al.* 1996; Torroni *et al.* 1998). However, these types of network analysis are not appropriate for 5S rDNA because they assume that the true relationship between the sequences is represented by one of the linear relationships within the network. This is not the case for 5S rDNA sequences, whose high incidence of character conflict probably results from recombination, so that each sequence has a multiple ancestry. This means that the true relationship between groups of sequences is reticulate rather than linear. The most appropriate method for phylogenetic analysis of 5S rDNA is therefore the network system used to describe recombining sequences at the *Adh* locus of *Drosophila melanogaster* (Simmons *et al.* 1989; Berry and Kreitman 1993) because this approach does not assume reproductive isolation and enables multiple ancestry, apomorphy, and extant ancestral states to be portrayed. In this article, we apply this type of network analysis to three levels of 5S rDNA organization in Triticum and Aegilops species: first, we examine specific spacer types within the *5S-DNA-A1* locus; second, we compare homeologous *5S-DNA-2* loci; and third, we analyze gene sequences from various *5S-DNA-1* and *5S-DNA-2* loci.

MATERIALS AND METHODS

Plant material: Seeds of *T. urartu* (catalog no. IPSR 1010011) and *T. sinskajae* (IPSR 1050001) were obtained from the Institute of Plant Science Collection of Wheats and Related Species, John Innes Centre, Norwich, United Kingdom; *T. dicoccoides* (Gat 601098) and *T. dicoccum* (Gat 17029) were obtained from the Institut für Pflanzengenetik und Kulturpflanzenforschung, Gatersleben, Germany; and *T. monococcum* ssp. *flavescens* was donated by Dr. Glynis Jones, University of Sheffield, United Kingdom.

DNA methods: Nucleic acids were extracted from grains of wheat using a modification of the CTAB protocol (Rogers and Bendich 1985) as described by Sallares *et al.* (1995). Polymerase chain reactions (PCRs) were 100 μl in volume and contained 300 ng of each primer, 150 μm each dNTP, 10× buffer (Promega, Madison, WI), 10–100 ng template DNA, and 2.5 units of *Taq* DNA polymerase (Promega). Cycling conditions were as follows: 2 min at 94°; 2 cycles of 2 min at 58°, 1 min at 74°, 1 min at 94°; 2 cycles of 2 min at 57°, 1 min at 74°, 1 min at 94°; 30 cycles of 2 min at 56°, 1 min at 74°, 1 min at 94°; 2 min at 56°; and 8 min at 74°. The primers were the A–C and B–C pairs described in Allaby and Brown (2000). Electrophoresis of PCR products was carried out at 3.33 V cm^{-1} in 3% NuSieve agarose (FMC BioProducts, Rockland, ME). Bands were excised and the DNA recovered by electroelution. PCR products were purified (High Pure PCR product purification kit; Roche Biochemicals, Indianapolis), restricted with *Bam*HI, precipitated with ethanol, ligated into M13mp18, and cloned in *Escherichia coli* XL1-Blue. Single-stranded DNA was prepared from recombinant plaques and sequenced (Sequenase 2.0; Amersham International, Arlington Heights, IL). Sequencing products were electrophoresed in 6% polyacrylamide gels and visualized by autoradiography. The resulting sequences are listed in Table 1, along with additional sequences obtained from the EMBL database. We use the Triticeae nomenclature of Miller (1987).

Network construction: Sequences were aligned using ClustalW (Thompson *et al.* 1994) and the alignments were checked and, where necessary, modified by eye. Nucleotide diversity (Nei and Li 1979) was calculated with Equation 9.5 of Li (1997). Sequences from different spacer size classes were analyzed separately because of the alignment problems (Kellogg and Appels 1995). To construct a network, a group of sequences that share phylogenetically informative characters were identified and these were positioned around a node. The node represents a consensus for that group of sequences, the connection between each sequence and the node indicating the nucleotide difference(s) compared with the consensus. The process was repeated for all groups of sequences in the dataset.

The methodology is illustrated in Figure 1. Network construction with this imaginary alignment would proceed as follows:

Sequence 1 is the consensus of the sequences in the alignment and is therefore used as the starting point (Figure 1A). Network construction is less complicated when the consensus sequence is used as the starting point, but this is not essential and the process can begin with any sequence.

Sequence 2 differs from sequence 1 at four positions and is linked to sequence 1 by a line in the network (Figure 1B).

Sequence 3 differs from sequence 1 at one position, this position being different from any of the substitutions in sequence 2. Sequence 3 is therefore linked directly to sequence 1 (Figure 1C).

Sequence 4 shares with sequence 3 the C → T substitution at position 11, but has an additional substitution at position 2. The line leading to sequence 3 is therefore extended and sequence 4 is placed at its terminus (Figure 1D).

Sequence 5 has one substitution compared with sequence 1, this being a unique substitution not seen so far. Sequence 5 is therefore linked directly to sequence 1 (Figure 1E).

Similarly, sequence 6 has three unique substitutions and is linked directly to sequence 1 (Figure 1F).

Sequence 7 has two differences from sequence 1, one of these being the C → T at position 11 seen in sequences 3 and 4. Sequence 7 therefore connects directly with sequence 3 (Figure 1G). In this part of the network, the line between sequences 1 and 3 represents the substitution at position 11, that between sequences 3 and 4 is the substitution at position 2, and that between sequences 3 and 7 is the substitution at position 24.

Sequence 8 shares with sequence 7 the A → G substitution at position 24 and has a unique substitution, not seen in any previous sequence, at position 9. Sequence 8 must therefore form a branch off of a line between sequences 1 and 7 (Figure 1H). This connection produces an "empty node" indicated by the small closed circle. Empty nodes are either sequences that are present in the 5S arrays being studied but that are not represented in the sequence dataset from which the network is constructed or ancestral sequences that no longer exist in the 5S arrays.

Sequence 9 has eight differences compared with sequence 1, but all are unique (note that the G → T substitution at position 18 is nonidentical with the position 18 substitution in sequence 2). Sequence 9 is therefore directly connected to sequence 1 (Figure 1I).

Sequence 10 shares one substitution with sequence 9 and has two unique ones. It therefore branches off from an empty node on the line between sequences 1 and 9 (Figure 1J).

Finally, sequence 11 has two unique substitutions and so forms a direct connection with sequence 1 (Figure 1K).

Because of character conflicts, assumed to arise from recombination events, individual sequences can be members of more than one group, introducing reticulations into the network and giving rise to "principal nodal sequences," which are defined as nodes to which a substantial number of other nodes and/or sequences are linked in a star-like pattern. This feature is illustrated by the networks shown in Figures 3 and 4.

