Plant Genetics and Genomics: Crops and Models

Volume 3

Plant Genetics and Genomics: Crops and Models

Series Editor: Richard A. Jorgensen

Forthcoming and planned volumes

Andrew H. Paterson
Editor

Genetics and Genomics of Cotton

Springer

Editor
Andrew H. Paterson
University of Georgia
Plant Genome Mapping Laboratory
111 Riverbend Road
Athens, GA 30602
USA
paterson@uga.edu

Preface

A few members of the *Gossypium* (cotton) genus are cultivated for the production of elongated single-celled fibers valued worldwide at about $20 billion annually at the farm gate, and which sustain one of the world's largest industries (textiles) with an annual worldwide economic impact of about $500 billion. In a number of ways, cotton production and the textile industry are closely tied to petrochemical usage. Practical application of cotton genomics offers the means to improve both sustainability of crop production, and utilization of bio-based alternatives to petrochemicals, as well as consumer satisfaction with the end product.

The *Gossypium* (cotton) genus also presents unique opportunities to advance our understanding of the natural world. In particular, the evolution of cultivated cottons from their wild ancestors has involved a fascinating series of events that offer scientists the opportunity to dissect the evolution of a novel organ, the 'lint fiber', and also to better understand the roles of polyploidy in generation of biodiversity and in crop productivity and quality.

In this book, advances of the past decade will be summarized and synthesized to elucidate the current state of knowledge of the structure, function, and evolution of the *Gossypium* genome, and progress in the application of this knowledge to cotton improvement. This book will address five broad topics. First, as a backdrop, it is important to understand the naturally-occurring diversity in the genus, its organization and distribution, and its evolutionary history. Of special importance is the formation of a single polyploid from two (among 8) diploid genome types, the radiation of this polyploid, and the independent domestication and improvement of two (among nearly 50) diploid and two (among 5) polyploid species.

Current and future cotton genomics studies are guided by a long history of classical genetics research, as well as nearly two decades of molecular genetics. Natural priorities in cotton improvement, including the genetic control of variation in cotton fiber yield and quality components, disease and pest responses, and abiotic stress responses, will be addressed.

Genomic tools for cotton biology and improvement have expanded dramatically in the past 5 years – a detailed summary of these tools and their early applications is central to this book.

The transition to DNA-based genetics has led to detailed characterization of the cotton genome, solving some mysteries but revealing others in the convoluted course of *Gossypium* evolution and improvement. A few unique features of *Gossypium* biology will be highlighted, illustrating how further study of *Gossypium* promises to advance our understanding of the natural world.

A natural long-term goal in such a diverse and complex genus as *Gossypium* is a synthesis of its evolutionary history and genetic potential into a mechanistic understanding of the key steps that have led to the productivity of cultivated cottons, and the range of diversity among wild relatives. The final chapter will attempt to point the way to such a synthesis, and the central role that genomic approaches will play.

Athens, Georgia, USA   Andrew H. Paterson

Contents

Contributors

Keith L. Adams UBC Botanical Garden and Centre for Plant Research and Department of Botany, University of British Columbia, 6270 University Blvd., Vancouver, BC V6T 1Z4, Canada

Randy D. Allen Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK, USA, randy.allen@okstate.edu

Ines Alvarez Real Jardín Botánico de Madrid, CSIC, Plaza de Murillo, 2, E-28014 Madrid, Spain, ines@rjb.csic.es

Dick Auld Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX, USA, dick.auld@ttu.edu

Jean-Pierre Baudoin Gembloux Agricultural University (FUSAGx), Unité de Phytotechnie tropicale et d'Horticulture, 2 passage des Déportés, B-5030 Gembloux, Belgium, konan.n@fsagx.ac.be; nguessanolivier@yahoo.fr; baudoin.jp@fsagx.ac.be; mergeai.g@fsagx.ac.be

Efrem Bechere Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX, USA, dick.auld@ttu.edu

Curt Brubaker Bayer BioScience N.V., Technologiepark 38, B-9052 Gent, Belgium, curt.brubaker@bayercropscience.com

B. Todd Campbell Coastal Plains Soil, Water, and Plant Research Center, USDA-ARS, Washington, DC, USA, Todd.Campbell@ars.usda.gov

Kent Chapman Department of Biological Sciences, University of Northern Texas, Denton, TX, USA, chapman@unt.edu

Peng W. Chee Department of Crop and Soil Sciences, University of Georgia, Athens, GA, USA, pwchee@uga.edu

Richard Cronn Pacific Northwest Research Station, USDA Forest Service, 3200 SW Jefferson Way, Corvallis, OR 97331, USA, rcronn@fs.fed.us

Angélique D'Hont Centre International en Recherche Agronomique pour le Développement (CIRAD), UMR1096, TA 70/03, Avenue Agropolis, 34398 Montpellier Cedex 5, France, angelique.dhont@cirad.fr

Lawrence C. Davis Biochemistry Department, Kansas State University, Manhattan, KS 66506, USA, ldavis@ksu.edu

Lalitadas A. Deshapande Head, Crop Improvement, Central Institute of Cotton Research, Nagpur, Maharashtra, India

Lex Flagel Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA

Mohamed Fokar Center for Biotechnology and Genomics, Texas Tech University, Lubbock, TX, USA, Mohamed.fokar@ttu.edu

Alan R. Gingle Plant Genome Mapping Lab, University of Georgia, Athens, GA, USA, agingle@uga.edu

Allan Green CSIRO Plant Industry, Clayton South, VIC, Australia, qing.liu@csiro.au; surinder.singh@csiro.au; allan.green@csiro.au

Candace H. Haigler Department of Crop Science and Department of Plant Biology, North Carolina State University, Raleigh, NC, USA, candace_haigler@ncsu.edu

Basavaraj M. Khadi Central Institute of Cotton Research, Nagpur, Maharashtra, India, cicrngp@rediffmail.com

N'Guessan Olivier Konan Gembloux Agricultural University (FUSAGx), Unité de Phytotechnie tropicale et d'Horticulture, 2 passage des Déportés, B-5030 Gembloux, Belgium, konan.n@fsagx.ac.be; nguessanolivier@yahoo.fr; baudoin.jp@fsagx.ac.be; mergeai.g@fsagx.ac.be

Venkatesh N. Kulkarni Senior Cotton Breeder, JK Agri-Genetics Ltd, 1-10-177, 4th Floor, Varun Towers, Begumpet, Hyderabad – 500 016, India, vnkulkarni@jkseeds.net

