Heredity (2012) 108, 500–506

www.nature.com/hdy

Heredity

ORIGINAL ARTICLE

Types, levels and patterns of low-copy DNA sequence divergence, and phylogenetic implications, for *Gossypium* genome types

J Rong[1,2], X Wang[1,3], SR Schulze[1], RO Compton[1], TD Williams-Coplin[1], V Goff[1], PW Chee[4] and AH Paterson[1]

To explore types, levels and patterns of genetic divergence among diploid Gossypium (cotton) genomes, 780 cDNA, genomic DNA and simple sequence repeat (SSR) loci were re-sequenced in *Gossypium herbaceum* (A1 genome), *G. arboreum* (A2), *G. raimondii* (D5), *G. trilobum* (D8), *G. sturtianum* (C1) and an outgroup, *Gossypioides kirkii*. Divergence among these genomes ranged from 7.32 polymorphic base pairs per 100 between *G. kirkii* and *G. herbaceum* (A1) to only 1.44 between *G. herbaceum* (A1) and *G. arboreum* (A2). SSR loci are least conserved with 12.71 polymorphic base pairs and 3.77 polymorphic sites per 100 base pairs, whereas expressed sequence tags are most conserved with 3.96 polymorphic base pairs and 2.06 sites. SSR loci also exhibit the highest percentage of 'extended polymorphisms' (spanning multiple consecutive nucleotides). The A genome lineage was particularly rapidly evolving, with the D genome also showing accelerated evolution relative to the C genome. Unexpected asymmetry in mutation rates was found, with much more transition than transversion mutation in the D genome after its divergence from a common ancestor shared with the A genome. This large quantity of orthologous DNA sequence strongly supports a phylogeny in which A–C divergence is more recent than A–D divergence, a subject that is of much importance in view of A–D polyploid formation being key to the evolution of the most productive and finest-quality cottons. Loci that are monomorphic within A or D genome types, but polymorphic between genome types, may be of practical importance for identifying locus-specific DNA markers in tetraploid cottons including leading cultivars.
Heredity (2012) **108,** 500–506; doi:10.1038/hdy.2011.111

Keywords: DNA diversity; SNP; evolution; speciation; cotton

INTRODUCTION

A suite of morphological and physiological adaptations that have permitted different *Gossypium* taxa to flourish in a wide range of natural environments offer potential solutions to a host of challenges to cotton cultivation, but the genetic basis of many of these adaptations remains to be discovered. About 45 diploid *Gossypium* species are recognized, classified into eight groups with genomes A–G and K based on chromosome paring in interspecific hybrids, chromosome size and geographical distribution (Beasley, 1940; Phillips and Strickland, 1966; Edwards and Mirza, 1979; Endrizzi *et al.*, 1985). The five tetraploid cotton species, including the two most important cultivated cottons, are all derived from a natural cross between A genome and D genome ancestors resembling the extant *G. herbaceum* and *G. raimondii*, about 1–2 million years ago (Wendel *et al.*, 1989; Wendel and Albert, 1992; Senchina *et al.*, 2003; Wendel and Cronn, 2003).

Genetic changes at the DNA level are the foundation of organic evolution. Both single-nucleotide substitution and insertion/deletion of multiple nucleotides are caused by a variety of genetic mechanisms, and their study provides insight into relationships among taxa, as well as rates and mechanisms of divergence, also providing DNA polymorphisms, which can be used in forward and association genetics studies. The *Gossypium* genus is an attractive model for studying the genetic changes associated with polyploid evolution, with a natural interspecific cross between taxa that remain extant followed by chromosome doubling having spawned five extant tetraploid species. Substantial divergence among the diploid genomes provides a backdrop for polyploid cotton evolution—for example, a twofold difference in genome size between the A and D genome progenitors is due largely to differential accumulation of retroelements in the A genome (Zhao *et al.*, 1995, 1998; Grover *et al.*, 2007) and biased accumulation of small deletions in the D genome (Grover *et al.*, 2007), with intergenomic exchange of these elements following polyploid formation (Wendel *et al.*, 1995; Cronn *et al.*, 1996; Hanson *et al.*, 1998; Zhao *et al.*, 1998).

An important element of invigorated efforts to relate DNA level information to its phenotypic consequences in cultivated (largely tetraploid) cottons and other polyploid crops, is to improve understanding of the rates and mechanisms of divergence among diploid relatives, especially those that contributed to polyploid formation. Although the diploid progenitors of polyploid cottons are clearly identified, the exact relationships among these and other diploids have remained controversial (Seelanan *et al.*, 1997; Cronn *et al.*, 2002), in part because only small numbers of genetic loci have been studied and evolutionary rates of different genes vary widely (Senchina *et al.*, 2003).

[1]Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA; [2]School of Agriculture and Food Science, Zhejiang A & F University, Linan, Zhejiang, PR China; [3]College of Sciences, Hebei Polytechnic University, Tangshan, Hebei, PR China and [4]Department of Crop and Soil Science, University of Georgia, Tifton, GA, USA
Correspondence: Dr AH Paterson, Plant Genome Mapping Laboratory, University of Georgia, 111 Riverbend Road, Room 228, Athens, GA 30602, USA.
E-mail: paterson@uga.edu
Received 6 August 2011; revised 22 September 2011; accepted 29 September 2011

Here, we explore types, levels and patterns of DNA polymorphism among a strategic sampling of diploid cottons, using DNA sequencing to characterize hundreds of orthologous loci derived from sequence-tagged sites that have previously been anchored to a cotton genetic map and broadly sample the genome (Rong *et al*., 2004).