It is important to recognize the difference between the *topology* of the network (the interconnectivity between different sequences) and its *spatial representation* (the way the network is drawn on paper). The former is important, the latter is not. If the construction is carried out correctly then there is only one possible topology for the network obtained for a particular set of sequences, but that topology can be drawn in many ways. For example, in Figure 1K, sequences 5 and 6 are both connected directly to sequence 1 with no intervening nodes: this is the unique topological relationship between these three sequences. However, the positioning of sequences 5 and 6 around sequence 1 is arbitrary and unimportant: a second spatial representation of Figure 1K could show sequence 6 positioned at "twelve-o'clock" compared to sequence 1 but this would have no effect on the topological relationship between the sequences.

RESULTS

Comparison of repeat types in a single array: *5S-DNA-A1* spacers: The PCRs that we carried out were designed to amplify specific types of *5S-DNA-1* spacer repeats. The B–C primer pair was used with *T. urartu* and polyploid wheats of the AABB and AABBDD lineages, because it

TABLE 1

5S rDNA sequences used in this study

Sequence	Species[a]	Locus	Genome	Accession no.	Reference
		A. *5S-DNA-A1* spacer sequences			
aes1	*Triticum aestivum*	*5S-DNA-A1*	A^u	Z11417	Appels *et al.* (1992)
aes2	*T. aestivum*	*5S-DNA-A1*	A^u	Z11450	Appels *et al.* (1992)
dcm1	*T. dicoccum*	*5S-DNA-A1*	A^u	AJ272302	This study
dco1	*T. dicoccoides*	*5S-DNA-A1*	A^u	AJ272300	This study
dco2	*T. dicoccoides*	*5S-DNA-A1*	A^u	AJ272301	This study
mon1	*T. monococcum*	*5S-DNA-A1*	A^m	Z11461	Baum and Appels (1992)
mon2	*T. monococcum*	*5S-DNA-A1*	A^m	AJ272294	This study
mon3	*T. monococcum*	*5S-DNA-A1*	A^m	AJ272293	This study
mon4	*T. monococcum*	*5S-DNA-A1*	A^m	AJ272295	This study
mon5	*T. monococcum*	*5S-DNA-A1*	A^m	AJ272291	This study
mon6	*T. monococcum*	*5S-DNA-A1*	A^m	AJ272292	This study
mon7	*T. monococcum*	*5S-DNA-A1*	A^m	AJ272290	This study
sin1	*T. sinskajae*	*5S-DNA-A1*	A^m	AJ272288	This study
sin2	*T. sinskajae*	*5S-DNA-A1*	A^m	AJ272289	This study
sin3	*T. sinskajae*	*5S-DNA-A1*	A^m	AJ272287	This study
ura1	*T. urartu*	*5S-DNA-A1*	A^u	AJ272298	This study
ura2	*T. urartu*	*5S-DNA-A1*	A^u	AJ272299	This study
ura3	*T. urartu*	*5S-DNA-A1*	A^u	AJ272296	This study
ura4	*T. urartu*	*5S-DNA-A1*	A^u	AJ272282	This study
ura5	*T. urartu*	*5S-DNA-A1*	A^u	AJ272284	This study
ura6	*T. urartu*	*5S-DNA-A1*	A^u	AJ272285	This study
ura7	*T. urartu*	*5S-DNA-A1*	A^u	AJ272286	This study
ura8	*T. urartu*	*5S-DNA-A1*	A^u	AJ272283	This study
		B. *5S-DNA-2* spacer sequences			
aes3	*T. aestivum*	*5S-DNA-2*	A^u, B or D	Z11423	Appels *et al.* (1992)
aes4	*T. aestivum*	*5S-DNA-2*	A^u, B or D	Z11424	Appels *et al.* (1992)
aes5	*T. aestivum*	*5S-DNA-2*	A^u, B or D	Z11425	Appels *et al.* (1992)
aes6	*T. aestivum*	*5S-DNA-2*	A^u, B or D	Z11426	Appels *et al.* (1992)
aes7	*T. aestivum*	*5S-DNA-2*	A^u, B or D	Z11427	Appels *et al.* (1992)
aes8	*T. aestivum*	*5S-DNA-2*	A^u, B or D	X66388	Vakhitov *et al.* (1989a)
mon8	*T. monococcum*	*5S-DNA-2*	A^m	Z11460	Baum and Appels (1992)
mon9	*T. monococcum*	*5S-DNA-2*	A^m	X66383	Vakhitov *et al.* (1989a)
mon10	*T. monococcum*	*5S-DNA-2*	A^m	X66391	Vakhitov *et al.* (1989b)
sea1	*Aegilops searsii*	*5S-DNA-2*	S^s	Z11462	Baum and Appels (1992)
sha1	*Ae. sharonensis*	*5S-DNA-2*	S^l	Z11481	Baum and Appels (1992)
spe1	*Ae. speltoides*	*5S-DNA-2*	S	X66387	Vakhitov *et al.* (1989a)
spe2	*Ae. speltoides*	*5S-DNA-2*	S	Z11464	Baum and Appels (1992)
squ1	*Ae. squarrosa*	*5S-DNA-2*	D	Z11465	Appels *et al.* (1992)
squ2	*Ae. squarrosa*	*5S-DNA-2*	D	X66381	Vakhitov *et al.* (1989a)
tim1	*T. timopheevi*	*5S-DNA-2*	A^u or G	X66385	Vakhitov *et al.* (1989a)
umb1	*Ae. umbellulata*	*5S-DNA-2*	U	Z11479	Baum and Appels (1992)
ura9	*T. urartu*	*5S-DNA-2*	A^u	X66384	Vakhitov *et al.* (1989a)
		C. 5S rDNA genes			
aes9	*T. aestivum*	*5S-DNA-1*	B or D	Z11415	Appels *et al.* (1992)
aes10	*T. aestivum*	*5S-DNA-A1*	A^u	Z11416	Appels *et al.* (1992)
aes11	*T. aestivum*	*5S-DNA-1*	B or D	Z11418	Appels *et al.* (1992)
aes12	*T. aestivum*	*5S-DNA-1*	B or D	Z11419	Appels *et al.* (1992)
aes13	*T. aestivum*	*5S-DNA-1*	B or D	Z11421	Appels *et al.* (1992)
aes14	*T. aestivum*	*5S-DNA-1*	B or D	Z11428	Appels *et al.* (1992)
aes15	*T. aestivum*	*5S-DNA-A1*	A^u	Z11449	Appels *et al.* (1992)
aes16	*T. aestivum*	*5S-DNA-1*	B or D	Z11454	Appels *et al.* (1992)
squ3	*Ae. squarrosa*	*5S-DNA-1*	D	Z11466	Appels *et al.* (1992)

Gene sequences were also used in C from the following units whose spacers are listed in A and B: aes1, aes2, aes3, aes4, aes5, aes6, aes7, aes8, mon1, mon8, mon9, mon10, sea1, sha1, spe1, spe2, squ1, squ2, tim1, umb1, and ura9.