Avishag Levi The RH Smith Institute of Plant Science and Genetics in Agriculture, The Hebrew University of Jerusalem, Rehovot 76100, Israel

Ginger G. Light Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX, USA, ginger.light@ttu.edu

Lifeng Lin Plant Biology Department, University of Georgia, Athens, GA, USA, phill@uga.edu; Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA, paterson@dogwood.botany.uga.edu

Qing Liu CSIRO Plant Industry, Clayton South, VIC, Australia, qing.liu@csiro.au, surinder.singh@csiro.au, allan.green@csiro.au

Edward L. Lubbers Department of Crop and Soil Sciences, University of Georgia, Tifton, Georgia, USA, elubbers@uga.edu

Manjula S. Maralappanavar Agricultural Research Station, Dharwad Farm, Dharwad – 580 007, Karnataka, India, manjula_arsd@yahoo.co.in

Guy Mergeai Gembloux Agricultural University (FUSAGx), Unité de Phytotechnie tropicale et d'Horticulture, 2 passage des Déportés, B-5030 Gembloux, Belgium, konan.n@fsagx.ac.be; nguessanolivier@yahoo.fr; baudoin.jp@fsagx.ac.be; mergeai.g@fsagx.ac.be.

S.S. Narayanan Consultant, JK Agri-Genetics Ltd, 1-10-177, 4th Floor, Varun Towers, Begumpet, Hyderabad – 500 016, India

Bay Nguyen Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX, USA, robert.wright@ttu.edu

Chen Niu Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX, USA, robert.wright@ttu.edu

Andrew H. Paterson Plant Biology Department, University of Georgia, Athens, GA, USA, phill@uga.edu; University of Georgia, Plant Genome Mapping Laboratory, 111 Riverbend Road, Athens, GA 30602, USA, paterson@dogwood.botany.uga.edu; paterson@dogwood.botany.uga.edu; jkrong@yahoo.com; paterson@uga.edu

Richard G. Percy USDA-ARS, Southern Plains Agricultural Research Center, College Station, Texas, USA, Richard.Percy@ars.usda.gov

Mehboob-ur-Rahman Plant Genomics & Mol Breeding Labs, National Institute for Biotechnology & Genetic Engineering (NIBGE), Faisalabad, Pakistan, mehboob_pbd@yahoo.com

Junkang Rong Plant Genome Mapping Laboratory, The University of Georgia, Athens, GA, USA, jkrong@yahoo.com

Yehoshua Saranga The RH Smith Institute of Plant Science and Genetics in Agriculture, The Hebrew University of Jerusalem, Rehovot 76100, Israel, saranga@agri.huji.ac.il

Bir Singh Department of Crop Science, North Carolina State University, Raleigh, NC, USA, bsingh4@ncsu.edu

Surinder Singh CSIRO Plant Industry, Clayton South, VIC, Australia, qing.liu@csiro.au, surinder.singh@csiro.au, allan.green@csiro.au

James McD. Stewart Department of Crop, Soil and Environmental Sciences, University of Arkansas, Fayetteville, AR 72701, USA, jstewart@uark.edu

Norma L. Trolinder Texas Tech University & Genes Plus, Lubbock, TX, USA, genes@chipshot.net

Joshua A. Udall Department of Plant and Wildlife Science, Brigham Young University, Provo, Utah, USA, jaudall@byu.edu

Guirong Wang Data Systems Group, Sacramento, CA, USA, guirong2004@yahoo.com

Jonathan F. Wendel Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, IA 50011, USA, jfw@iastate.edu

Robert J. Wright Department of Plant and Soil Sciences, Texas Tech University, Lubbock, TX, USA, robert.wright@ttu.edu; Texas AgriLife Research, Texas A&M University, College Station, TX, USA

Yusuf Zafar National Institute for Genomics & Advanced Biotechnoogy (NIGAB), NARC, Islamabad, Pakistan, y_zafar@yahoo.com

Deshui Zhang Ventria Bioscience, Fort Collins, CO, USA, dzhang@ventria.com

Part I
Natural History and Genetic Diversity

Evolution and Natural History of the Cotton Genus

Jonathan F. Wendel, Curt Brubaker, Ines Alvarez, Richard Cronn, and James McD. Stewart

Abstract We present an overview of the evolution and diversity in *Gossypium* (the cotton genus). This framework facilitates insight into fundamental aspects of plant biology, provides the necessary underpinnings for effective utilization of cotton genetic resources, and guides exploration of the genomic basis of morphological diversity in the genus. More than 50 species of *Gossypium* are distributed in arid to semi-arid regions of the tropics and subtropics. Included are four species that independently have been domesticated for their fiber, two each in Africa-Asia and the Americas. *Gossypium* species exhibit extraordinary morphological variation, ranging from trailing herbaceous perennials to ∼15 m trees with a diverse array of reproductive and vegetative characteristics. A parallel level of cytogenetic and genomic diversity has arisen during the global radiation of the genus, leading to the evolution of eight groups of diploid (n = 13) species (genome groups A through G, and K). Data implicate an origin for *Gossypium* about 5–10 million years ago and a rapid early diversification of the major genome groups. Allopolyploid cottons appear to have arisen within the last 1–2 million years, as a consequence of trans-oceanic dispersal of an A-genome taxon to the New World followed by hybridization with an indigenous D-genome diploid. Subsequent to formation, allopolyploids radiated into three modern lineages, two of which contain the commercially important species *G. hirsutum* and *G. barbadense*.

1 Introduction to *Gossypium* diversity

Because the cotton genus (*Gossypium* L.) is so important to economies around the world, it has long attracted the attention of agricultural scientists, taxonomists, and evolutionary biologists. Accordingly, and notwithstanding the

J.F. Wendel (✉)
Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames, Iowa 50011
e-mail: jfw@iastate.edu

remaining gaps in our knowledge, a great deal is understood about the origin and diversification of the genus. Especially in the last two decades, modern molecular technologies have been brought to bear on such classic questions as the origin of the polyploid species, the relationships among species and species groups, and the origins of the domesticated forms from their wild progenitors.