MATERIALS AND METHODS

Plant materials and DNA extraction

The plant materials used in this experiment were two *Gossypium* A genome species (*G. herbaceum*, A1-97 and *G. arboreum*, A2-47), two D genome species (*G. raimondii*, D5-1 and *G. trilobum*, D8-1), one C genome species (*G. sturtianum,* C1-4) and one out-group (*Gossypioides kirkii*). Genomic DNA used for sequence amplification and analysis was extracted as reported elsewhere (Rong *et al*., 2004).

Oligonucleotide primer design

The DNA sequences used to design primers come from 15 sources that include 4 classes: expressed sequence tag (EST), simple sequence repeat (SSR), low-copy genomic DNA (gDNA) and Cot-filtered non-coding (CFNC) DNA (Supplementary Table 1). Most ESTs, gDNA and SSRs are from genetically mapped sequences (Rong *et al*., 2004). CFNC sequences were extracted from *G. raimondii* DNA using renaturation kinetics to obtain putatively low-copy DNA then sequencing random clones and filtering out the subset containing recognizable functional domains (Rong *et al*., 2011). The eprimer3 interface to the primer3 software program (Rozen and Skaletsky, 2000) was used to design primers in batches. Low complexity or microsatellite regions were screened using the program Cross Match before running eprimer3.

PCR amplification and sequencing

The primer sequences and annealing temperatures used during PCR amplification are listed in Supplementary Table 1. Amplification of PCR products was performed in 30-μl reactions, including 3 μl 10× PFU buffer, 1.8 μl 25 mM $MgCl_2$, 3 μl 2 mN dNTPs, 1 μl Taq polymerase with 5% PFU DNA polymerase (an enzyme found in the hyperthermophilic archaeon *Pyrococcus furiosus*,), 1 μl DNA, 1.5 μl primer and 17.2 μl H_2O, using the following procedure: initial denaturation at 94 °C for 4 min, then 32 cycles of denaturation at 94 °C for 1 min, annealing at 52 °C for 1 min and extension at 72 °C for 1 min. In the final cycle, extension at 72 °C was prolonged for 7 min. In an effort to increase the throughput and quality of PCR, 'Touchdown PCR' was used for some loci (Supplementary Table 1), decreasing the annealing temperature 1° every second cycle until a minimum temperature is reached (Rong *et al*., 2011). Amplicon quality was determined by running a 5 μl aliquot of each product on a 1% agarose gel. Amplicons were only deemed suitable for sequencing if single high quality bands were clearly observed in the agarose electrophoresis gels.

Before sequencing, PCR products were cleaned up by adding 4 μl 1% Exonuclease I (Exo) and 9.1% Shrimp Alkaline Phosphatase to each well of the plate that has PCR product. Plate is put on thermocycler and the 'ExoSap' program is run for 30 min (37 and 80 °C for 15 min each). Sequencing reactions used the ABI Big Dye 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA). Unincorporated dye terminators were removed from the reactions by first adding a mild detergent (12 μl 0.003% SDS solution) to each well, denaturing, and then filtering using Sephadex G-50 filter plates. The cleaned products were filtered directly into Perkin-Elmer MicroAmp Optical 96-well reaction plates, then sequenced on an ABI Prism 3730xI (Applied Biosystems, Foster City, CA, USA).

Sequence alignment

A bioinformatics pipeline was developed in which sequences produced from the six cotton relatives by each primer pair were treated as a single project. Using computational tools developed in Python, high-quality traces (meeting a minimum threshold of 50 base pairs with a Phred score greater than 20) were distributed into folders specific to the primer pair from which they were derived. The software programs Phred and Phrap were then run in such a way that consensus sequences for each genotype were first produced via the alignment of forward and reverse reads and quality files for these alignments were generated. This quality file reflects the combined scores of the reads by increasing the consensus *Q* score at base positions that overlap and match, and likewise decreasing *Q* score at mismatches. To further reduce the number of false-positive polymorphism calls, base pairs were trimmed from either end of a sequence that was below a quality score of 16. The six consensus sequences for each diploid locus were then aligned using T-Coffee and/or Clustalw (Thompson *et al*., 1994; Notredame *et al*., 2000).

Polymorphism analysis

We performed both pairwise and three-way comparisons of polymorphism and evolutionary rates between different diploid cotton relatives. Pairwise comparisons were done between all combinations of (sub) species, revealing: single-nucleotide polymorphism (SNP), multiple extended SNP, single-nucleotide indel, extended (multiple nucleotide) indel, and mixed indels and nucleotide substitutions. Three parameters were mainly analyzed to indicate divergence among different genomes, including SNPs, polymorphic sites (nucleotides with divergence due to all polymorphism types listed above, and abbreviated as 'poly site' in the figures and tables) and polymorphic base pairs, abbreviated as 'poly bp' in a 100-bp comparison window. All bases with sequence quality scores $\geqslant 25$ were counted, and the quality control threshold was reduced when two matched bases between compared sequences were identical with the sum of the two scores $\geqslant 30$.

Three-way comparison was performed to reveal nucleotide variations in sequences from two (sub)genomes with reference to their closest outgroup genome, to try to provide more resolution about substitution types and ancestral states. A three-node tree ((X1, X2), O)) defined a specific three-way comparison between (sub)genomes X1 and X2 with reference to their outgroup O. The following three-way combinations were considered in our analysis: ((A1, A2), C1), ((D5, D8), C1), ((A1, C1), D5), ((A1, D5), Ki), and ((C1, D5), Ki). At a specific site, the matched trios (bases and/or indels) were subjected to quality control: if a trio was formed by three identical bases, the data were only considered if the summed quality score was $\geqslant 35$; if two bases and one indel, the summed quality score of the two bases $\geqslant 30$; if only one base and two indels, the quality score of the base $\geqslant 25$. For three-way comparison of ((A, D), Ki), when the two A and two D subgenomes were each monomorphic but differ between A and D, then ancestral sequences were inferred by referring to the Ki sequence.

