Sequencing the cotton genomes

Discipline: Molecular Biology

Andrew H. Paterson

Plant Genome Mapping Laboratory

University of Georgia,

Athens GA 30602 (USA)

Email: paterson@uga.edu

Phone: 1-706-583-0162

FAX: 1-706-583-0160

The author appreciates financial support from the National Science Foundation Plant Genome Research Program and US Department of Agriculture Plant Genome Program for cotton genetic and physical mapping, and the commitment of the US Department of Energy Joint Genome Institute 'Community Sequencing Program' to conduct exploratory sequencing of cotton. He also thanks numerous colleagues, collaborators, and staff for valuable suggestions.

Blind Cover Page

Sequencing the cotton genomes

Abstract

The genomes of most major crops, including cotton, will be fully sequenced in the next few years. Cotton is unusual, although not unique, in that we will need to sequence not only cultivated (tetraploid) genotypes but their diploid progenitors, to understand how elite cottons have surpassed the productivity and quality of their progenitors. Some technical questions remain about the most efficient approach by which to sequence the *Gossypium* (cotton) genomes, and different members of the genus may require different strategies. The potential benefits of the post-genomic era in cotton are real and large – improved quality, productivity, and stability; reduced input needs that improve sustainability and environmental stewardship of cotton production; and value-added features tailored to human needs. However, *the greatest challenge facing the cotton community is the conversion of 'sequence' to 'knowledge,' and will require investment, creativity, investment, energy, investment, coordination, investment, patience, and investment*. We will quickly identify much cotton sequence to be repetitive "junk DNA" –this cannot be dismissed as unimportant, but will be relatively low in unique information content. We will also quickly convert some cotton sequence to information based on similarity to known sequences (from *Arabidopsis* in particular). However, few if any other seedborne epidermal plant cells reach 1-2" in length and >90% cellulose. *To understand and manipulate the unique features of cotton will require a host of enabling tools, technologies, and resources to be developed and creatively deployed; in particular targeting portions of the sequence that are substantially different from those of other organisms.*

Keywords: comparative genomics, computational biology, functional genomics

The genomes of most major crops, including cotton, will be fully sequenced in the next 10 years. Plant genome sizes vary over at least 1000-fold, from 125 Mbp for the haploid genome of *Arabidopsis thaliana* to 125 Gbp for the lily *Fritillaria assyriaca*(Bennett and Smith, 1991). The average genome size of 70 crops for which estimates are available is about 2×10^9 bp (Paterson, 2006b), with a current sequencing cost of ~$10-20 million based on the most economical (whole-genome shotgun) approach, and assuming 8x average coverage of each nucleotide. Consequently, today, the decision to sequence a crop genome remains a complex equation that balances genome size with scientific, economic and social impact; the phylogenetic distance from previously sequenced plants (i.e. the new information that is likely to be yielded); relevant information from prior studies (such as availability of genetic or physical maps); and the persuasiveness of individual (or groups of) investigators.

Angiosperm (flowering plant) genomes recently comprised ~13.5% of Genbank sequence data (Paterson, 2006b); if this fraction is maintained - and if we sustain the 60% per year average increase in sequence information that has been realized since the 1980s – then the sequencing of the 200 most important domesticated plants would take a relatively short 14 years. Emerging technologies offer the means to maintain or perhaps even accelerate the increase in sequence information (Margulies et al., 2005; Shendure et al., 2005). This will eventually reduce dramatically the complexity of the decision to sequence a genome, making it possible to sequence multiple genomes routinely in the context of everyday research programs,

Cotton is unusual, although not truly unique, in that we will need to sequence not only cultivated (tetraploid) genotypes but their diploid progenitors, to understand how tetraploid cottons have come to 'transgress' the productivity and quality of their progenitors. The genus *Gossypium*

occurs naturally throughout tropical and subtropical regions of the world. According to meiotic pairing and chromosome size, the 45 diploid species ($2n = 26$) fall into genomic groups A, B, C, D, E, F, G, or K. The A-genome clade, also including B, E, and F genome types distinguished from one another based on pairing behavior, chromosome sizes, and relative fertility in interspecific hybrids (Beasley, 1942) occur naturally in Africa and Asia, while the D-genome clade occurs in America. A third diploid clade exists in Australia, including C, G, and K genome types.

Allotetraploid cottons originated in the New World, from interspecific hybridization between an A-genome African diploid species resembling *G. herbaceum*, and a D-genome American diploid species (Skovsted, 1934; Beasley, 1940) resembling *G. raimondii* or *G. gossypioides* (Gerstel, 1958; Phillips, 1963). Based on over 50 genes sequenced in both diploid and allopolyploid cotton (Senchina et al., 2003b), A- and D-genome groups are estimated to have diverged from a common ancestor 5-10 MYA, then been reunited via polyploidization in an A-genome cytoplasm (Wendel, 1989; Small and Wendel, 1999) about 1-2 MYA (Wendel and Cronn, 2003) following trans-oceanic dispersal to the New World of an A-genome propagule closely resembling the extant species *G. herbaceum*. Following hybridization with a native D-genome diploid resembling *G. raimondii* and chromosome doubling, the nascent allopolyploid spread throughout the American tropics and subtropics, radiating into different lineages now represented by three clades and five species.