Perhaps the most striking aspect of this history is that it is so widespread in scope, involving ancient human cultures on several continents and a convergent or parallel plant domestication process from divergent and geographically isolated wild ancestors. This parallel domestication process involved four species, two from the Americas, *G. hirsutum* and *G. barbadense,* and two from Africa-Asia, *G. arboreum* and *G. herbaceum*. In each of these four cases, aboriginal peoples discovered thousands of years ago that the unique properties of cotton fibers made them useful for ropes, textiles and other applications. Each of the domesticated species has its own unique history of domestication, diversification, and utilization, as detailed in many papers describing various stages in the domestication process, the origin of present patterns of genetic diversity, the shape and severity of genetic bottlenecks that accompanied the development of landraces and cultivars, and the influence of recent human history on geographic patterns of cultivation (Brubaker, Bourland and Wendel 1999; Brubaker and Wendel 1994, 2001; Hutchinson 1951, 1954, 1959; Hutchinson, Silow and Stephens 1947; Percy and Wendel 1990; Wendel, Brubaker and Percival 1992; Wendel, Olson and Stewart 1989).

This rich history involved human shaping and molding of naturally occurring diversity that originated through the process of evolutionary diversification over a period of millions of years, a legacy we continue to exploit today through deliberate introgression of alien germplasm from diverse, wild gene pools. Given current threats to global ecosystems, it has never been more critical than it is today to recognize this legacy of speciation and diversification. Toward this end a synopsis of the present understanding of the evolutionary history and taxonomy of *Gossypium* is given, and in the process an entry into the relevant body of literature is provided for the serious reader.

2 Origin of the Genus *Gossypium*

The cotton genus belongs to a small taxonomic tribe, the *Gossypieae,* that includes only eight genera (Fryxell 1968; Fryxell 1979). Four of these genera are small with restricted geographic distributions (Fryxell 1968; Fryxell 1979) including *Lebronnecia* (Marquesas Islands), *Cephalohibiscus* (New Guinea, Solomon Islands), *Gossypioides* (east Africa, Madagascar), and *Kokia* (Hawaii). The tribe also includes four moderately sized genera with broader geographic ranges: *Hampea*, with 21 neotropical species; *Cienfuegosia*, a diverse genus with 25 species from the neotropics and parts of Africa; *Thespesia*, with 17 tropical species; and last but not least, *Gossypium*, the largest and most widely distributed

genus in the tribe with more than 50 species (Fryxell 1992; Stewart, Craven, Brubaker and Wendel 2008; Stewart and Ulloa unpublished).

Molecular phylogenetic analyses have clarified several aspects of the evolutionary history of the tribe (Cronn, Small, Haselkorn and Wendel 2002; Seelanan, Schnabel and Wendel 1997). Most important has been the demonstration that the group of species recognized as belonging to *Gossypium* do in fact constitute a single natural lineage, despite their global distribution and extraordinary morphological and cytogenetic diversity. A second discovery has been the identity of the closest relatives of *Gossypium*, i.e., the African-Madagascan genus *Gossypioides* and the Hawaiian endemic genus *Kokia*. These latter genera may thus be used as phylogenetic outgroups for studying evolutionary patterns and processes within *Gossypium*. A third insight concerns the temporal component to the genealogy, which is evident in sequence divergence data that serve as a proxy for time. Using this "molecular clock", Seelanan, Schnabel and Wendel (1997) suggested that *Gossypium* branched off from *Kokia* and *Gossypioides* approximately 12.5 million years ago (mya), in agreement with a later, more extensive data set based on 10 different nuclear genes (Cronn, Small, Haselkorn and Wendel 2002). Thus, *Gossypium* appears to have diverged from its closest relatives during the Miocene, perhaps 10–15 mya, subsequently spreading around the world via trans-oceanic dispersal to acquire its modern geographic range.

3 Diversification and Speciation of the Diploid Cotton Species

From its origin millions of years ago, *Gossypium* underwent speciation and diversification, achieving a nearly worldwide distribution with several primary centers of diversity in the arid or seasonally arid tropics and subtropics (Table 1). Species-rich regions include Australia, especially the Kimberley region in NW Australia, the Horn of Africa and southern Arabian Peninsula, and the western part of central and southern Mexico. Recognition of these groups of related species and their individual constituents reflects accumulated scientific understanding that emerged from basic plant exploration and taxonomic and evolutionary study. The taxonomy of the genus has been well-studied (Cronn, Small, Haselkorn and Wendel 2002; Fryxell 1979, 1992; Hutchinson, Silow and Stephens 1947; Saunders 1961; Seelanan, Schnabel and Wendel 1997; Watt 1907), with the most modern and widely followed taxonomic treatments being those of Fryxell (1979; 1992) in which species are grouped into four subgenera and eight sections (Table 1). This classification system is based primarily on morphological and geographical evidence, although most infra-generic alignments are congruent with cytogenetic and molecular data sets as well (see Chapter 11 by Mergeai, this volume).

At present, *Gossypium* includes approximately 50 species (Fryxell 1992), but remarkably, new species continue to be discovered (Fryxell 1992; Stewart, Craven, Brubaker and Wendel 2008; Ulloa, Stewart, Garcia, Godoy, Gaytán

Table 1 Diversity and geographic distribution of the major lineages of *Gossypium*. Genomic placements of species enclosed by parentheses are yet to be determined

Genome group	Number of species	Recognized species	Geographic distribution
A	2	*G. arboreum,* *G. herbaceum*	Africa, Asia
B	3 (4)	*G. anomalum* *G. triphyllum* *G. capitis-viridis* *(G. trifurcatum)*	Africa, Cape Verde Islands
C	2	*G. sturtianum* *G. robinsonii*	Australia
D	13(14)	*G. thurberi* *G. armourianum* *G. harknessii* *G. davidsonii* *G. klotschianum* *G. aridum* *G. raimondii* *G. gossypioides* *G. lobatum* *G. trilobatum* *G. laxum* *G. turneri* *G. schwendimanii* *(G. sp.nov.)*	Primarily Mexico; also Peru, Galapagos Islands, Arizona
E	5 (9)	*G. stocksii* *G. somalense* *G. areysianum* *G. incanum* *G. trifurcatum* *(G. benidirense)* *(G. bricchettii)* *(G. vollesenii)* *(G. trifurcatum)*	Arabian Penisula, Northeast Africa, Southwest Asia
F	1	*G. longicalyx*	East Africa
G	3	*G. bickii* *G. australe* *G. nelsonii*	Australia
K	12	*G. anapoides* *G. costulatum* *G. cunninghamii* *G. enthyle* *G. exiguum* *G. londonderriense* *G. marchantii* *G. nobile* *G. pilosum* *G. populifolium* *G. pulchellum* *G. rotundifolium*	NW Australia, Cobourg Peninsula, NT

Table 1 (continued)