Reconstruction of phylogeny

We pooled all the alignments of genes together to produce a concatenated alignment, on which the genome and species phylogeny was reconstructed. The concatenated alignment was constructed by selecting aligned columns without gaps and with matched bases having quality scores $\geqslant 25$. The phylogeny was reconstructed using several approaches, including neighbor-joining, maximum parsimony, minimum evolution and UPGMA, each implemented in MEGA 4 (Tamura *et al*., 2007). Default parameter settings for each approach were adopted. To compare potential phylogenies by rearranging the locations of diploid A, D and C genome species to produce different trees [T1: ((((A1, A2), C1), (D5, D8)), Ki); T2: ((((A1, A2), (D5, D8), C1)), Ki); T3: ((((A1, A2), (D5, D8)), C1), Ki)], we performed maximum likelihood evaluation by running DNAML using default settings implemented in PHYLIP (http://evolution.genetics.washington.edu/). The difference of logarithms of maximum likelihoods for two different phylogenies follows a normal distribution as previously described (Kishino and Hasegawa, 1989).

RESULTS

General divergence of DNA sequences among diploid Gossypium genomes

A total of 780 loci have been amplified from 6 diploid cotton relatives and their amplicons have been sequenced and aligned (Supplementary Figure 1 is an example of such alignment), resulting in 15 possible pairwise comparisons for each locus. Because of differences in success of PCR amplification and DNA sequencing among species, the number of informative loci vary among different pairwise comparisons, from 372 (D8/Ki) to 552 (A1/A2) (Table 1 and Supplementary Table 2). Average comparable sequence length is similar among pairwise comparisons, ranging from 209 (C1/D5) to 229 bp

Table 1 DNA divergence in the pairwise comparisons among six cotton relatives

Comp.[a]	*Seq pair*	*Total bp*	*Seq length (bp)*	*SNP/100 bp*	*Poly site/100 bp*	*Poly bp/100 bp*
A1:A2	552	125 545	227.4	0.34	0.72	1.44
A1:C1	512	112 407	219.5	1.33	1.95	3.80
A1:D5	518	113 572	219.3	1.54	2.29	4.50
A1:D8	512	112 244	219.2	1.60	2.35	4.29
A1:Ki	366	78 974	215.8	2.26	3.48	7.32
A2:C1	536	114 736	214.1	1.32	1.97	3.84
A2:D5	537	114 193	212.6	1.51	2.25	4.40
A2:D8	531	113 096	213.0	1.63	2.40	4.40
A2:Ki	371	78 971	212.9	2.23	3.41	7.25
C1:D5	526	109 982	209.1	1.27	1.96	3.80
C1:D8	529	111 081	210.0	1.29	2.02	4.11
C1:Ki	374	80 148	214.3	2.05	3.11	6.51
D5:D8	544	116 422	214.0	0.68	1.21	2.86
D5:Ki	376	80 777	214.8	2.19	3.35	6.93
D8:Ki	372	81 879	220.1	2.18	3.30	6.63
A:D	423	81 904	193.6	1.50	1.85	3.08

Abbreviations: A, A genome ancestor; Comp., comparison; D, D genome ancestor.
[a]Pairwise comparisons of six diploid cotton relatives, one pair of sequence per locus in comparison. A1: *G. herbaceum* (A1-97); A2: *G. arboretum* (A2-47); D5: *G. raimondii*, (D5-1); D8: *G. trilobum* (D8-1); C1: *G. sturtianum* (C1-4) and Ki: *Gossypioides kirkii*.

Table 2 DNA divergence of different locus type in pairwise comparisons among six cotton relatives

Locus type	*Seq. pair*	*Total bp*	*Seq. length*	*SNP/100 bp (average; range)*	*Poly site/100 bp (average; range)*	*Poly bp/100 bp (average; range)*
EST	4200	996 182	237.2	1.32; 0.29 (A1:A2)–2.06 (A1:K)	2.06; 0.67 (A1:A2)–3.30 (A1:K)	3.96; 1.38 (A1:A2)–6.73 (D5:K)
gDNA	2155	421 957	195.8	1.75; 0.34 (A1:A2)–2.71 (D8:K)	2.54; 0.65 (A1:A2)–3.85 (D8:K)	4.99; 1.12 (A1:A2)–8.32 (A1:K)
CFNC	569	90 104	158.4	1.86; 0.82 (A1:A2)–3.44 (D8:K)	2.66; 1.20 (A1:A2)–5.08 (D8:K)	5.81; 1.66 (A1:A2)–11.68 (D5:K)
SSR	232	35 784	154.2	2.01; 0.41 (A1:A2)–3.66 (C1:K)	3.77; 1.61 (A1:A2)–5.47 (C1:K)	12.71; 6.17 (A1:A2)–25.3 (A2:C1)
Total				1.48	2.27	4.55

Abbreviations: CFNC, Cot-filtered non-coding; EST, expressed sequence-tag; gDNA, genomic DNA; Seq., sequence; SNP, single-nucleotide polymorphism; SSR, simple sequence repeat.

(A1/A2). SSR-derived loci have the shortest comparable sequence lengths (154.2, 136.8–250.5) followed by CFNC loci (158.4, 131.0–178.9) and gDNA (195.8, 184.5–204.8), with EST loci having the longest (237.2, 227.7–254.4) (Table 2).