Domesticated AD-tetraploid cottons appeared in the New World by 3,500–2,300 B.C. (Hutchinson et al., 1947). Domesticated A-genome diploids existed in the Old World by 2700 BC (Chowdhury and Burth, 1971), and remain intensively bred and cultivated in India, China, and Pakistan. The D-genome cotton species do not produce spinnable fiber, but have had a

significant impact on fiber traits in the allotetraploids as evidenced by marker-assisted QTL localization (Jiang et al., 1998) and chromosome substitution line performance (Saha et al., 2006).

Sequencing of representatives from each diploid clade, and preferably each genome, will be important to molecular dissection of numerous evolutionary patterns and biological phenomena, including the genomic and morphological diversity that has permitted species within the genus to adapt to a wide range of ecosystems in warmer, arid regions of the world. Sequences from diploid species, especially certain A and D genome species, will aid AD genome sequence assembly, and could prove to be invaluable in revealing differences in gene content and expression patterns across the ploidy levels, and providing insight into polyploid genome evolution. The high degree of conservation of gene order and sequence between diploids and tetraploids suggests that the vast majority of data from diploids will extrapolate directly to tetraploids.

Some technical questions remain regarding the most cost-efficient approach by which to sequence the Gossypium (cotton) genomes, and different members of the genus may require different strategies. Given the expected continuing progress in improving sequencing throughput and reducing cost (Paterson, 2006a), a strong case can be made for complete sequencing of one or more representatives of each *Gossypium* genome type, including a tetraploid. Efficient approaches to capturing the unique information available from the genus will need to consider several constraints, as follows, that may require the use of different sequencing strategies for different taxa.

1. The diploid clades diverged approximately 5-10 mya (Cronn et al., 2002), have preserved a high degree of genomic content and arrangement. At the whole-genome level, a high degree of colinearity and synteny among the A, D, and tetraploid genomes (Reinisch et al., 1994; Brubaker et al., 1999; Rong et al., 2004; Desai et al., 2006) suggests that complete sequencing of a small number of genotypes together with reduced-representation sequencing of representatives of additional nodes might be a cost-effective interim step. There has been some additional rearrangement of tetraploid chromosome structure relative to their diploid progenitors (Brubaker et al., 1999; Rong et al., 2004; Desai et al., 2006), including some evidence of cryptic rearrangements that may not be obvious based on genetic maps (Waghmare et al., 2005). In partial summary, the high degree of transferability of information about gene content and order among the respective genome types suggests that whole-genome efforts in favorable taxa will provide strong guidance for future efforts in the most difficult taxa.

2. Efficient strategies for capturing the sequence diversity represented within the *Gossypium* genus will be greatly influenced by large differences in genome size and organization across the genus. The diploid genomes vary about 3-fold in DNA content, but have the same chromosome number and similar gene content. This variation in genome size appears to have accumulated in about 5-10 million years since the diploid clades are thought to have diverged from a common ancestor (Senchina et al., 2003a). The smallest haploid genome size is estimated to be ~880-Mb for *G. raimondii* Ulbrich, with a size of ~1.75-Gb for *G. arboreum* L., and ~2.5 Gb for tetraploid *G. hirsutum* L. (Hendrix and Stewart, 2005). DNA content of the allopolyploids is approximately the sum of those of the A and D-genome

progenitors, and nearly all of >22,000 AFLP fragments surveyed are additive in the allopolyploids (Liu et al., 2001). The variation in DNA content in the diploid species might be the net result of both increases and decreases in copy number of various repeat families.

3. Much of the size variation among the diploid genome types is due to dispersed repetitive DNA (Zhao et al., 1998a), which appears to be largely retrotransposon-like elements (Hawkins et al., 2006). There appears to have been large expansions of repetitive DNA content in the A/B/E/F and C/G/K genome clades in the 5-10 million years since the divergence of the diploid clades, thus many of these element families may include large numbers of relatively recently-derived members that are problematic for whole-genome shotgun approaches. By contrast, the D genome clade appears to have only a minimum of such recently-amplified repetitive DNA, and may be more amenable to whole-genome shotgun approaches. A survey of about 100 of the most abundant families in the tetraploid genome showed only 4 to be abundant in the D genome but rare or absent in the A genome. Thus, most high-copy repetitive DNA families in the D genome are older than the A-D divergence (5-10 million years old), an age that renders them likely to be amenable to assembly by a whole-genome shotgun approach. By contrast, the alternative A genome progenitor contains about 50 repetitive element families that are rare or absent from the D genome, suggesting that these families amplified in this same 5-10 million year period. Most of these A-genome repetitive element families contain thousands of members, and have continued to amplify and transpose since polyploid formation about 1-2 mya (Zhao et al., 1998a), rendering the A and tetraploid 'AD' genomes potentially less amenable to whole-genome shotgun approaches.