Genome group	Number of species	Recognized species	Geographic distribution
AD	5	*G. hirsutum* *G. barbadense* *G. tomentosum* *G. mustelinum* *G. darwinii*	New World tropics and subtropics, including Hawaii

and Acosta 2006). The genus is extraordinarily diverse; species morphologies range from fire-adapted, herbaceous perennials in NW Australia to trees in SW Mexico that escape the dry season by dropping their leaves. Corolla colors span a rainbow of blue to purple (*G. triphyllum*), mauves and pinks ("Sturt's Desert Rose", *G. sturtianum*, is the official floral emblem of the Northern Territory, Australia), whites and pale yellows (NW Australia, Mexico, Africa-Arabia) and even a deep sulphur-yellow (*G. tomentosum* from Hawaii). Seed coverings range from nearly glabrous to the naked eye (e.g., *G. klotzschianum* and *G. davidsonii*), to short stiff, dense, brown hairs that aid in wind-dispersal (*G. australe, G. nelsonii*), to the long, fine white fibers that characterize highly improved forms of the four cultivated species. There are even seeds that produce fat bodies to facilitate ant-dispersal (Seelanan, Brubaker, Stewart, Craven and Wendel 1999).

As the genus diversified and spread, it underwent extensive chromosomal evolution (see Chapter 11 by Konan et al. and Chapter 11 by Mergeai, this volume). Although all diploid species share the same chromosome number ($n = 13$), there is more than three-fold variation in DNA content per genome (Hendrix and Stewart 2005). Chromosome morphology is similar among closely related species, and this is reflected in the ability of related species to form hybrids that display normal meiotic pairing and sometimes high F_1 fertility. In contrast, crosses among more distant relatives may be very difficult to effect, and those that are successful are characterized by meiotic abnormalities. The collective observations of pairing behavior, chromosome sizes, and relative fertility in interspecific hybrids led to the designation of single-letter genome symbols (Beasley 1941) for related clusters of species. Presently, eight diploid genome groups (A through G, plus K) are recognized (Endrizzi, Turcotte and Kohel 1985; Stewart 1995). This cytogenetic partition of the genus is largely congruent with taxonomic and phylogenetic divisions. A brief introduction to the major groups of diploid cotton species follows.

3.1 Australian Species

Australian cottons (subgenus *Sturtia*) comprise 16 named species as well as a new species whose description is in press (Stewart, Craven, Brubaker and

Wendel 2008). Collectively, these taxa comprise the C-, G-, and K-genome groups, with two, three, and twelve species, respectively. These three groups of species are implicated by DNA sequence data (Liu, Brubaker, Green, Marshall, Sharp and Singh 2001; Seelanan, Brubaker, Stewart, Craven and Wendel 1999; Seelanan, Schnabel and Wendel 1997) to be natural lineages, consistent with their formal alignments into the taxonomic sections *Sturtia* (C-genome), *Hibiscoidea* (G-genome), and *Grandicalyx* (K-genome). Relationships among the three groups, however, remain unclear. Some data place *G. robinsonii* as basal within the entire assemblage of Australian species (DeJoode and Wendel 1992), suggesting that radiation of *Gossypium* in Australia proceeded eastward from the westernmost portion of the continent. Whether this basal position will withstand the scrutiny of other data sets is an open question, as the most recent analyses (Liu, Brubaker, Mergeai, Cronn and Wendel 2001; Seelanan, Brubaker, Stewart, Craven and Wendel 1999; Seelanan, Schnabel and Wendel 1997) are equivocal in this regard.

With respect to the taxonomy within each of the three Australian genome groups, there is little uncertainty for the C- and G-genome groups, as these are well represented in collections and have been thoroughly studied (DeJoode and Wendel 1992; Fryxell 1979, 1992; Liu, Brubaker, Mergeai, Cronn and Wendel 2001; Seelanan, Brubaker, Stewart, Craven and Wendel 1999; Seelanan, Schnabel and Wendel 1997; Wendel, Stewart and Rettig 1991). Much less certain is the taxonomy of the K-genome species, which are all placed in section *Grandicalyx*. Collecting expeditions to the Kimberley area have enhanced our understanding of diversity within the group and have resulted in the discovery of at least seven new species (Fryxell, Craven and Stewart 1992; Stewart, Craven, Brubaker and Wendel, 2008). These unusual species have a distinctive geography, morphology and ecology, and exhibit a syndrome of features that are characteristic of fire-adaptation. In particular, they are herbaceous perennials with a bi-seasonal growth pattern whereby vegetative growth dies back during the dry season to underground rootstocks that initiate a new cycle of growth with the onset of the next wet season or following a fire. Species in section *Grandicalyx* have flowers that are upright when open but that become pendant following pollination so that at maturity, the capsules release sparsely haired, ant-dispersed seeds that bear elaiosomes (fat bodies) to aid in attracting ants. Many of these species are poorly represented in collections and not well understood taxonomically. Molecular phylogenetic analyses have yielded conflicting results regarding interspecific relationships in this group (Liu, Brubaker, Mergeai, Cronn and Wendel 2001; Seelanan, Brubaker, Stewart, Craven and Wendel 1999).

3.2 African-Asian Species

Fourteen species from Africa and Arabia are recognized in the most recent taxonomic treatment of the genus (Fryxell 1992), collectively in subgenus

Gossypium. The taxonomic section *Gossypium* contains four subsections, whereas section *Serrata* contains only *G. trifurcatum* from the desert area of eastern Somalia. The presence of dentate leaves raised the possibility that it may not belong in *Gossypium*, but recent molecular work clearly established this poorly known entity as a *bona fide* if unusual cotton species (Rapp, Alvarez and Wendel 2005). This latter example underscores the provisional nature of much of the taxonomy of the African-Arabian species of *Gossypium*, which are sorely in need of basic plant exploration and systematic study. Within section *Pseudopambak*, species recognition and definition are in some cases based on limited herbarium material (e.g., *G. benadirense, G. bricchettii, G. vollesenii*) and seeds have not been collected. Consequently, no analyses have been conducted on cytogenetic characteristics nor molecular phylogenetic affinities.