[a] According to the nomenclature of Miller (1987).

```
   1   5   10   15   20   25   30   35   40
1  GCATTACTTTCGGTAGCGCGGAAAGGCGACGGGACTTATA
2  ............C...C.......C.....A........
3  ..........T.............................
4  .T........T.............................
5  ......................T..................
6  ......T.....C......................G......
7  ..........T............G.................
8  ........C..............G.................
9  ...C-..........T.TAA.......CC............
10 ..................A..G...............T..
11 ............T......................A....
```

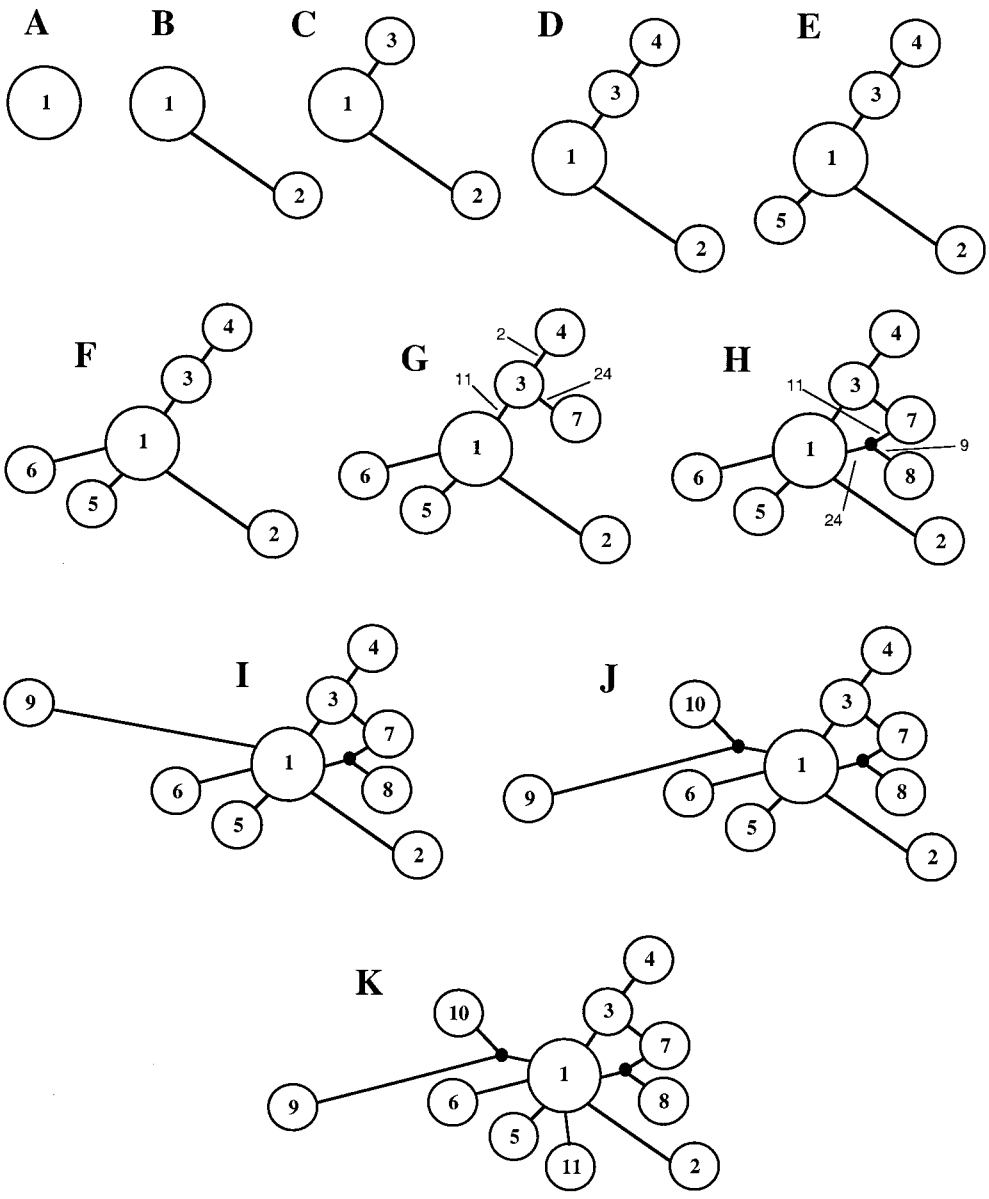

FIGURE 1.—Illustration of the methodology used for network construction. The alignment (top) contains 11 imaginary sequences with dots indicating identities with sequence 1 and dashes indicating deletions. A–K illustrate the steps in network construction, as described in the text.

specifically amplifies a repeat type that is found only on the A^u genome; it does not give products with *T. monococcum* or other wheats containing the A^m genome (ALLABY and BROWN 2000). The A–C pair, which is plesiomorphic and amplifies spacers of all *5S-DNA-1* loci, was used in this project to obtain additional sequences from *T. urartu*, as well as sequences from the A^m-type genomes of diploid Triticum species. The sequences that were obtained, plus three published sequences that were included in this part of the study, are listed in