Patterns of DNA divergence among the six cotton relatives (Table 1 and Supplementary Table 2) agree generally with expected phylogenetic relationships (Hawkins *et al*., 2006). The two D genome species (D5 vs D8) exhibited nearly twice the DNA divergence of the two A genome species (A1 vs A2), with 0.68 vs 0.34 SNPs, 1.21 vs 0.72 polymorphic sites and 2.86 vs 1.44 polymorphic bp per 100 comparable bp, respectively (Table 1 and Supplementary Table 2). Among different genome types, the outgroup species (*G. kirkii*) was more distant from all other genotypes than each was from one another (6.51–7.32 polymorphic bp/100 bp). Importantly the divergence between A and D species is about 4.4 polymorphic bp/100 bp, whereas that between A and C species is about 3.8 (Table 1 and Supplementary Table 2).

The types and levels of divergence among loci varied widely. The pairwise comparison between G. *herbaceum* (A1) and *G. trilobum* (D8) is used as an example to illustrate this point. A total of 512 loci were successfully amplified and sequenced resulting in 112 244 comparable bp from A1 and D8, with average comparable sequence length per locus of 219.2 (Table 1). Among the 512 comparable loci, 35 comprising 4693 bp (134 bp/locus) were identical. The remaining 477 revealed a total of 4814 polymorphic base pairs, derived from 2641 polymorphic sites including 1798 SNPs, 168 single base indels and 675 extended polymorphisms (Figure 1). So, the most abundant

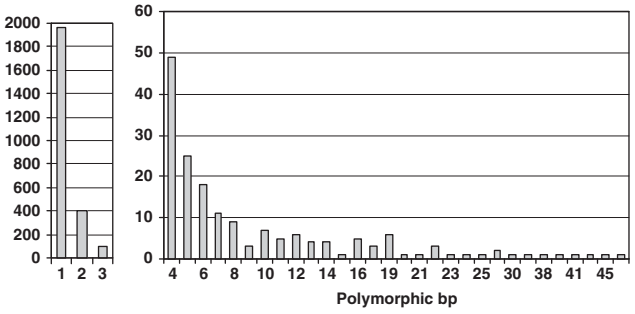

Figure 1 Length (bp) of polymorphic sites in pairwise comparison between A1 and D8 cotton species.

mutations are single-nucleotide substitutions (Figure 1). However, SNP frequencies varied considerably among different loci, ranging from 0.21 to 10.7 per 100 bp, with about 50% in a range of 0.5–1.5 SNPs/100 bp (Figure 2). Extended polymorphisms averaged 4.2 bp in length with the longest being a 65 bp deletion in *G. trilobum* at locus A1270. The number of polymorphic sites per 100 bp ranges from 0.01 to 14.99 among different loci, with most loci having 0.5–2.99 polymorphic sites per 100 bp (Figure 3).

Comparison of DNA element classes

Primer pairs derived from different types of sequence-tagged sites (EST, gDNA, CFNC DNA and SSR) showed different polymorphism

levels. ESTs have the smallest DNA divergence followed by gDNA and CFNC (Table 2). SSR-derived loci showed the largest DNA divergence in most pairwise comparisons with a few exceptions (Supplementary Table 2, Table 2). For example, in the comparison of A1 vs A2, CFNC showed greater SNP number than SSR (Supplementary Table 2). SSR-derived loci had almost twice as much divergence as ESTs in most pairwise comparisons and most mutation types (Supplementary Table 2). In particular, SSR loci have more polymorphic sites larger than two bp, resulting in a larger total number of polymorphic bp than other types of loci (Figure 3). For example, SSR loci have 3.66 SNP in the most divergent pairwise comparison (C1:K), only ~1.5 times more than ESTs. However, SSR loci have 6.17 polymorphic bp in the least divergent pairwise comparison (A1 vs A2), ~4.5 times more than EST (1.38, Table 2, Supplementary Table 2). Similarly, in the most divergent pairwise comparison, SSR loci have almost 4 times more polymorphic bp than ESTs (25.3 vs 6.73, Table 2).

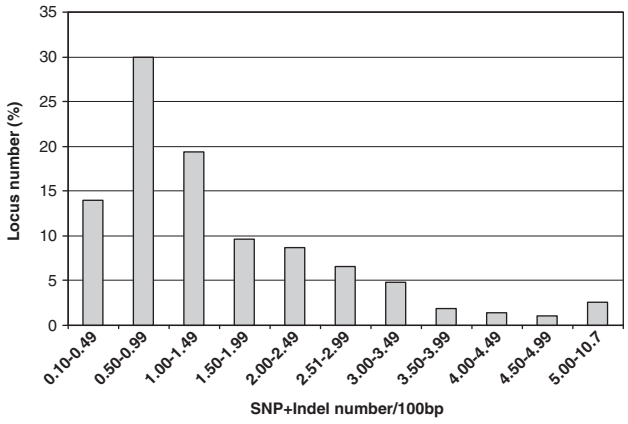

Figure 2 Frequencies of loci with different numbers of single bp mutations (SNP+Indel) per 100 bp comparison between A1 and D8 cotton species.

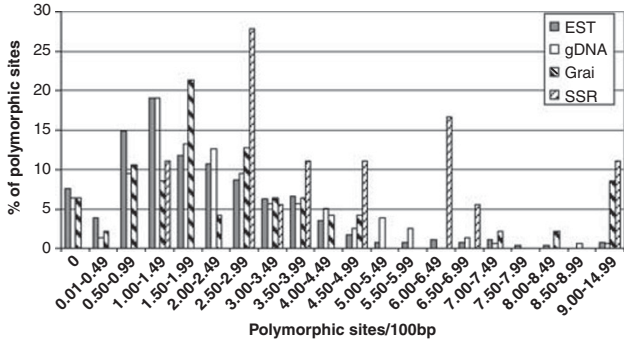

Figure 3 Comparison of polymorphic site abundance between A1 and D8 cotton species in four DNA element classes.