From a cytogenetic standpoint, the African-Arabian species exhibit considerable diversity, collectively accounting for four of the eight genome groups (A-, B-, E-, and F-). The A-genome group comprises the two cultivated cottons of subsection *Gossypium*, *G. arboreum* and *G. herbaceum*. Three African species in subsection *Anomala* (*G. anomalum, G. captis-viridis* and *G. triphyllum*) comprise the B genome. *Gossypium trifurcatum* may also belong to the B genome, but this has not been established. The sole F-genome species, *G. longicalyx*, is cytogenetically distinct (Phillips 1966), morphologically isolated (Fryxell 1971; Fryxell 1992) and, according to Fryxell (1979), is perhaps adapted to more mesic conditions than any other diploid *Gossypium* species. The remaining African-Asian species, those of subsection *Pseudopambak*, are considered to possess E-genomes, although this has yet to be verified.

3.3 American Diploid Species

Subgenus *Houzingenia* contains two sections and six subsections, whose species collectively represent the New World D-genome diploids. These species have been more thoroughly studied than most, and consequently their taxonomy is reasonably well-understood. This subgenus has also received considerable phylogenetic attention (Álvarez, Cronn and Wendel 2005; Cronn, Small, Haselkorn and Wendel 2002; Cronn, Small, Haselkorn and Wendel 2003; Small and Wendel 2005; Wendel and Albert 1992), which provides support for the naturalness of most of the subsections. Evolutionary relationships among the apparently natural subsections are less certain, however (Álvarez, Cronn and Wendel 2005), although available evidence suggests that *G. gossypioides* is basal-most within the subgenus (Cronn, Small, Haselkorn and Wendel 2003).

Twelve of the 14 D-genome diploid species are endemic to western Mexico, thus this area is the center of diversity of the D genome. Likely, the lineage became established and initially diversified in this region. Later range extensions are inferred to have arisen from relatively recent (probably Pleistocene) long-distance dispersals, leading to the evolution of endemics in Peru (*G. raimondii*) and the Galapagos Islands (*G. klotzschianum*).

4 Origin and Diversification of the Polyploid Cottons

Classic cytogenetic investigations demonstrated that the American tetraploid species are allopolyploids containing two resident genomes, an A-genome from Africa or Asia, and a D-genome similar to those found in the American diploids (Beasley 1940; Denham 1924; Harland 1940; Skovsted 1934, 1937; Webber, 1935). Additional support for the hypothesis of an allopolyploid origin of the American tetraploids emerged in subsequent decades from numerous sources of evidence, including the synthesis of additional experimental allotetraploids (Stewart 1995). This history and evidence is detailed in Endrizzi, Turcotte and Kohel (1985), Wendel and Cronn (2003) and Chapter 11 by Mergeai, this volume.

When did allopolyploid cottons first form, and how did this happen given that the two diploid genomes involved (A and D) presently exist in species from different hemispheres? This question was a classic botanical mystery for over half a century (see Endrizzi, Turcotte and Kohel 1985), a mystery at least partially solved through the recent use of molecular technologies. With respect to the first part of the question, that of "when", gene sequence data convincingly demonstrate that allopolyploid *Gossypium* originated prior to the evolution of modern humans but relatively recently in geological terms, perhaps 1–2 mya, or in the mid-Pleistocene (Cronn, Small, Haselkorn and Wendel 2002; Seelanan, Schnabel and Wendel 1997; Senchina, Alvarez, Cronn, Liu, Rong, Noyes, Paterson, Wing, Wilkins and Wendel 2003). With respect to the second part of the question, that of polyploid parentage, it is now clear that both extant A-genome species (*G. arboreum, G. herbaceum*) are equally divergent from the A-genome of allopolyploid cottons and that the closest living relative of the progenitor D-genome donor is *G. raimondii* (Endrizzi, Turcotte and Kohel 1985; Wendel and Cronn 2003). One aspect of the history of the polyploid cottons that has become clear is that they all contain an A-genome cytoplasm, and most likely from a single source (Galau and Wilkins 1989; Small and Wendel 1999; Wendel 1989). Studies using nuclear (bi-parentally inherited) genes lead to the same conclusion. Hence, evidence indicates that natural allopolyploid cottons all derive from a single lineage.

Given a Pleistocene origin for allopolyploid cotton species, one may infer that their morphological diversification and spread must have been relatively rapid following polyploidization. At present, five allopolyploid species are recognized. *Gossypium darwinii* is native to the Galapagos Islands, where it may form large and continuous populations in some areas (Percy and Wendel 1990). *Gossypium tomentosum*, from the Hawaiian Islands, has a more diffuse population structure, occurring mostly as scattered individuals and small populations on several islands (DeJoode and Wendel 1992). A third allopolyploid, *G. mustelinum*, is an uncommon species restricted to a relatively small region of northeast Brazil (Wendel, Rowley and Stewart 1994). In addition to these three truly wild species, there are two cultivated species (*G. barbadense* and

G. hirsutum), each of which has a large indigenous range, collectively encompassing a wealth of morphological forms that span the wild-to-domesticated continuum (Brubaker and Wendel 1993, 1994, 2001; Fryxell 1979; Hutchinson 1951; Percy and Wendel 1990). *Gossypium hirsutum* is widely distributed in Central and northern South America, the Caribbean, and even reaches distant islands in the Pacific (Solomon Islands, Marquesas). *Gossypium hirsutum* is thought to have a more northerly distribution than *G. barbadense,* with wild populations occurring as far north (27°38'N) as Tampa Bay (Stewart, personal observation). However, recently Stewart and Bertoni (2007 unpublished) collected a wild population of *G. hirsutum* in the Chaco state of Presidente Hayes, Paraguay (~22°S). *Gossypium barbadense* has a more southerly indigenous range, centered in the northern third of South America but with a large region of range overlap with *G. hirsutum* in the Caribbean.

Consideration of the distribution of the allopolyploid species suggests that polyploidy led to the invasion of a new ecological niche. Fryxell (1965, 1979) noted that in contrast to the majority of diploid species, allopolyploid species typically occur in coastal habitats, at least those forms that arguably are truly wild. Two species, both island endemics (*G. darwinii* and *G. tomentosum*), are restricted to near coastlines, and for two others (*G. barbadense* and *G. hirsutum*), wild forms occur in littoral habitats ringing the Gulf of Mexico, northwest South America, and even on distant Pacific Islands. Fryxell (1965, 1979) speculated that following initial formation, adaptation of the newly evolved allopolyploid to littoral habitats enabled it to exploit the fluctuating sea levels that characterized the Pleistocene. This ecological innovation is envisioned to have facilitated initial establishment of the new polyploid lineage, and also may have provided a means for the rapid dispersal of the salt-water tolerant seeds. The recent discovery of *G. hirsutum* in Paraguay does not negate this hypothesis, in that the population was located next to an area of seasonal water accumulation suspected of being saline due to the many years of water accumulation followed by evaporation. The timing and source of the initial introduction of cotton to this area are unknown, but could involve bird dispersal.