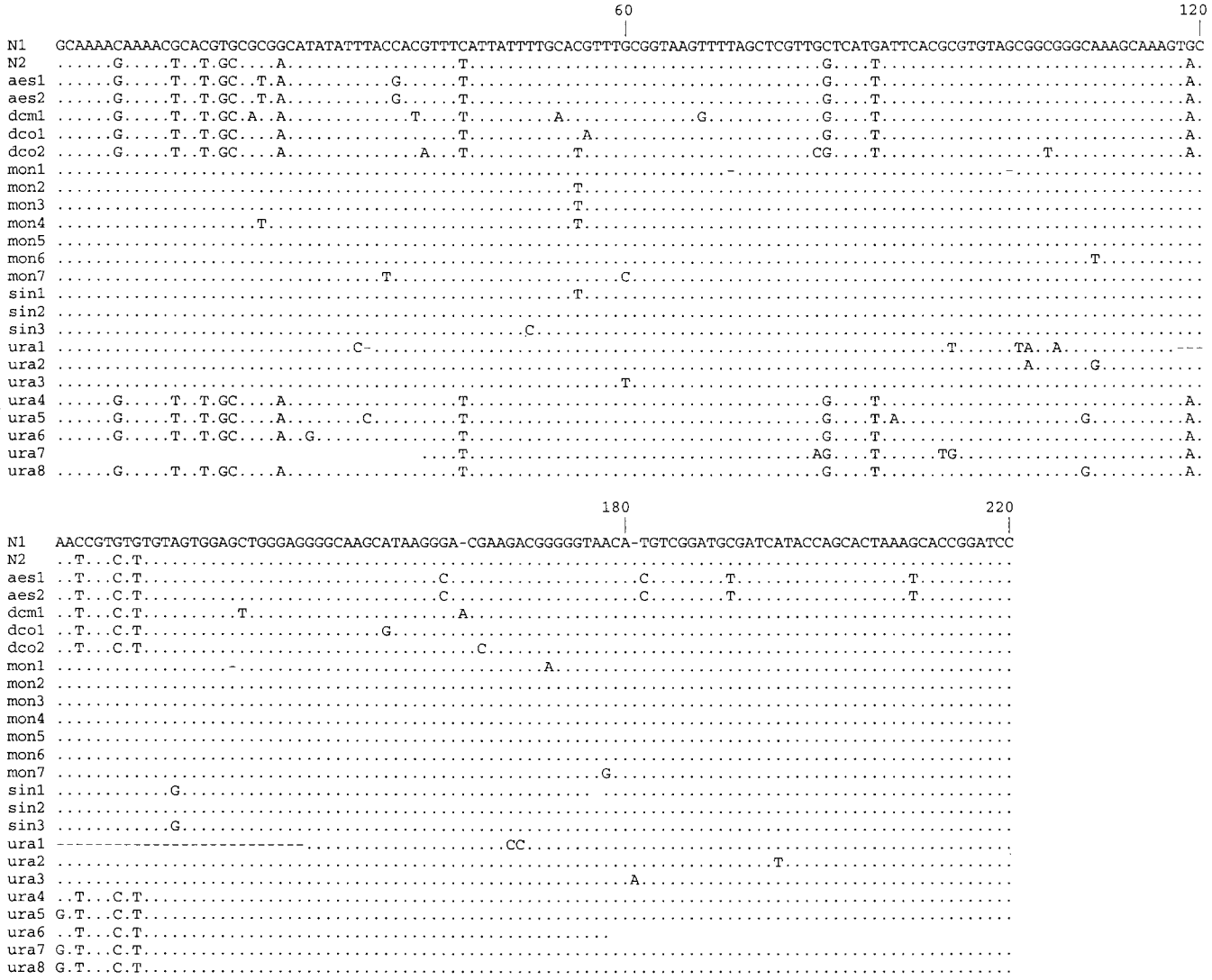

FIGURE 2.—Alignment of the *5S-DNA-A1* spacer sequences. Sequence names are defined in Table 1A. Dots indicate identities with the principal nodal sequence N1 (see Figure 3), dashes indicate deletions, and empty spaces are unsequenced regions.

Table 1A and aligned in Figure 2. The network constructed from these sequences is shown in Figure 3.

The network contains two principal nodal sequences, N1 and N2, which correspond to the spacer types amplified by the A–C and B–C primer pairs, respectively. Three of the 23 sequences (13%) are located at these nodes. Both principal nodes support star-like phylogenies, indicating that they represent ancestral sequences from which there are multiple apomorphies. The group of sequences associated with the N1 node contains all the sequences obtained from A^m genomes, along with three from the A^u genome of *T. urartu*. The N2 group contains the other A^u sequences, including the five sequences from polyploid wheats. The two nodal sequences are relatively distant (13 substitutions in the 220-bp alignment), implying that each is of considerable antiquity, and the presence of just three branches between the two parts of the network indicates that the N1 and N2 groups have evolved with a large degree of independence from one other.

The presence of both A^m and A^u sequences in the N1 group suggests that some of these sequences predate the split between the two genomes. It is therefore appropriate to subdivide the sequence data from a single array into the N1 and N2 groups and analyze each paralogous group separately when making an orthologous comparison between the spacer sequences of different genomes. This conclusion has important implications. The π-value calculated from the 10 A^m sequences included in our study is 0.0104 ± 0.0059, and the value for the 13 A^u sequences (those associated with both the N1 and N2 nodes) is 0.0583 ± 0.036. The mean intragenomic diversity of 0.0344 ± 0.017 is significantly less than the mean nucleotide diversity between the two genomes, which is 0.0589 ± 0.030. When analyzed in this way, the results support the conclusion of KELLOGG and APPELS

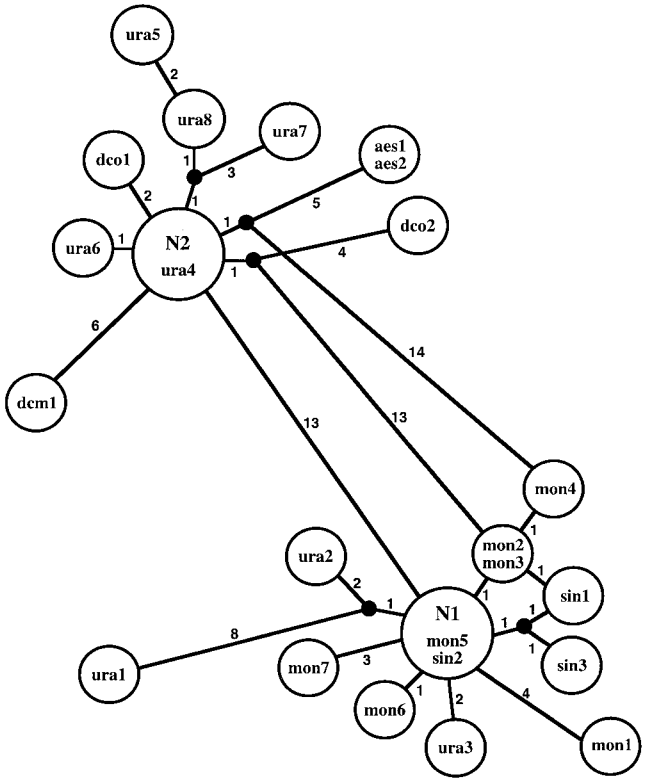

FIGURE 3.—Network of *5S-DNA-A1* spacer sequences. Each sequence is represented by an open circle except for the two principal nodal sequences (N1 and N2), which are shaded. Each interconnecting line has an attached number, which is the number of nucleotide differences between the two sequences linked by that line.