4. The tetraploid clades combine the properties of the A and D genome diploids with modification by intergenomic concerted evolution. Concerted evolution of the repetitive DNA fraction (Wendel et al., 1995b; Wendel et al., 1995a; Cronn et al., 1996; Zhao et al., 1998a) has been clearly shown. The possibility of intergenomic exchange of low-copy DNA remains somewhat unclear, with evidence for (Reinisch et al., 1994) and against it (Cronn et al., 1999), but growing data from other taxa strongly suggest that it may be an important dimension of polyploid evolution (Hughes and Hughes, 1993; Moore and Purugganan, 2003; Gao and Innan, 2004; Chapman et al., 2006; Wang et al., 2007.). The possibility of intergenomic concerted evolution, much like the presence of recently-amplified repetitive DNA families, would tend to support the need for a BAC-based rather than a whole-genome shotgun approach in the affected genome(s).

Given these four considerations, one logical conclusion is that the whole-genome shotgun sequence of the smallest *Gossypium* genome would be likely to expediently provide fundamental information about gene content and organization across the genus. *G. raimondii* has the smallest genome size in the genus (~880Mb) and lowest amount of repetitive DNA sequences, with most of its repetitive DNA relatively old and therefore likely to be comprised of well-differentiated family members. A fully sequenced *G. raimondii* genome would establish an initial 'template/backbone' toward the long term goal of characterizing the spectrum of diversity among the eight *Gossypium* genome types and three polyploid clades. Whole-genome shotgun based characterization of this smallest genome is in theory the most cost effective and easiest of the whole genome approaches at present. For these and other reasons the U.S. Department of

Energy Joint Genome Institutes has selected *G. raimondii* for a pilot study for shotgun sequencing 0.5x coverage, to better define the genome and a workable strategy for its complete sequencing.

The economic importance of cotton fibers and scientific interests in polyploidy suggest an ultimate goal of sequencing a *G. hirsutum* tetraploid. The possibility of intergenomic concerted evolution, much like the presence of recently-amplified repetitive DNA families, would tend to support the need for a BAC-based rather than a whole-genome shotgun approach. Using a finished diploid genome as a template and guide, a BAC-based sequence of a tetraploid will elucidate the types and frequencies of changes that have distinguished polyploid from diploid cottons. A reasonable approach is to establish a minimum tiling path of finger-printed BAC contigs (FPC) using genetically-anchored DNA markers and BAC-end sequences largely as described for other taxa (Chen et al., 2002; Bowers et al., 2005). While hybridization probes may anchor multiple homoeologous loci, 5-10 million years of divergence will provide for adequate differentiation of BAC contigs in most instances. Any exceptions can be resolved using routine techniques (Lin et al., 2000). Contig assemblies might be further validated, and any rogue contigs lacking genetic markers anchored to their chromosomal locations, using BAC fluorescence *in situ* hybridization (FISH) (Hanson et al., 1995; Stelly et al., 1995; Zwick et al., 1998; Kim et al., 2005b; Kim et al., 2005a; Wang et al., 2006).

As noted above, a completed sequence of the A-diploid genome is also essential. The exact method by which to pursue this awaits further characterization of the levels and patterns of diversity among members of its many large families of repetitive DNA elements.

We will quickly identify much of the sequence as repetitive "junk DNA" – however this cannot be dismissed as unimportant. Two prior analyses using DNA renaturation kinetics (Walbot and Dure, 1976; Geever et al., 1989) show the *G. raimondii* genome to have three kinetic components.

(1) About 3-7% of the DNA renatures virtually instantaneously, suggesting intramolecular associations such as palindromic sequences (Walbot and Dure, 1976; Geever et al., 1989).

(2) About 30-32% of the total genomic DNA is repetitive, with a kinetic complexity (approximately equivalent to sequence complexity) of 1.6×10^6 bp and an average iteration frequency of ~120 copies per haploid genome (Geever et al., 1989). Others (Walbot and Dure, 1976) subdivided the repetitive fraction into a highly-repetitive component of about 5% of the genome, composed of elements in 10,000 or more copies; and a middle-repetitive component accounting for 27% of the genome. A random sampling of 0.04% of the tetraploid cotton genome, enough to sample repetitive element families that occur in 2500 or more copies, revealed 4 D-genome (*G. raimondii*) derived elements ranging in copy number up to about 15,000 based on slot-blot hybridization analysis (Zhao et al., 1998b).

(3) About 60-63% of the genomic DNA is single- or low-copy (Walbot and Dure, 1976; Geever et al., 1989).