5 Phylogenetic Relationships in the Genus

A genealogical framework for the genus has been provided by multiple molecular phylogenetic investigations (reviewed in Wendel and Cronn 2003). Each of these studies shows that genealogical lineages of species are congruent with genome designations and geographical distributions. Cytogenetic studies (reviewed in Chapter 11, this volume) further support this conclusion. Accordingly, each genome group corresponds to a single natural lineage, and in most cases, these lineages are also geographically cohesive. This information is summarized in a depiction of our present understanding of relationships (Fig. 1).
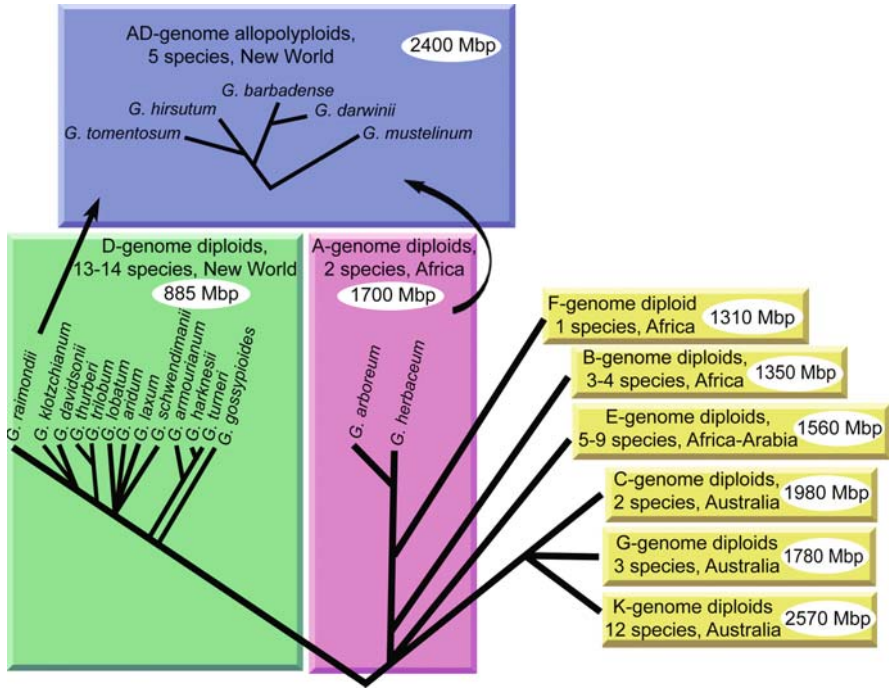
Fig. 1 Evolutionary history of *Gossypium*, as inferred from multiple molecular phylogenetic data sets. The closest relative of *Gossypium* is a lineage containing the African-Madagascan genus *Gossypioides* and the Hawaiian endemic genus *Kokia*. Following its likely origin 5–10 mya, *Gossypium* split into three major diploid lineages: the New World clade (D-genome); the African-Asian clade (A-, B-, E- and F-genomes); and the Australian clade (C-, G-, and K-genomes). This global radiation involved several trans-oceanic dispersal events and was accompanied by morphological, ecological, and chromosomal differentiation (2C genome sizes shown in white ellipses). Interspecific hybridization is implicated in the evolution of approximately one-fourth of the genus. Allopolyploid cottons formed following trans-oceanic dispersal of an A-genome diploid to the Americas, where the new immigrant underwent hybridization, as female, with a native D-genome diploid similar to modern *G. raimondii*. Polyploid cotton probably originated during the Pleistocene (1–2 mya), with the five modern species representing the descendants of an early and rapid colonization of the New World tropics and subtropics (*See* Color Insert)

Several aspects of the phylogenetic history of *Gossypium* bear highlighting. *First*, there exist four major lineages of diploid species corresponding to three continents: Australia (C-, G-, K-genomes), the Americas (D-genome), and Africa/Arabia (two lineages: one comprising the A-, B-, and F-genomes, and a second containing the E-genome species). *Second*, the earliest divergence event in the genus separated the New World D-genome lineage from the ancestor of all Old World taxa. Thus, New World and Old World diploids are phylogenetic sister groups. Following this basal-most split in the genus, cottons comprising the Old World lineage divided into three groups, namely, the Australian cottons

(C-, G-, and K-genome species), the African-Arabian E-genome species, and the African A-, B-, and F-genome cottons. *Third*, the African F-genome clade, which consists of the sole species *G. longicalyx*, is definitively diagnosed as sister to the A-genome species. This identifies the wild forms most closely related to those first domesticated in the A-genome species *G. arboreum* and *G. herbaceum*. Because this relationship is revealed, prospects are raised for ultimately understanding the genetic basis of the origin of useful, long lint. *Fourth*, the major lineages of *Gossypium* were established in relatively rapid succession shortly after the genus originated and diverged from the *Kokia-Gossypioides* clade. The evolutionary picture thus envisioned is that there was a rapid and global radiation early in the history of the genus, with temporally closely spaced divergence events.

Allopolyploid cottons, including the two species of great commercial importance today (*G. hirsutum* and *G. barbadense*) are implicated to have formed following a biological reunion of two genomes (A and D) that descended from the earliest split in the genus. That is, the two constituent genomes of allopolyploid cotton evolved first in different hemispheres and diverged for millions of years, in isolation from one another. Allopolyploid cottons thus contain duplicated but slightly divergent copies of most genes. Senchina et al. (2003) studied 48 pairs of these duplicates, and showed that on average, there is about 3–4% sequence divergence between copies, although there is considerable variation about this mean.

Consideration of the phylogeny of Fig. 1 in a temporal context and in light of plate tectonic history leads to an inference of multiple intercontinental dispersal events and other episodes of trans-oceanic travel during the evolutionary history of the *Gossypium*. These include at least one dispersal between Australia and Africa, another to the Americas (probably Mexico) leading to the evolution of the D-genome diploids, and a second, much later colonization of the New World by the A-genome ancestor of the AD-genome allopolyploids. Long-distance dispersal played a role not only in diversification of major evolutionary lines but also in speciation within *Gossypium* genome groups. Examples include dispersals from southern Mexico to Peru (*G. raimondii*), from northern Mexico to the Galapagos Islands (*G. klotzschianum*), from western South America to the Galapagos Islands (*G. darwinii*), from Africa to the Cape Verde Islands (*G. capitis-viridis*), and from the neotropics to the Hawaiian Islands (*G. tomentosum*).