Comparison of A and D genomes

The two A and two D genome species studied were chosen for several reasons. First, in each case the two species provide a diverse sampling of variation within the respective genome types. Loci that are polymorphic between the A and D genome types but monomorphic within each type may be especially attractive for SNP discovery in tetraploid cottons, being locus-specific. For each genome type, reference maps have been made from crosses between the two species we studied (Brubaker *et al*., 1999; Rong *et al*., 2004; Desai *et al*., 2006). Accordingly, SNPs that differentiate between the two species within these genome types are potential genetic markers for ongoing studies, especially of value in the A-genome that is cultivated and improved in some parts of the world.

A total of 423 loci were amplified from the quartet of two A and two D genome species, producing 81 904 comparable base pairs across average sequence lengths of 193.6 bp (Table 3). A total of 2328 base pairs distributed across 1513 sites appear potentially informative in the distinction between A and D genomes. Among polymorphic sites, 1229 SNPs were detected between A and D diploids, which might be used to distinguish A and D genomes in tetraploid cotton (Table 1, Supplementary Table 3). In addition, 213 genome-specific indel loci were also found—the largest indel useful in the segregation of ancestral tetraploid genomes is 65 bp in length at locus A1270. Unexpectedly, SSR loci showed the lowest rate of SNPs useful in the distinction between A and D genomes, even fewer than in comparison of A1 vs A2. More in line with other comparisons, however, SSR loci have the highest percentage of polymorphisms extending over multiple nucleotides, have the second highest frequency of polymorphic sites (2.28/100 bp) and have the highest frequency of polymorphic base pairs (5.87/100 bp).

A total of 552 loci comprising 125 545 comparable bp (227.4 bp per locus on average) were available for finding polymorphism between *G. arboreum* and *G. herbaceum*. A total of 359 loci (65%) contained polymorphisms between these two species. CFNC loci showed the highest ratio of SNPs, twice the second one (SSR loci) at 0.82 vs 0.41(Supplementary Table 2). ESTs had the lowest SNP frequency.

A total of 544 loci comprising 116 422 comparable bp (214 per locus on average) were available for finding polymorphism between *G. raimondii* and *G. trilobum*, and 430 (79%) loci contained polymorphisms. As noted above, polymorphism between the two D genome species is much higher than between the two A genome species for all types of loci, being two times more on average (0.68 vs 0.34, Table 1, Supplementary Table 2). In the comparison of D genome species, SSR, rather than CFNC in the A genome, showed the highest frequency of SNPs (Supplementary Table 2) among the four classes of loci.

Comparative sequence divergence rates among diploid *Gossypium* genomes

To explore relative differences in frequencies of different mutational types among genomes and species, three way comparisons were analyzed using an appropriate outgroup (Figure 4 and Supplementary

Table 3 Percentage of transitions and transversions in six diploid cotton relatives detected in three way comparisons

	A[a]	*D*[a]	*A1*	*A2*	*A1*	*C1*	*A1*	*D5*	*D5*	*C1*	*D5*	*D8*
Transversion	46.1	37.2	43.8	42.5	46.2	44.0	45.5	41.2	41.6	41.5	41.9	38.6
Transition	53.9	62.9	56.2	57.5	53.8	56.0	54.5	58.8	58.4	58.5	58.1	61.4
Transi-transv	7.8	25.7	12.4	15.0	7.7	11.9	9.0	17.6	16.9	17.1	16.2	22.9
SNP no.	420	309	193	178	810	563	580	534	539	439	339	345

[a]Inferred A or D ancestral genome. For species abbreviations, refer to Table 1.

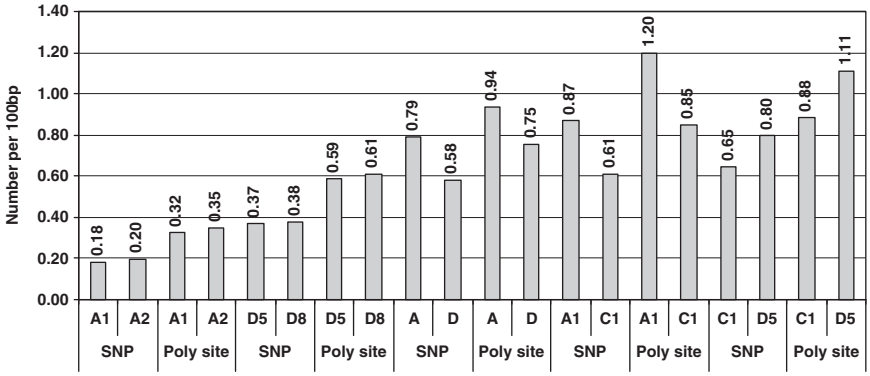

Figure 4 Types and levels of DNA polymorphism in species and genome types inferred in three way comparisons using outgroup species.

Table 4). In the comparisons between two species with the same genome (A or D), *G. sturtianum* (C genome) was used as an outgroup. No obvious differences were observed in frequencies of SNPs (0.18 vs 0.20 SNPs/100 bp of A1 vs A2 and 0.37 vs 0.38 of D5 vs D8) or polymorphic sites (0.32 vs 0.35 bp/100 bp of A1 vs A2; 0.59 vs 0.61 of D5 vs D8) between two species with the same genome type, suggesting that the respective species evolved at similar rates after speciation (Figure 4, Supplementary Table 4).