(1995) that wheat 5S rDNA spacers display higher nucleotide diversity between rather than within species and suggest that this pattern holds true even when closely related genomes are compared. However, the analysis is erroneous, because it fails to take account of the paralogous nature of the N1 and N2 groups of spacer sequences. A truly orthologous comparison can be made if the N1 sequences are analyzed on their own. The π-value for the A^m sequences remains at 0.0104 $\pm$ 0.0059, because all the A^m sequences are in the N1 group, but the π-value for the A^u sequences is now 0.0354 $\pm$ 0.012. The mean intragenomic nucleotide diversity is 0.0229 $\pm$ 0.0088 and the intergenomic value is 0.0237 $\pm$ 0.0151. These two values are closely similar, casting doubt on the hypothesis that spacers show greater inter- compared with intragenomic variation and suggesting that the pattern of spacer diversity is, in fact, the same as that observed for the gene sequences (KELLOGG and APPELS 1995), diversity between species being about equal to diversity within a species. The discrepancy is clearly due to the inaccuracies introduced into the nucleotide diversity values when nonorthologous sequences are included in the comparison.

Comparison of sequences from homeologous loci: *5S-DNA-2* spacers: Spacer regions from 18 published *5S-DNA-2* sequences (Table 1B) were aligned. Nucleotide diversities were calculated for spacers from taxa represented by two or more sequences and mean pairwise differences between pairs of taxa were determined. *T. aestivum* sequences were excluded from this analysis because these might be on different genomes. The diversities within taxa (0.0514–0.1502) covered a similar range to the diversities between taxa (0.0905–0.1664), and in several cases the diversity within a taxon was greater than the diversity seen when that taxon was compared with a second taxon.

The network constructed from the *5S-DNA-2* spacer sequences (Figure 4) is complex, due to a high incidence of conflicting characters. The network divides into three segments, each segment comprising a star-like phylogeny associated with one of three principal nodal sequences, C1, C2, and C3. The C1 group includes sequences from the diploids *T. urartu* (A^uA^u genomes), *Ae. speltoides* (SS), and *Ae. squarrosa* (DD), whose genomes are thought to be ancestral to the A, B, and D genomes of hexaploid *T. aestivum*, along with sequences from *T. monococcum* (A^mA^m) and *T. timopheevi* (A^uA^uGG). None of the reticulatory branches indicating differences from the C1 sequence are fixed to any one species, suggesting that these nucleotide changes occurred in the common ancestor of the A, S, and D genomes. From our analysis of HMW glutenin gene diversity (ALLABY *et al.* 1999), we have estimated that the A, B, D, and G genomes diverged ~6 million years ago, so the genetic diversity described by the reticulations of the C1 part of the network appears to have been maintained for at least this length of time.

The C2 group contains three additional sequences from the S genomes of members of the Sitopsis section of Aegilops (*Ae. searsii*, *Ae. sharonensis*, and *Ae. speltoides*) as well as sequences from *Ae. squarrosa* (DD), *Ae. umbellulata* (UU), and *T. aestivum*. The C3 node is ancestral to three sequences, two from *T. monococcum* and one from *T. aestivum*. Reticulations occur between the three nodal clusters but with much less frequency than within individual clusters.

Sequences from *T. aestivum* are linked exclusively to the C2 and C3 nodes, with the possible exception of aes8, which is almost equidistant between C1 and C2, being one nucleotide closer to C2 but closely affiliated with the *T. urartu* sequence ura9, which is clearly a part of the C1 cluster. Because of the size of the dataset (18 sequences represent <0.1% of the repeats in a single 5S rDNA array), the results are subject to sampling errors, but the binomial probability is 0.016 (significant at the 5% level) of obtaining no *T. aestivum* sequences in the C1 cluster if these sequences are evenly distributed between the three clusters, suggesting that the distribution shown in Figure 4 is genuine. In contrast, 3 of the 6 sequences from genomes ancestral to *T. aestivum* are associated with the C1 cluster, and the other three are in the C2 cluster.