6 Seed and Fiber Diversity and Seed Dispersal

An additional perspective offered by the phylogeny of the genus (Fig. 1) concerns the evolutionary history of seeds and their associated single-celled epidermal trichomes, known as hairs in the wild species and lint and fuzz fibers in the cultivated taxa. The seeds and their coverings are extraordinarily
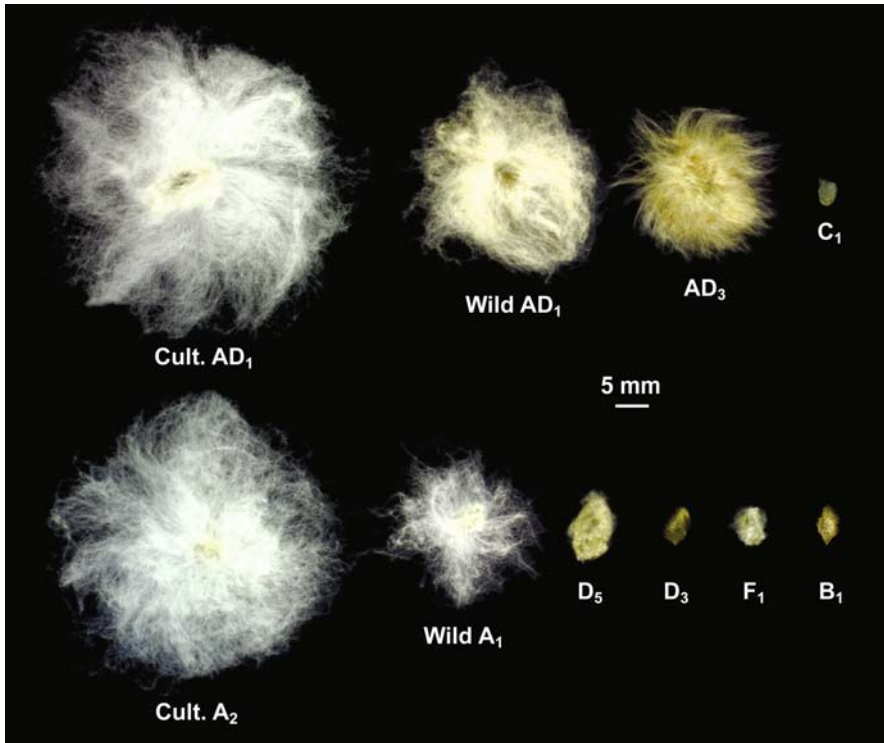
Fig. 2 Representative seed and trichome diversity in *Gossypium*. Seed and trichome size and morphology are exceedingly variable in the genus. Most wild species have relatively small seeds (<5 mm in any dimension) with equally short fibers. Long (spinnable) fiber evolved only once, in the ancestor of modern A-genome cottons, which subsequently donated this capacity to modern tetraploid species, including the commercially important *G. hirsutum* and *G. barbadense*, at the time of allopolyploid formation in the mid-Pleistocene. See text for additional detail. Key to species: Cult. AD_1 = *G. hirsutum* TM1; Wild AD_1 = *G. hirsutum* Tx2094 from the Yucatan Peninsula; AD_3 = *G. tomentosum* WT936 from Hawaii; C_1 = *G. sturtianum* C_1-4 from Australia; Cult. A_2 = *G. arboreum* AKA8401; Wild A_1 = *G. herbaceum* subsp. *africanum* from Botswana; D_5 = *G. raimondii* from Peru; D_3 = *G. davidsonii* D_{3d}-32 from Baja California; F_1 = *G. longicalyx* F_1-3 from Tanzania; B_1 = G. *anomalum* B_1-1 from Africa (*See* Color Insert)

diverse in *Gossypium*, as shown in Fig. 2, which shows a sampling of the different lineages of wild species and two cultivated species (*G. arboreum* and *G. hirsutum*) for comparison. Both wild and cultivated cottons produce fiber on the seed coat, but there are striking morphological and structural differences between these fibers, the most obvious of which is their size. Some D-genome species (*G. thurberi, G. trilobum, G. davidsonii,* and *G. klotzschianum*) do not possess obvious seed hairs, but they actually are present as developmentally repressed structures not visible to the unaided eye (Applequist, Cronn and Wendel 2001). Similarly, the three D-genome species of subsection *Cauducibracteolata* give the appearance of being hairless but, in fact,

have seed hairs that are tightly appressed to the seed. Cultivated lint fiber is a single cell of almost pure cellulose, which elongates during development, depending on the species, to a final length of two to six centimeters (Kim and Triplett 2001). The wild-type fiber cell is composed mostly of cellulose and suberin, and elongates to less than one centimeter (Ryser and Holloway 1985; Applequist, Cronn and Wendel 2001). These and many other differences (Fryxell 1979; Hutchinson and Stephens 1945) reflect both natural evolutionary processes as well as human-mediated selection during domestication. The duration of the elongation phase and the timing of onset of secondary wall synthesis appear to be key determinants of the final length of the fiber in both wild and cultivated plants (Applequist, Cronn, and Wendel, 2001). Phylogenetic analysis of growth rates has shown that the evolutionary innovation of prolonged elongation arose in the F-genome/A-genome lineage, which may have facilitated the original domestication of the A-genome cottons. This trait of prolonged elongation was passed on to the allopolyploids, which in turn was a key component of their eventual domestication (Applequist, Cronn, and Wendel 2001). Relatively little is known about the developmental and genetic underpinnings of the diverse morphologies illustrated in Fig. 2, but an improved understanding of the changes that occurred during evolution and domestication may have implications for crop improvement.

In nature, seed dispersal in *Gossypium* often follows a "shaker" model, where erect, mature capsules dehisce along the sutures and the seeds are distributed near the parental plant as wind shakes the branches. This likely is the ancestral method of seed dispersal in the genus, as it occurs in all species of the B-, C-, E-, F-, and D- (in part) genome groups and in one species (*G. bickii*) of the G-genome group. The A-, AD-, some D-genome species, and G- and K-genome species have evolved other mechanisms of seed dispersal, with perhaps the most noteworthy innovation being the fat bodies to facilitate ant dispersal on the seeds of the K-genome species (Seelanan, Brubaker, Stewart, Craven and Wendel 1999). The development of spinnable fibers apparently has occurred only once in the history of *Gossypium,* in the ancestor of the two A-genome species which became the progenitor of the A-subgenome of the tetraploids.