The repetitive portions of the *G. raimondii* sequence will be quickly revealed by using established techniques that will also shed light on their antiquity, and the nature of their evolution –i.e. differentiating a sudden rapid burst of amplification, from slow progressive growth (and probably also elimination) (Wicker et al., 2005).

While much of the repetitive DNA is thought to be 'junk DNA' that continues to exist because of its ability to multiply rapidly (Doolittle and Sapienza, 1980), some proximally-repeated elements serve essential functions (centromeres), or encode products needed in large quantities (rDNA). Moreover, there is growing evidence of roles of repetitive DNA in the regulation of gene expression, and even some highly-repetitive regions of a genomc contain occasional genes (Nagaki et al., 2004). Therefore, while the repetitive fraction of the genome will be a relatively low priority for functional analysis, it cannot be summarily dismissed.

We will convert some cotton sequence to information by identifying similarities to other well-studied genomes (Arabidopsis in particular). The relatively close relationship of cotton and *Arabidopsis*, detailed genetic map for cotton, and potential importance of using functional genomic information and tools from *Arabidopsis* to aid in dissecting economically-important pathways in cotton make this system an excellent case study for exploring comparisons of gene order among divergent taxonomic families. Research into the genetic control of cotton fiber development may benefit from progress in understanding the growth and development of hair-bearing epidermal cells (trichomes) in *Arabidopsis*. Indeed, *Gossypium* and *Arabidopsis* are thought to have shared common ancestry about 83-86 million years ago (Benton, 1993), and cotton may be the best crop outside of the Brassicales in which to employ 'translational genomics' from *Arabidopsis*.

A total of 2162 (92.5% of) probes detecting 2800 (92.8% of) loci in a recent cotton reference map could be sequenced. Among these, 1738 (62.1%) of the sequenced loci had one or more unambiguous homologs in the *Arabidopsis* genome based on 7437 BLAST matches that met a threshold of $E< 10^{-10.}$ At least 59% of the cotton map and 53% of the *Arabidopsis* transcriptome

show correspondence in multi-locus gene arrangements (Rong et al., 2005), suggesting that the well-annotated *Arabidopsis* genome sequence and its gene set, the best-studied of any plant, will provide considerable guidance in deducing the structure and function of those cotton genes that are relatively conserved.

To understand and manipulate the features that make cotton unique will require a host of enabling tools, technologies, and resources; in particular targeting portions of the sequence that are substantially different from those of other organisms. In that the basic gene set for angiosperms has been revealed by sequencing of several botanical models, a natural priority in sequencing cotton will be to reveal genes are related to its unique features. There are few if any other examples of seedborne epidermal plant cells that reach 1-2" or more in length and are nearly pure cellulose. How will we recognize the genes that confer these features, and how will we determine how they work?

Rapid gene evolution may be due to a lack of structural or functional constraint, or to strong positive selection for functional divergence. Established statistical approaches allow one to distinguish clearly between these possibilities (Yang, 1997; Nielsen and Yang, 1998; Yang, 1998; Yang et al., 2000a). Genes under strong positive selection are an important complement to the highly conserved, functionally important genes amenable to comparative genomics. For example, rapidly evolving genes in Drosophila, mammals, and several other species are vital to reproductive success, cell-cell recognition, and cellular response to pathogens (e.g., (Yang et al., 2000b; Swanson et al., 2001b; Swanson et al., 2001a)).

However, recognition of genes that have evolved rapidly, will not by itself reveal their functions. More generally, there is every reason to expect that many cotton genes may have

different (or at least partly different) functions than Arabidopsis genes with similar sequences. Even 'diploid' cottons are actually paleopolyploids, having incurred a large-scale (presumably whole-genome) duplication since their divergence from Arabidopsis. As a consequence of genome duplication(s) and associated gene loss, gene linkage relationships in cotton are often different than in *Arabidopsis* (Rong et al., 2005). Further, there is every reason to anticipate that the functions of some genes have been subdivided [*subfunctionalized* – (Lynch and Force, 2000)] between duplicated *Gossypium* copies, while other duplicated copies may have evolved completely new functions (neofunctionalization) that do not exist in *Arabidopsis* or other outgroups. Finally, Arabidopsis itself is not nearly so 'simple' a genome as was once thought, having been through at least one whole-genome duplication itself since its divergence from cotton – thus, both *Arabidopsis* genes AND Gossypium genes may have changed function and tissue-specificity since their divergence from common ancestral genes. Ongoing sequencing of additional angiosperm genomes may identify better models than *Arabidopsis* for deducing ancestral and derived (modern) functions of groups of related genes.

In partial summary, *to understand and manipulate the features that make cotton unique will require new enabling tools, technologies, and resources.* A few particularly-high priorities among these are likely to include (in random order):

1. Large-scale expression profiling of the full set of cotton genes (indeed, preferably the entire genome) across a comprehensive sampling of Gossypium species, tissues, organs and developmental states, using a common platform such as has been used in other taxa (Persson et al., 2005), to permit deductions about gene function based on coordinated expression patterns.