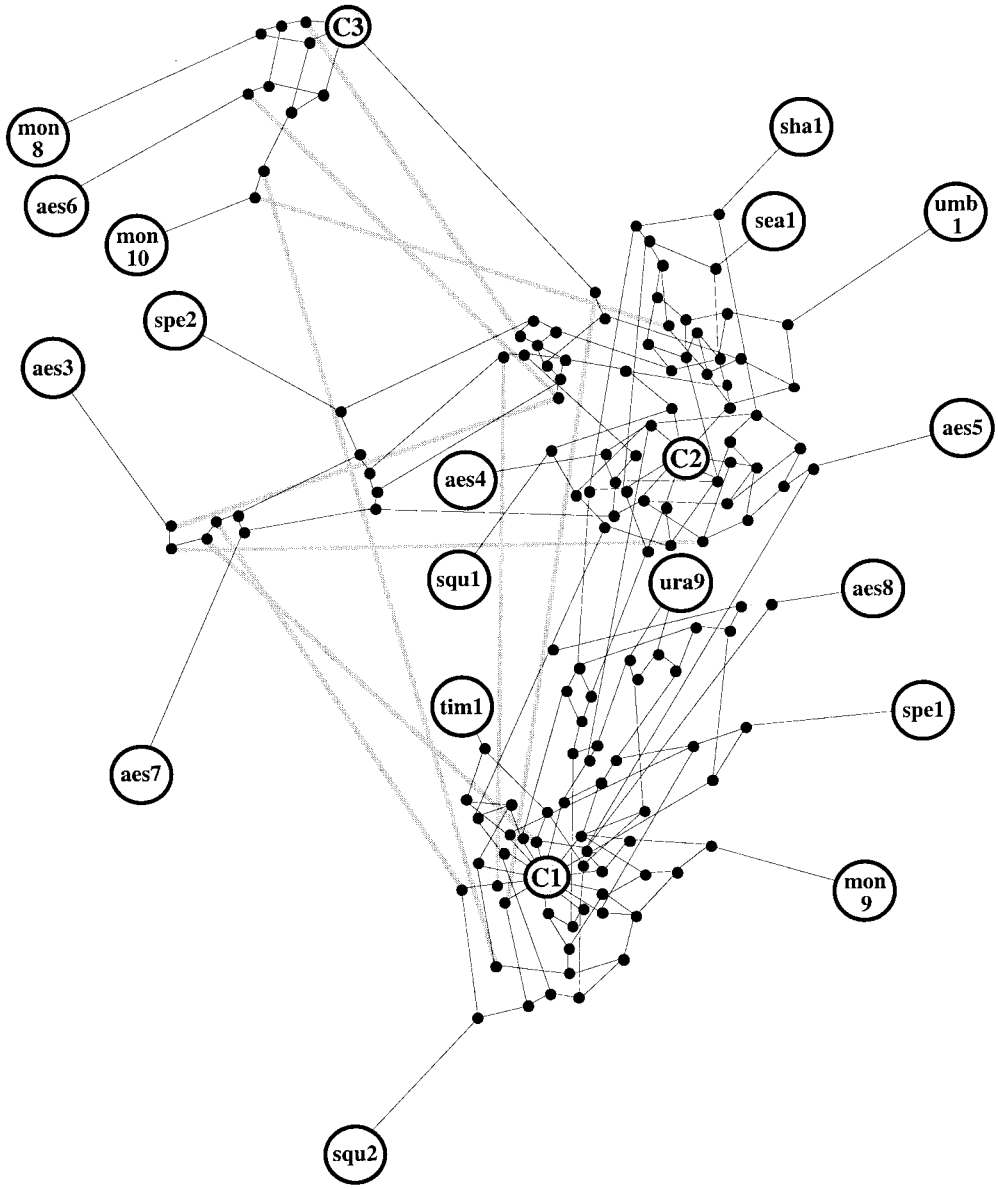

FIGURE 4.—Network of the *5S-DNA-2* spacer sequences. Each sequence is represented by an open circle and the three principal nodal sequences (C1, C2, and C3) are depicted by shaded circles. For clarity, the numbers of nucleotide differences are not marked on the interconnecting lines (full details are available from the authors). The lengths of the thin lines are roughly proportional to the numbers of nucleotide differences that they represent, though there are several places where the topology (which in effect is three-dimensional) cannot be accurately drawn in two dimensions. Shaded lines denote more distant relationships.

There have been attempts to use genetic distance to allocate 5S rDNA sequences to particular loci. For example, APPELS *et al.* (1992) affiliated aes4 with the D genome of hexaploid wheats because of its similarity with squ1. Our analysis suggests that allocations made on this basis are unreliable, because in most cases apomorphisms that are not exclusive to a single sequence are also not exclusive to a single genome. The network roots aes4 and squ1 at a node that is just one nucleotide different from the C2 sequence, showing that aes4 and squ1 are quite plesiomorphic and do not share a single apomorphy to the exclusion of all other sequences in the network. In fact, aes4 shares as many sequence features with a group that includes the two S genome sequences sea1 and sha1 and so a B genome origin could be argued. We suggest that the accurate and informative description of aes4 is "C2 repeat type."

Comparison of paralogous and orthologous loci: 5S rDNA genes: An accurate alignment of the spacers of the *5S-DNA-1* and *5S-DNA-2* loci is not possible and, because of frequent indels, it is difficult even to align the two size classes of spacers corresponding to the *5S-DNA-1* units on chromosomes 1A and 1B/1D (ALLABY and BROWN 2000). To compare these various loci, it is necessary to examine the gene sequences rather than the spacers. The network for 30 published *5S-DNA-1* and *5S-DNA-2* gene sequences (Table 1C) is shown in Figure 5. As with the other two networks, there is a principal nodal sequence, which in this case corresponds to the correct (*i.e.*, functional) copy of the 5S rRNA gene (which has the same sequence in Triticum and Aegilops species), at the center of a star-like phylogeny. Four of the 30 sequences (13.3%) in the dataset are correct copies of the gene and hence are located at the node.

The network shows that gene sequences from a single

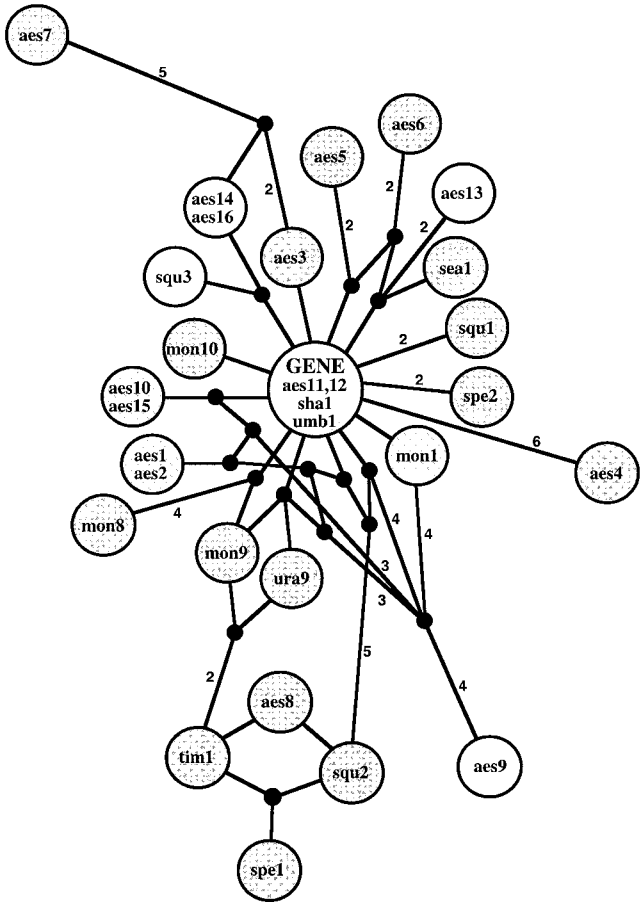

FIGURE 5.—Network of the gene sequences. The largest open circle is the correct gene sequence and the smaller open circles are variants. These are shaded as follows: dark stippling, genes from *5S-DNA-2* loci; light cross-hatching, genes from *5S-DNA-A1* loci; no shading, genes from *5S-DNA-B1* and *5S-DNA-D1* loci. Numbers indicate the numbers of nucleotide differences between pairs of sequences; for clarity, lines linking sequences with just a single nucleotide difference do not have numbers attached.

locus in a single species do not group together. For example, the two sequences from *Ae. speltoides*, spe1 and spe2, are at distant parts of the network and both are closely related to different sequences from *Ae. squarrosa*, squ2 and squ1, respectively, implying that the origin of these sequence types predates the common ancestor of these two species. The reticulations between *T. aestivum* sequences from different loci can be explained in three ways: mutations in the common ancestor of the genomes that are involved, parallel mutations in the different genome lineages, or homogenization between the *5S-DNA-1* and *5S-DNA-2* loci.