The C genome of *G. sturtianum* appears to have evolved more slowly than either A or D genomes (Figure 4 and Supplementary Table 4). Among 452 loci including 92 842 comparable bp, 810 SNPs (0.87 SNP/100 bp) were detected in A1 vs 563 (0.61SNP/100 bp) in C1 when *G. raimondii* was used as an outgroup, a highly significant difference (P=2.6E−11). Similarly, among 331 loci (67 649 bp), 539 SNPs (0.80 SNPs/100 bp) were found in D5 vs 439 (0.65 SNP/100 bp) in C1 when *G. kirkii* was used as an outgroup, also a highly significant difference. Across all types and measures of mutation, the C genome always showed less than A or D (Supplementary Table 4).

We analyzed the mutation rate of A and D ancestral genomes using *G. kirkii* as an outgroup. If A1 and A2, or D5 and D8, are monomorphic but there is polymorphism between A and D, comparison to *G. kirkii* permits us to infer ancestral state at the locus. On the basis of such inference, the A genome experienced faster evolution than the D genome, with significantly more SNPs and polymorphic sites (Figure 4, Supplementary Table 4). Across the various types of mutations, the A genome seemed to have more insertion and less deletion than the D genome (Supplementary Table 4). Interestingly, CFNC loci were incongruent with other types of loci, with more SNPs (16 vs 5) and polymorphic sites (22 vs 5) inferred to have evolved in the D than the A genome. In addition, all extended deletions (1), extended mixtures (3), single deletions (1) and insertions (1) differentiating the two genome types occurred in D, not in A (Supplementary Table 4).

Transition and transversion mutations

The relative frequencies of the 12 possible single base pair mutations were analyzed using three-way comparisons. Consistent with conventional dogma, transition mutations occurred in higher frequency than transversions in all genomes (Table 3). However, different genomes exhibited considerable difference in the ratios of transition to transversion. For example, the A ancestor genome had only 7.8% more transitions than transversions, whereas the D ancestor genome had 25.7% more (Table 3). Across all pairwise comparisons among A1, A2, D5, D8 and C1, the total SNP number is negatively correlated with the difference between transversion and transition rates (r=−0.445, P=0.074) (Table 3). There is pronounced asymmetry in mutational events, with the D genome showing about 30% more A to G and T to C mutations than the A genome, but with rates of G to A and C to T mutation being similar in the two genomes (Supplementary Figure 2). Among the eight possible transversion mutations, G to T, C to A, and C/T to G/A occurred in obviously higher frequencies in A than in D, with the others at similar frequencies in the two genomes.

Phylogenetic relationship among diploid Gossypium genomes

To reconstruct the phylogeny of the cotton genomes and species studied, we made a concatenated alignment by selecting aligned columns without gaps and meeting the quality threshold. The concatenated alignment contained 34 223 sites from 255 genes. All approaches used clearly and strongly supported the same tree topology T1: ((((A1, A2), C1), (D5, D8)), Ki), grouping C and A species together (Figure 5). This topology is further supported by maximum likelihood analysis, significantly (P-value=0.003) rejecting alternative tree topologies.

DISCUSSION

Levels and patterns of divergence of DNA sequences

Resequencing of hundreds of orthologous loci that broadly sample the genomes of diploid cottons sheds light on the types and rates of evolutionary mechanisms operating in the ~7 million years (Small *et al.*, 1998; Cronn *et al.*, 2002) since divergence of these species from common ancestors. Different DNA sequence fragments ranged widely in levels of polymorphism. For example in the comparison between A1 and D8 the ratio of polymorphic site varied from zero to 14.7 per 100 bp (Figure 4) among different loci. In all 15 pairwise comparisons, more than half of mutant sites are single-nucleotide polymorphisms, except for A1: A2 which showed nearly 50% SNPs (Table 1, Figure 2). However, polymorphisms spanning multiple nucleotides (ranging up to 274 bp insertion in C genome in the SSR locus of BNL1046 compared with A and D genome) account for 63.9% of total nucleotide divergence.

When DNA 'islands' comprising different classes of loci were analyzed separately, SSR regions showed more divergent sites and larger polymorphic size than other classes, and ESTs the least, as expected. A new type of DNA fragment extracted using Cot filtration (named CFNC), had higher divergence than other classes, in a few comparisons even exhibiting higher polymorphism than SSRs. The merits of DNA marker discovery using CFNCs has been described in detail elsewhere (Rong *et al.*, 2011). Although higher polymorphism in non coding regions (SSR, CFNC and PstI) is presumed to be related to purifying selection acting on coding regions (EST), even gene

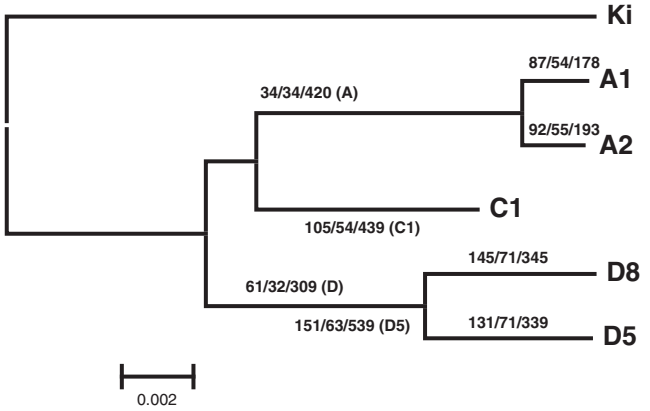

Figure 5 Phylogenetic relationship among diploid *Gossypium* genomes inferred from nucleotide alignments of 34 223 sites from 255 genes. The scale represents nucleotide substitutions per position (Maximum likelihood) estimated by PAML. The numbers above or below the branches are total sites of extended polymorphism/single indel/SNP (last) in the corresponding three way comparisons of all involved sequences, reflecting DNA sequence divergence occurring along the indicated branch of the tree.