2. Large-scale sampling of patterns of between-species divergence and within-species diversity of the full set of cotton genes (indeed, preferably the entire genome), providing the means to distinguish among genes that show evolutionary patterns such as:
    a. Divergence to novel function in a particular clade (for example, the A-genome diploids), followed by purifying selection within that clade suggesting that the new function is under strong selection;
    b. Divergence to new function in a clade, with continuing positive selection within the clade such as might be expected in the ongoing 'arms war' between plants and their pests;
    c. Conservative evolution across otherwise divergent clades, suggesting that the ancestral function is broadly adaptive and under purifying selection.

3. Comprehensive mutant resources. Strategies for *Gossypium* functional genomics need to anticipate that many genes may be implicated in crop improvement by association genetics approaches that would benefit from functional validation. Comprehensive mutant populations, using established techniques (McCallum et al., 2000; Till et al., 2003; Slade et al., 2005; Comai and Henikoff, 2006) that are likely to become much faster and less costly using future-generation resequencing technologies, can provide a means by which functional analysis of *Gossypium* genes can be carefully-targeted to complement and supplement more extensive resources for *Arabidopsis* and other botanical models. This approach will provide for both the study of genes/gene families that are less tractable in other plants, and also for targeting functional analyses to specific genes implicated in key cotton traits by association genetics or other approaches. Such resources are ideally needed for each of the two cultivated tetraploids (to permit study of

duplicated gene fates during all-important adaptation to the polyploid state) and each of the diploid genome types, with priority placed on the progenitor A and D genomes that contributed to the tetraploid.

The potential benefits of the post-genomic era in cotton are real and large – improved quality, productivity, and stability; reduced input needs that improve sustainability and environmental stewardship; and value-added features tailored to human needs rather than natural adaptation. The 8 divergent genomes in the *Gossypium* (cotton) genus are well known to enjoy a broad spectrum of morphological and physiological diversity that has permitted species within the genus to adapt to a wide range of ecosystems in warmer, arid regions of the world. Virtually all of this diversity is conferred by genes that are not yet identified, and the vast majority is found in taxa that are presently beyond the reach of mainstream breeding programs. Identification of genes native to *Gossypium* that confer desirable adaptations or traits, together with their rapid and specific transfer to elite genotypes using highly-targeted transformation approaches, may provide a means to harness this variability in a manner that is minimally subject to public concerns.

Moreover, better understanding of *Gossypium* genome evolution may provide means to broaden the reach of breeding programs into the secondary and tertiary gene pools. Convergent patterns of gene loss following whole-genome duplications in angiosperms, yeast, and even vertebrates suggest the existence of taxon-independent principles of molecular evolution that contribute to the evolutionary fates of whole-genome duplications. By better understanding these principles through the analysis of additional genomes and functional dissection in favorable models, we might gain understanding of how ‘synthetic polyploids’ combining genomes and

genotypes that are not known to have occurred naturally, might be more widely used in crop improvement.

The greatest challenge facing the cotton community is the conversion of 'sequence' to 'knowledge,' a challenge that will require investment, creativity, investment, energy, investment, coordination, investment, patience, and investment. The inevitable sequencing of the cotton genome(s) will mark the culmination of one era, and the beginning of another. The sequence(s) will lay bare the secrets of the genetic potential of the *Gossypium* genus, if we are clever enough to find appropriate ways to recognize them. In the 'simple' botanical model *Arabidopsis thaliana*, publication of its sequence in 2000 (Initiative, 2000) was followed shortly by the inception of the *Arabidopsis* 2010 project by the US National Science Foundation, and similar projects in other countries, with the goal of determine the function of each of the (~30,000) *Arabidopsis* genes by the year 2010. To date, the *Arabidopsis* 2010 project alone has invested more than $168 million toward this goal (www.nsf.gov/bio/pubs/awards/2010awards.htm), with additional investments made in other countries, and by private firms. While the cotton genome will derive some benefit from Arabidopsis 2010 (detailed above), we must anticipate that the greater complexity of cotton will require a similar level of investment, in order to realize the potential benefits of its sequencing. Some investments may be in commercializable intellectual property that are appropriately made in the private sector. However, many will be in enabling tools that might most efficiently be produced in the public domain or by public-private consortia, which engage core competencies of public researchers as a 'virtual research and development network' that offers new opportunities for small and medium-sized businesses and enhances

existing opportunity for large businesses, by providing new tools, information, and young scientists with the expertise to put these resources to work.

REFERENCES

Beasley, J.O. 1940. The origin of American tetraploid *Gossypium* species. Amer Naturalist 74:285-286.

Beasley, J.O. 1942. Meiotic chromosome behavior in species hybrids, haploids, and polyploids of Gossypium. Genetics 27:25-54.