DISCUSSION

The network analysis describes relationships between 5S rDNA sequences in a more realistic fashion than conventional tree building because network analysis makes fewer assumptions about the direction of evolution, the extent of sexual isolation, and the pattern of ancestry and descent. The networks presented here show that the 5S rDNA sequences of Triticum and Aegilops species are related in a reticulate manner around principal nodal sequences. The spacer networks have multiple principal nodes of considerable antiquity, but the gene network has just one principal node corresponding to the functional gene sequence.

Principal nodal sequences and their star-like phylogenies: A principal nodal sequence represents a root within a network and corresponds to the common ancestor of the sequences in the associated star-like phylogeny. These nodal sequences appear to be of considerable antiquity. The longevity is most clearly apparent in the *5S-DNA-A1* spacer network (Figure 3), the average genetic distance between the extant spacer sequences and the nodal sequence being 0.01060 for the N1 group and 0.01403 for the N2 group, implying divergence times of 1.1–1.8 and 1.4–2.5 million years ago, respectively. The possibility that recombination between adjacent repeat types (LASSNER and DVORÁK 1986) eliminates some new mutations forces the conclusion that the real age of the N1 and N2 nodal sequences is actually much greater than these values. The C1, C2, and C3 nodal sequences of the *5S-DNA-2* spacer network (Figure 4) represent the roots of larger families of spacer sequences, encompassing all the major genomes of Triticum and Aegilops. This network clearly shows that the majority of nucleotide substitutions are not restricted to a single species, which again implies that the genetic variation indicated by the network, and consequently the nodal sequences, are maintained on a long-term basis. KELLOGG and APPELS (1995) have highlighted the paradox that there appears to be little selection on any particular gene in an array although conservation of gene sequences across species implies that selection of some description is occurring. Our results indicate that a similar paradox applies to the spacer sequences: the extant sequences are highly variable, showing that, as expected for a nonfunctional region of DNA, there is little if any selection acting on individual spacers, but the existence of nodal consensus sequences of considerable antiquity indicates that some type of conservation process is in operation.

One important feature of network analysis is that it enables orthologous and paralogous groups of spacer sequences to be distinguished. This reveals that it is possible for an ortholog group to be lost from an array, as appears to have occurred with the N1 group of *5S-DNA-A1* spacers during the evolution of polyploid wheats. The differential loss of ortholog groups can lead to overestimation of genetic diversity, but the networks indicate which sequences should be used to achieve a genuine orthologous comparison.

Evolution of 5S rDNA arrays: The 5S rDNA spacer region is relatively devoid of function, compared with

the gene sequence, so the functional constraint on the spacers will be relatively small. However, our study shows that the genes and spacers have important aspects in common, both evolving in a manner that involves little direct selection on individual sequences but that results in conservation of consensus sequences of considerable antiquity and leads to similar patterns of intra- and interspecific diversity for both components of the repeat unit. These observations suggest that the gene and spacer sequences are evolving together. An insight into this evolutionary process might be provided by the nature of the principal nodal sequences. In the networks for the *5S-DNA-A1* spacers and the gene sequences, ~13% of the extant sequences are identical to a principal nodal sequence. Arguably, an inherent feature of a tandem array is the conservation of sequence types, presumably through recombination, regardless of selection pressure. Viewed in this way, a tandem array is a dual-acting system that combines inherent sequence conservation with a large store of mutant variations. The implication is that the number of correct gene copies within an array is essentially the result of neutral processes that act on both gene and spacer regions, this interpretation in part explaining why fixation of correct gene copies within an array has not been observed. The molecular mechanisms that underlie these neutral processes could include homogenization by crossing over and replication slippage, contrary to the conclusion of Kellogg and Appels (1995) that these mechanisms are not of significant importance in 5S rDNA gene evolution.

The only difference between the spacer and gene sequences is that the consensus sequences of the spacer regions are probably arbitrary sequences, whereas the consensus sequence of the gene is functionally constrained. Exactly how natural selection acts on the consensus sequences of the 5S rDNA genes, or how such a largely degenerate array functions efficiently, remains unclear. These questions cannot be rigorously addressed until it has been shown whether the functional 5S rRNAs are indeed homogeneous, or if minor variants corresponding to some or all of the "incorrect" sequences are also active.

Founder effects during 5S rDNA evolution in polyploid wheats: We discovered two examples of repeat-type bias for spacer sequences in polyploid wheats. First, the *5S-DNA-A1* sequences of polyploid wheats were all of the N2 type, although both N1 and N2 types were present in *T. urartu*, the diploid donor of the A^u genome of these wheats (Figure 3). Second, *5S-DNA-2* spacer sequences of the C1 type were well represented in diploid species containing genomes ancestral to the polyploid wheats, but no *T. aestivum* sequences of this type were seen (Figure 4). These biases probably indicate that bottlenecks have occurred during the evolution of polyploid wheats, there being a greater opportunity for sequences to be fixed or lost if the population size is relatively small. These founder effects, which can be detected by network analysis, are potentially important for understanding the dynamics of polyploidization and inferring the events involved in evolution of the hexaploid bread wheats, which do not exist in the wild, subsequent to the origin of agriculture.

We thank Glynis Jones, University of Sheffield, for donating plant material. This work was supported by a grant from the UK Natural Environment Research Council.

LITERATURE CITED

Allaby, R. G., and T. A. Brown, 2000 Identification of a 5S rDNA spacer type specific to *Triticum urartu* and wheats containing the *T. urartu* genome. Genome **43:** 250–254.

Allaby, R. G., M. Banerjee and T. A. Brown, 1999 Evolution of the high-molecular-weight glutenin loci of the A, B, D and G genomes of wheat. Genome **42:** 296–307.

Appels, R., B. R. Baum and B. C. Clarke, 1992 The 5S DNA units of bread wheat (*Triticum aestivum*). Plant Syst. Evol. **183:** 183–194.

Bandelt, H.-J., P. Forster, B. C. Sykes and M. B. Richards, 1995 Mitochondrial portraits of human populations using median networks. Genetics **141:** 743–753.

Bandelt, H.-J., P. Forster and A. Röhl, 1999 Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. **16:** 37–48.