sequences showed variable DNA divergence. For example, between two diploid A genome species, Gate3AE04, from a fiber related EST library (Arpat *et al*., 2004) has 1.8 SNP/100 bp and 2.6 polymorphic sites/100 bp including three extended indels with one spanning 58 base pairs. However, many sequences from the same library do not have any polymorphisms between A1 and A2 species. Senchina *et al*. (2003) proposed that silent site rate variation among genes may be related to genomic location, local levels of recombination and mutation, variable codon usage biases, and perhaps poorly understood differences in chromatin structure that alter in unknown ways the rate of fixation of mutations.

Reassessing the relationships among diploid cottons

The phylogeny of diploid cottons, especially the relationship between A and D genomes that give rise to tetraploid cottons, has been somewhat unclear. Prior studies used only small numbers of nucleotides either from chloroplast (Wendel and Albert, 1992; Small *et al*., 1998) or nuclear DNA (Seelanan *et al*., 1997; Small *et al*., 1999; Cronn *et al*., 2002; Wendel and Cronn, 2003). Early studies using cpDNA (Wendel and Albert, 1992; Small *et al*., 1998) suggested the C genome to have diverged first from the common ancestor shared by both A and D genome species. However, later research using genomic DNA suggested D genome species to have diverged first (Cronn *et al*., 2002; Wendel and Cronn, 2003). The latter conclusion is supported by our results, with the 780 loci used here being by far the largest data set we are aware of to date, suitable to address this question. Indeed, our data clearly reject all alternative phylogenies.

Different *Gossypium* genome types appear to be experiencing appreciably different evolutionary rates, with the A genome evolving more rapidly than the C genome, and the D genome in between (Supplementary Table 3). These differences appear to characterize genome types—species within the A and D genome types appear to be experiencing mutation at similar rates. We can identify no particular correlate with this variation—neither phylogenetic placement nor genome size appears related to it. An attractive opportunity for further study is to investigate additional genome types—for example are those closely related to A (that is B, E, F) also experiencing relatively rapid change, or are those close to C (K) also evolving slowly? It is interesting to note that the evolution of spinnable fibers, the primary economic product of cotton and also a potential means of long-range dispersal, occurred in the most rapidly evolving (A) of the genomes studied to date.

Striking differences among genome types in the frequencies of specific base pair changes were found. We found an excess of transition over transversion substitutions, as is common in metazoans (Keller *et al*., 2007). However, the A ancestor genome has only 7.8% more transitions than transversions, whereas the D ancestor genome has 25.7% more, due largely to abundance of A to G and T to C mutations. This could contribute to the higher rate of C to T (16.6%) than T to C (12.1%) mutations in the A ancestor genome. As in metazoans, part of this bias could be related to greater DNA methylation in the retrotransposon-rich A genome. Methylated cytosines (mC) are subject to UV light-induced mC→T transition mutations (Bird, 1980), at 10–50 fold greater frequency than other substitution mutations (Zhao and Jiang, 2007). The CFNC elements showed an unique pattern of base pair mutation, in the D ancestor showing significantly more SNPs (16 vs 5) and polymorphic sites (22 vs 5) than in the A ancestor. This contrasts with SSR, gDNA and EST elements, which showed significantly more mutations in the A than the D ancestor. However, only a sample of CFNC fragments (5) could be used in this comparison. In the comparisons between A1 and A2 where a larger number of CFNC loci (37) could be analyzed, A2 showed significantly more SNP and polymorphic sites than A1 (21 vs 6 and 29 vs 14), but no difference was detected in the other three types of elements. In addition, no such bias was detected for CFNC fragments between D5 and D8.

SNP markers for genomics research and plant breeding

SNPs are the most abundant source of genetic variation and the most elemental difference between genotypes, preferable to SSRs for many applications (Wang *et al*., 1998; Altshuler *et al*., 2000) and steadily becoming more economical to the genotype. Large scale projects in such model organisms as human, *Oryza sativa* and *Arabidopsis thaliana* are driving the development of low-cost, high-efficiency genotyping (Langridge and Fleury, 2011). The diploid A and D genomes are progenitors of tetraploid cotton At and Dt subgenomes. SNPs between two diploid A species, or two diploid D species are useful in gene discovery and mapping within the respective diploid genomes. Indeed, the two A genome species and two D genome species studied in this research were each used to construct genetic maps (Brubaker *et al*., 1999; Rong *et al*., 2004; Desai *et al*., 2006).

Loci that are monomorphic within A or D genome types, but polymorphic between genome types, may be of practical importance for identifying locus-specific markers in tetraploid cottons including leading cultivars. Although some loci that are monomorphic within genome types may be evolutionarily constrained at the diploid level, the presence of a homologous copy may render them less constrained at the tetraploid level. Although SNP frequencies within A or D genomes are the lowest of the 15 possible pair wise comparisons in this study, we nonetheless identified a total of 1229 SNPs from 423 sequence-tagged sites that differentiated between the A and D genome types. At these loci, genome-specific primers or oligonucleotides might be designed, simplifying SNP discovery in tetraploid cottons by mitigating the complication of having a second homologous copy of the locus present elsewhere in the genome. Specific amplification of the At or Dt genome loci from different tetraploid genotypes may facilitate the discovery of allelic sequence diversity which can be used in genetic and evolutionary studies.