Bennett, M., and J. Smith. 1991. Nuclear DNA amounts in angiosperms. Philos Trans R Soc London B 334:309-345.

Benton, M.J. 1993. The fossil record 2. Chapman and Hall, New York.

Bowers, J.E., M.A. Arias, R. Asher, J.A. Avise, R.T. Ball, G.A. Brewer, R.W. Buss, A.H. Chen, T.M. Edwards, J.C. Estill, H.E. Exum, V.H. Goff, K.L. Herrick, C.L.J. Steele, S. Karunakaran, G.K. Lafayette, C. Lemke, B.S. Marler, S.L. Masters, J.M. McMillan, L.K. Nelson, G.A. Newsome, C.C. Nwakanma, R.N. Odeh, C.A. Phelps, E.A. Rarick, C.J. Rogers, S.P. Ryan, K.A. Slaughter, C.A. Soderlund, H.B. Tang, R.A. Wing, and A.H. Paterson. 2005. Comparative physical mapping links conservation of microsynteny to chromosome structure and recombination in grasses. Proceedings of the National Academy of Sciences of the United States of America 102:13206-13211.

Brubaker, C.L., A.H. Paterson, and J.F. Wendel. 1999. Comparative genetic mapping of allotetraploid cotton and its diploid progenitors. Genome 42:184-203.

Chapman, B.A., J.E. Bowers, F.A. Feltus, and A.H. Paterson. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicality to angiosperm genome duplication. Proc Natl Acad Sci U S A 103:2730-5.

Chen, M.S., G. Presting, W.B. Barbazuk, J.L. Goicoechea, B. Blackmon, F.C. Fang, H. Kim, D. Frisch, Y.S. Yu, S.H. Sun, S. Higingbottom, J. Phimphilai, D. Phimphilai, S. Thurmond, B. Gaudette, P. Li, J.D. Liu, J. Hatfield, D. Main, K. Farrar, C. Henderson, L. Barnett, R. Costa, B. Williams, S. Walser, M. Atkins, C. Hall, M.A. Budiman, J.P. Tomkins, M.Z. Luo, I. Bancroft, J. Salse, F. Regad, T. Mohapatra, N.K. Singh, A.K. Tyagi, C. Soderlund, R.A. Dean, and R.A. Wing. 2002. An integrated physical and genetic map of the rice genome. Plant Cell 14:537-545.

Chowdhury, K.A., and G.M. Burth. 1971. Linn. Soc. London Biol. J. 3:303-312.

Comai, L., and S. Henikoff. 2006. TILLING: practical single-nucleotide mutation discovery. Plant Journal 45:684-694.

Cronn, R.C., R.L. Small, and J.F. Wendel. 1999. Duplicated genes evolve independently after polyploid formation in cotton. Proc Natl Acad Sci U S A 96:14406-11.

Cronn, R.C., X. Zhao, A.H. Paterson, and J.F. Wendel. 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.

Cronn, R.C., R.L. Small, T. Haselkorn, and J.F. Wendel. 2002. Rapid diversification of the cotton genus (Gossypium : Malvaceae) revealed by analysis of sixteen nuclear and chloroplast genes. American Journal of Botany 89:707-725.

Desai, A., P.W. Chee, J. Rong, O.L. May, and A.H. Paterson. 2006. Chromosome structural changes in diploid and tetraploid A genomes of Gossypium. Genome 49:336-45.

Doolittle, W., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm, and genome evolution. Nature 284:601-603.

Gao, L.Z., and H. Innan. 2004. Very low gene duplication rate in the yeast genome. Science 306:1367-70.

Geever, R., F. Katterman, and J. Endrizzi. 1989. DNA hybridization analyses of *Gossypium* allotetraploid and two closely related diploid species. Theor Appl Genet 77:553-559.

Gerstel, D.U. 1958. Chromosomal translocations in interspecific hybrids of the genus *Gossypium*. Evolution 7:234-244.

Hanson, R.E., M.S. Zwick, S. Choi, M.N. Islam-Faridi, T.D. McKnight, R.A. Wing, H.J. Price, and D.M. Stelly. 1995. Fluorescent in situ hybridization of a bacterial artificial chromosome. Genome 38:646-51.

Hawkins, J.S., H. Kim, J.D. Nason, R.A. Wing, and J.F. Wendel. 2006. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252-61.

Hendrix, B., and J.M. Stewart. 2005. Estimation of the nuclear DNA content of gossypium species. Ann Bot (Lond) 95:789-97.

Hughes, M.K., and A.L. Hughes. 1993. Evolution of duplicate genes in a tetraploid animal, *Xenopus laevis*. Mol Biol Evol 10:1360-9.

Hutchinson, J.B., A.R. Silow, and S.G. Stephens. 1947. The evolution of *Gossypium* and differentiation of the cultivated cottons Oxford University Press, London.