Baum, B. R., and R. Appels, 1992 Evolutionary change at the 5S DNA locus of species in the *Triticeae*. Plant Syst. Evol. **183:** 195–208.

Berry, A., and M. Kreitman, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134:** 869–893.

Cox, A. V., M. D. Bennett and T. A. Dyer, 1992 Use of the polymerase chain reaction to detect spacer size heterogeneity in plant 5S-rRNA gene clusters and to locate such clusters in wheat (*Triticum aestivum* L.). Theor. Appl. Genet. **83:** 684–690.

Dinman, J. D., and R. B. Wickner, 1995 5S rRNA is involved in fidelity of translational reading frame. Genetics **141:** 95–105.

Dover, G. A., 1982 Molecular drive: a cohesive mode of species evolution. Nature **299:** 111–117.

Dover, G. A., 1986 Molecular drive in multigene families: how biological novelties arise, spread and are assimilated. Trends Genet. **2:** 159–165.

Dubcovsky, J., and J. Dvorák, 1995 Ribosomal RNA multigene loci: nomads of the *Triticeae* genomes. Genetics **140:** 1367–1377.

Dvorák, J., H.-B. Zhang, R. S. Kota and M. Lassner, 1989 Organization and evolution of the 5S ribosomal RNA gene family in wheat and related species. Genome **32:** 1003–1015.

Erdmann, V. A., and J. Wolters, 1986 Collection of published 5S, 5.8S and 4.5S ribosomal RNA sequences. Nucleic Acids Res. **14** (Suppl.): r1–59.

Excoffier, L., P. E. Smouse and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

Ford, P. J., and E. M. Southern, 1973 Different sequences for 5S RNA in kidney cells and ovaries in *Xenopus laevis*. Nature **241:** 7–10.

Forster, P., R. Harding, A. Torroni and H.-J. Bandelt, 1996 Origin and evolution of native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. **59:** 935–945.

Gerlach, W. L., and T. A. Dyer, 1980 Sequence organisation of the repeating units in the nucleus of wheat which contain 5S rRNA genes. Nucleic Acids Res. **8:** 4851–4865.

Goldsborough, P. B., T. H. N. Ellis and C. A. Cullis, 1981 Organisation of the 5S RNA genes in flax. Nucleic Acids Res. **9:** 5895–5904.

Hillis, D. M., C. Moritz, C. A. Porter and R. J. Baker, 1991 Evidence for biased gene conversion in concerted evolution of ribosomal DNA. Science **252:** 308–310.

Kanazin, V., E. Ananiev and T. Blake, 1993 The genetics of 5S

rRNA encoding multigene families in barley. Genome **36:** 1023–1028.

Kellogg, E. A., and R. Appels, 1995 Intraspecific and interspecific variation in 5S RNA genes are decoupled in diploid wheat relatives. Genetics **140:** 325–343.

Kolchinsky, A., V. Kanazin, E. Yakovleva, A. Gazumyan, C. Kole *et al.*, 1990 5S-RNA genes of barley are located on the second chromosome. Theor. Appl. Genet. **80:** 333–336.

Lagudah, E. S., B. C. Clarke and R. Appels, 1989 Phylogenetic relationships of *Triticum tauschii*, the D-genome donor to hexaploid wheat. 4. Variation and chromosomal location of 5S DNA. Genome **32:** 1017–1025.

Lassner, M. W., and J. Dvorák, 1986 Preferential homogenisation between adjacent and alternate subrepeats in wheat rDNA. Nucleic Acids Res. **14:** 5499–5512.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer, Sunderland, MA.

Miller, T. E., 1987 Systematics and evolution, pp. 1–30 in *Wheat Breeding—Its Scientific Basis*, edited by F. G. H. Lupton. Chapman & Hall, London.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nei, M., and W.-H. Li, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

Prim, R. C., 1957 Shortest connection networks and some generalisations. Bell Syst. Tech. J. **36:** 1389–1401.

Reddy, P., and R. Appels, 1989 A second locus for the 5S multigene family in *Secale* L.: sequence divergence in two lineages of the family. Genome **32:** 456–467.

Richards, M., H. Corte-Real, P. Forster, V. Macaulay, H. Wilkinson-Herbots *et al.*, 1995 Palaeolithic and Neolithic lineages in the European mitochondrial gene pool. Am. J. Hum. Genet. **59:** 185–203.

Rogers, S. O., and A. J. Bendich, 1985 Extraction of DNA from milligram amounts of fresh herbarium and mummified plant tissues. Plant Mol. Biol. **5:** 69–76.

Sallares, R., R. G. Allaby and T. A. Brown, 1995 PCR-based identification of wheat genomes. Mol. Ecol. **4:** 509–514.

Sastri, D. C., K. Hilu, R. Appels, E. S. Lagudah, J. Playford *et al.*, 1992 An overview of evolution in plant 5S DNA. Plant Syst. Evol. **183:** 169–181.

Schlötterer, C., and D. Tautz, 1994 Chromosomal heterogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr. Biol. **4:** 777–783.

Scoles, G. J., B. S. Gill, Z.-Y. Xin, B. C. Clarke, C. L. McIntyre *et al.*, 1988 Frequent duplication and deletion events in the 5S RNA genes and associated spacer regions of the *Triticeae*. Plant Syst. Evol. **160:** 105–122.

Simmons, G. M., M. E. Kreitman, W. F. Quattlebaum and N. Miyashita, 1989 Molecular analysis of the alleles of alcohol dehydrogenase along a cline in *Drosophila melanogaster*. I. Maine, North Carolina and Florida. Evolution **43:** 393–409.

Sykes, B., A. Leiboff, J. Low-Beer, S. Tetzner and M. Richards, 1995 The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. Am. J. Hum Genet. **57:** 1463–1475.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Torroni, A., H.-J. Bandelt, L. D'Urbano, P. Lahermo, P. Moral *et al.*, 1998 mtDNA analysis reveals a major late palaeolithic population expansion from southwestern to northeastern Europe. Am. J. Hum. Genet. **62:** 1137–1152.

Vakhitov, V. A., and Yu. M. Nikonorov, 1989a Nucleotide sequences of polyploid wheat species and *Aegilops* species. Mol. Biol. **23:** 431–440.

Vakhitov, V. A., and Yu. M. Nikonorov, 1989b Nucleotide sequence of large 5S repeat in diploid wheat *Triticum monococcum*. Biopolym. Cell **5:** 58–62.

Communicating editor: M. A. Asmussen