DATA ARCHIVING

Data have been deposited at Dryad: doi:10.5061/dryad.fb5hk394.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank many members of the Paterson lab for their assistance, and Bayer Crop Science and the Consortium for Plant Biotechnology Research for their financial support.

Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L *et al.* (2000). An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513–516.

Arpat AB, Waugh M, Sullivan JP, Gonzales M, Frisch D, Main D *et al.* (2004). Functional genomics of cell elongation in developing cotton fibers. *Plant Mol Biol* **54**: 911–929.

Beasley JO (1940). The production of polyploids in *Gossypium*. *J Hered* **31**: 39–48.

Bird AP (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* **8**: 1499–1504.

Brubaker CL, Paterson AH, Wendel JF (1999). Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. *Genome* **42**: 184–203.

Cronn RC, Small RL, Haselkorn T, Wendel JF (2002). Rapid diversification of the cotton genus (Gossypium: Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. *Am J Bot* **89**: 707–725.

Cronn RC, Zhao XP, Paterson AH, Wendel JF (1996). Polymorphism and concerted evolution in a tandemly repeated gene family: 5S ribosomal DNA in diploid and allopolyploid cottons. *J Mol Evol* **42**: 685–705.

Desai A, Chee PW, Rong JK, May OL, Paterson AH (2006). Chromosome structural changes in diploid and tetraploid A genomes of Gossypium. *Genome* **49**: 336–345.

Edwards GA, Mirza MA (1979). Genomes of the Australian wild species of cotton. II. The designation of a new G genome for *Gossypium bickii*. *Can J Genet Cytol* **21**: 367–372.

Endrizzi J, Turcotte EL, Kohel RJ (1985). Genetics, cytol cytology, and evolution of Gossypium. *Adv Genet* **23**: 272–375.

Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2007). Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (Gossypium). *Plant J* **50**: 995–1006.

Hanson RE, Zhao XP, Islam-Faridi MN, Paterson AH, Zwick MS, Crane CF *et al.* (1998). Evolution of interspersed repetitive elements in Gossypium (Malvaceae). *Am J Bot* **85**: 1364–1368.

Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. *Genome Res* **16**: 1252–1261.

Keller I, Bensasson D, Nichols RA (2007). Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genet* **3**: e22.

Kishino H, Hasegawa M (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol* **29**: 170–179.

Langridge P, Fleury D (2011). Making the most of 'omics' for crop breeding. *Trends Biotechnol* **9**: 33–40.

Notredame C, Higgins DG, J H (2000). T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Evol* **302**: 205–217.

Phillips LL, Strickland MA (1966). The cytology of a hybrid between *Gossypium hirsutum* and *G. longicalyx*. *Can J Genet Cytol* **8**: 91–95.

Rong JK, Abbey C, Bowers JE, Brubaker CL, Chang C, Chee PW *et al.* (2004). A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). *Genetics* **166**: 389–417.

Rong JK, Robertson JS, Schulze SR, Paterson AH (2011). Cot-based sampling of genomes for polymorphic low-copy DNA. *Mol Breeding* (in press).

Rozen S, Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds). *Bioinformatics Methods and Protocols: Methods in Molecular Biology*. Humana Press: Totowa, NJ, pp 365–386.

Seelanan T, Schnabel A, Wendel JF (1997). Congruence and consensus in the cotton tribe (Malvaceae). *Syst Bot* **22**: 259–290.

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong JK, Noyes RD *et al.* (2003). Rate variation among nuclear genes and the age of polyploidy in Gossypium. *Mol Biol Evol* **20**: 633–643.

Small RL, Ryburn JA, Cronn RC, Seelanan T, Wendel JF (1998). The tortoise and the hare: choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction in a recently diverged plant group. *Am J Bot* **85**: 1301–1315.

Small RL, Ryburn JA, Wendel JF (1999). Low levels of nucleotide diversity at homoeologous Adh loci in allotetraploid cotton (Gossypium L.). *Mol Biol Evol* **16**: 491–501.

Tamura K, Dudley J, Nei M, Kumar S (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596–1599.

Thompson JD, Higgins DG, Gibson TJ (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.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R *et al.* (1998). Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome. *Science* **280**: 1077–1082.

Wendel JF, Albert VA (1992). Phylogenetics of the cotton genus (Gossypium)—character-state weighted parsimony analysis of chloroplast-DNA restriction site data and its systematic and biogeographic implications. *Syst Bot* **17**: 115–143.

Wendel JF, Cronn RC (2003). Polyploidy and the evolutionary history of cotton. *Adv Agron* **78**: 139–186.

Wendel JF, Olson PD, Stewart JM (1989). Genetic diversity, introgression, and independent domestication of old-world cultivated cottons. *Am J Bot* **76**: 1795–1806.

Wendel JF, Schnabel A, Seelanan T (1995). Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (Gossypium). *Proc Natl Acad Sci USA* **92**: 280–284.

Zhao XP, Si Y, Hanson RE, Crane CF, Price HJ, Stelly DM *et al.* (1998). Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. *Genome Res* **8**: 479–492.

Zhao XP, Wing RA, Paterson AH (1995). Cloning and characterization of the majority of repetitive DNA in cotton (Gossypium L). *Genome* **38**: 1177–1188.

Zhao ZM, Jiang CZ (2007). Methylation-dependent transition rates are dependent on local sequence lengths and genomic regions. *Mol Biol Evol* **24**: 23–25.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)