Initiative, T.A.G. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796-815.

Jiang, C., R.J. Wright, K.M. El-Zik, and A.H. Paterson. 1998. Polyploid formation created unique avenues for response to selection in *Gossypium*. Proc Natl Acad Sci U S A 95:4419-24.

Kim, J.S., P.E. Klein, R.R. Klein, H.J. Price, J.E. Mullet, and D.M. Stelly. 2005a. Chromosome identification and nomenclature of Sorghum bicolor. Genetics 169:1169-73.

Kim, J.S., M.N. Islam-Faridi, P.E. Klein, D.M. Stelly, H.J. Price, R.R. Klein, and J.E. Mullet. 2005b. Comprehensive molecular cytogenetic analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics 171:1963-76.

Lin, Y.-R., X. Draye, X. Qian, S. Ren, L. Zhu, and A. Paterson. 2000. Fine-scale mapping and sequence-ready contig assembly in highly-duplicated genomes, using the BAC-RF method. Nucleic Acids Research 28:e23. http://www3.oup.co.uk/nar/methods/Volume_28/Issue_07/gnd023_gml.abs.html.

Liu, B., G. Brubaker, R.C. Cronn, and J.F. Wendel. 2001. Polyploid formation in cotton is not accompanied by rapid genomic changes. Genome 44:321-330.

Lynch, M., and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459-473.

Margulies, M., M. Egholm, W.E. Altman, S. Attiva, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes,

B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P.I. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. Mcdade, M.P. Mckenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380.

McCallum, C.M., L. Comai, E.A. Greene, and S. Henikoff. 2000. Targeting induced local lesions in genomes (TILLING) for plant functional genomics. Plant Physiology 123:439-442.

Moore, R.C., and M.D. Purugganan. 2003. The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A 100:15682-7.

Nagaki, K., Z.K. Cheng, S. Ouyang, P.B. Talbert, M. Kim, K.M. Jones, S. Henikoff, C.R. Buell, and J.M. Jiang. 2004. Sequencing of a rice centromere uncovers active genes. Nature Genetics 36:138-145.

Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929-936.

Paterson, A.H. 2006a. Leafing through the genomes of our major crop plants: strategies for capturing unique information. Nat Rev Genet 7:174-84.

Paterson, A.H. 2006b. Leafing through the genomes of our major crop plants: strategies for capturing unique information. Nature Reviews Genetics 7:174-184.

Persson, S., H.R. Wei, J. Milne, G.P. Page, and C.R. Somerville. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proceedings of the National Academy of Sciences of the United States of America 102:8633-8638.

Phillips, L.L. 1963. The cytogenetics of *Gossypium* and the origin of New World cottons. Evolution 17:460-469.

Reinisch, A.J., J.M. Dong, C.L. Brubaker, D.M. Stelly, J.F. Wendel, and A.H. Paterson. 1994. A detailed RFLP map of cotton, *Gossypium hirsutum* x *Gossypium barbadense*: chromosome organization and evolution in a disomic polyploid genome. Genetics 138:829-47.

Rong, J., J.E. Bowers, S.R. Schulze, V.N. Waghmare, C.J. Rogers, G.J. Pierce, H. Zhang, J.C. Estill, and A.H. Paterson. 2005. Comparative genomics of Gossypium and Arabidopsis: Unraveling the consequences of both ancient and recent polyploidy. Genome Research 15:1198-1210.

Rong, J., C. Abbey, J.E. Bowers, C.L. Brubaker, C. Chang, P.W. Chee, T.A. Delmonte, X. Ding, J.J. Garza, B.S. Marler, C.H. Park, G.J. Pierce, K.M. Rainey, V.K. Rastogi, S.R. Schulze, N.L. Trolinder, J.F. Wendel, T.A. Wilkins, T.D. Williams-Coplin, R.A. Wing, R.J. Wright, X. Zhao, L. Zhu, and A.H. Paterson. 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.

Saha, S., J.N. Jenkins, J. Wu, J.C. McCarty, O.A. Gutierrez, R.G. Percy, R.G. Cantrell, and D.M. Stelly. 2006. Effects of chromosome-specific introgression in upland cotton on fiber and agronomic traits. Genetics 172:1927-38.

Senchina, D.S., I. Alvarez, R.C. Cronn, B. Liu, J. Rong, R.D. Noyes, A.H. Paterson, R.A. Wing, T.A. Wilkins, and J.F. Wendel. 2003a. Rate variation among nuclear genes and the age of polyploidy in Gossypium. Mol Biol Evol 20:633-43.

Senchina, D.S., I. Alvarez, R.C. Cronn, B. Liu, J.K. Rong, R.D. Noyes, A.H. Paterson, R.A. Wing, T.A. Wilkins, and J.F. Wendel. 2003b. Rate variation among nuclear genes and the age of polyploidy in Gossypium. Molecular Biology and Evolution 20:633-643.

Shendure, J., G.J. Porreca, N.B. Reppas, X.X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309:1728-1732.

Skovsted, A. 1934. Cytological studies in cotton. II. Two interspecific hybrids between Asiatic and New World cottons. J. Genet. 28:407-424.

Slade, A.J., S.I. Fuerstenberg, D. Loeffler, M.N. Steine, and D. Facciotti. 2005. A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. Nature Biotechnology 23:75-81.

Small, R.L., and J.F. Wendel. 1999. The mitochondrial genome of allotetraploid cotton (Gossypium L.). Journal of Heredity 90:251-253.

B. S. Gill and J. Raupp (ed.) 1995. Classical and Molecular Cytogenetic Analysis of Cereal Genomes.

Swanson, W.J., Z.H. Zhang, M.F. Wolfner, and C.F. Aquadro. 2001a. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proceedings of the National Academy of Sciences of the United States of America 98:2509-2514.

Swanson, W.J., A.G. Clark, H.M. Waldrip-Dail, M.F. Wolfner, and C.F. Aquadro. 2001b. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proceedings of the National Academy of Sciences of the United States of America 98:7375-7379.

Till, B.J., S.H. Reynolds, E.A. Greene, C.A. Codomo, L.C. Enns, J.E. Johnson, C. Burtner, A.R. Odden, K. Young, N.E. Taylor, J.G. Henikoff, L. Comai, and S. Henikoff. 2003. Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Research 13:524-530.

Waghmare, V.N., J. Rong, C.J. Rogers, G.J. Pierce, J.F. Wendel, and A.H. Paterson. 2005. Genetic mapping of a cross between Gossypium hirsutum (cotton) and the Hawaiian endemic, Gossypium tomentosum. Theor Appl Genet 111:665-76.

Walbot, V., and L.S. Dure. 1976. Developmental Biochemistry of Cotton Seed Embryogenesis and Germination .7. Characterization of Cotton Genome. Journal of Molecular Biology 101:503-536.

Wang, K., X. Song, Z. Han, W. Guo, J.Z. Yu, J. Sun, J. Pan, R.J. Kohel, and T. Zhang. 2006. Complete assignment of the chromosomes of Gossypium hirsutum L. by translocation and fluorescence in situ hybridization mapping. Theor Appl Genet 113:73-80.

Wang, X., H. Tang, J.E. Bowers, F.A. Feltus, and A.H. PAterson. 2007. Extensive concerted evolution of rice paralogs and the road to regaining independence. . Genetics ###:###-### (accepted).

Wendel, J.F. 1989. New World Tetraploid Cottons Contain Old-World Cytoplasm. Proceedings of the National Academy of Sciences of the United States of America 86:4132-4136.

Wendel, J.F., and R.C. Cronn. 2003. Polyploidy and the evolutionary history of cotton, p. 139-186 Advances in Agronomy, Vol 78, Vol. 78.

Wendel, J.F., A. Schnabel, and T. Seelanan. 1995a. An unusual ribosomal DNA sequence from *Gossypium gossypioides* reveals ancient, cryptic, intergenomic introgression. Mol Phylogenet Evol 4:298-313.

Wendel, J.F., A. Schnabel, and T. Seelanan. 1995b. Bidirectional interlocus concerted evolution following allopolyploid speciation in cotton (*Gossypium*). Proc Natl Acad Sci U S A 92:280-4.

Wicker, T., J.S. Robertson, S.R. Schulze, F.A. Feltus, V. Magrini, J.A. Morrison, E.R. Mardis, R.K. Wilson, D.G. Peterson, A.H. Paterson, and R. Ivarie. 2005. The repetitive landscape of the chicken genome. Genome Research 15:126-136.

Yang, Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568-573.

Yang, Z., R. Nielsen, N. Goldman, and A. Krabbe Pedersen. 2000a. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155.

Yang, Z.H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences 13:555-556.

Yang, Z.H., R. Nielsen, N. Goldman, and A.M.K. Pedersen. 2000b. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431-449.

Zhao, X.P., Y. Si, R.E. Hanson, C.F. Crane, H.J. Price, D.M. Stelly, J.F. Wendel, and A.H. Paterson. 1998a. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Res 8:479-92.

Zhao, X.P., Y. Si, R.E. Hanson, C.F. Crane, H.J. Price, D.M. Stelly, J.F. Wendel, and A.H. Paterson. 1998b. Dispersed repetitive DNA has spread to new genomes since polyploid formation in cotton. Genome Research 8:479-492.

Zwick, M.S., M.N. Islam-Faridi, D.G. Czeschin, Jr., R.A. Wing, G.E. Hart, D.M. Stelly, and H.J. Price. 1998. Physical mapping of the liguleless linkage group in Sorghum bicolor using rice RFLP-selected sorghum BACs. Genetics 148:1983-92.