The evolutionary "purpose" of epidermal seed hairs, or trichomes, is a matter of speculation. Fryxell (1979) suggested that elongated fibers aid dispersal by birds, an hypothesis that gains credibility by observations by one of us (J. Stewart) of a bird's nest in NW Puerto Rico that contained numerous seeds of feral *G. hirsutum*, as well as a collection of *G. darwinii* from a finch's nest in the Galapagos Islands (JFW, unpubl.). One might also speculate that fibers serve to inhibit germination unless there is sufficient moisture to saturate the fibers; should germination occur following a light rain, there might not be sufficient water for subsequent survival of the seedling. In this respect the waxy coating of the fibers would repel water, to a point, and thus prevent premature germination. A related possibility is that seed hairs function as "biological incubators" to facilitate germination only when ecological conditions are

appropriate, by recruiting particular microbial communities under appropriate moisture regimes. Finally, two G-genome species (*G. australe, G. nelsonii*) have evolved stiff straight seed hairs that facilitate wind dispersal in that the stiffening hairs function to extrude seeds from the locules of the dehisced capsule.

7 Genetic Diversity in the Cultivated Cottons

Domestication has altered the genetic composition of the four cultivated species in comparison to their wild relatives. As is typical of most wild *Gossypium* species, the pre-domesticated cotton species probably had restricted distributions (Brubaker, Bourland and Wendel 1999), although we cannot be certain of this because post-domestication, human-mediated germplasm diffusion has partially (*G. barbadense* and *G. hirsutum*) or completely (*G. herbaceum* and *G. arboreum*) obscured the geographic and genetic origins of each species. It is clear, however, that the indigenous ranges of the four cotton species expanded considerably as cultigens. Eventually this brought *G. barbadense* and *G. hirsutum* into sympatry in the New World, and in the Old World the ranges of *G. herbaceum* and *G. arboreum* also intersected. The consequence was interspecific gene flow prior to the initiation of deliberate bi-specific breeding programs (*e.g*., the Acala and Pima cultivar families). Among indigenous New World cotton populations, natural introgression patterns appear to be biased toward the transfer of *G. barbadense* alleles into *G. hirsutum*, particularly in the Caribbean where the ranges of the two species overlap to the greatest extent (Wendel, Brubaker and Percival 1992; Brubaker and Wendel 1994; Brubaker, Koontz and Wendel 1993; Percy and Wendel 1990). Among the modern elite cultivars, however, the converse is more prevalent. Estimates based on RFLP data suggest that up to 8.9% of alleles in modern *G. barbadense* cultivars may have originated from *G. hirsutum* (Wang, Dong and Paterson 1995). The Old World cotton species co-occur in India, northern Africa and China, and there is evidence for low levels of interspecific gene flow, which may explain in part the considerable morphological overlap between the two species (Wendel, Olsen and Stewart 1989).

While pre- and post- domestication interspecific gene flow may have mingled the genetic composition of cotton species to some extent, the genetic winnowing that accompanied human selection progressively narrowed the total available genetic diversity. In the earliest stages of domestication, appreciable levels of genetic diversity were captured among geographic foci of agronomic development, often recognized as "races" (Hutchinson 1951; Hutchinson 1954; Hutchinson and Ghose, 1937; Hutchinson 1962; reviewed by Brubaker, Bourland and Wendel 1999), but the modern elite cultivars now contain only a fraction of this original diversity. As facultative self-pollinators, cotton species tolerate inbreeding, and the most recent stages of domestication accelerated the progressive purge of alleles, resulting in modern elite cultivars with reduced levels of genetic diversity.

Table 2 Comparative genetic diversity estimates for cotton species

	Description	Genome	No of Loci	Percent Polymorphic Loci (P)	Mean # Alleles per Locus (A)	Mean Panmictic Heterozygosity (H_T)	Reference
	449 Plant Species			50	1.96	0.15	Hamrick & Godt 1990
	Endemics			40	1.80	0.10	Hamrick & Godt 1990
	Island endemics			25	1.32	0.06	Hamrick & Godt 1990
G. arboreum	feral, commensal, modern improved cultivars	A	40	28	2.36	0.066	Wendel et al., 1989
G. herbaceum	wild, feral, commensal, modern improved cultivars	A	40	25	2.00	0.039	Wendel et al., 1989
G. barbadense	wild, feral, commensal, modern improved cultivars	AD	59	41	1.69	0.06	Percy & Wendel 1990
	modern improved cultivars	AD	59	36	1.39	0.058	
G. darwinii	wild island endemic	AD	59	27	1.34	0.062	Wendel & Percy 1990
G. hirsutum	wild, feral, commensal, modern improved cultivars	AD	50	60	2.30	0.163	Wendel et al., 1992
	modern improved cultivars	AD	50	28	1.28	0.056	
G. mustelinum	rare Brazilian endemic	AD	50	12	1.14	0.08	Wendel et al. 1994
G. tomentosum	wild island endemic	AD	50	22	1.24	0.033	DeJoode & Wendel 1992

Despite the ready availability of molecular markers over the past ten years, the most informative set of comparative data regarding genetic diversity of *Gossypium* species can be found in a series of early isozyme studies (Table 2). The numbers indicate that the wild tetraploid species have levels of genetic diversity consistent with their occurrence as island endemics (*G. darwinii* and *G. tomentosum*) or their rarity (*G. mustelinum*). To the extent that the progenitor populations of the cotton species were similarly restricted in distribution, they might have been similarly depauperate genetically. The cotton species as they currently exist, however, contain appreciable albeit unremarkable levels of genetic diversity relative to plants in general (Table 2). Of the four cotton species, *G. hirsutum* contains the highest levels of genetic diversity, and it is here that the genetic consequences of domestication are most evident, as the modern upland cultivars contain only half of the genetic diversity available. *Gossypium barbadense* cultivars, having undergone less intensive selection, retain a greater proportion of the diversity present in the species as a whole. More recent molecular marker studies (using RFLPs, RAPDs, AFLPs, or SSRs), while not directly comparable and while focused on *G. hirsutum,* have validated these initial genetic "portraits" (Abdalla, Reddy, El-Zik and Pepper 2001; Brubaker and Wendel 1994; Hawkins, Pleasants and Wendel 2005; Iqbal Reddy, El-Zik and Pepper 2001; Khan, Hussain, Askari, Stewart, Malik and Zafar 2000; Lacape, Dessauw, Rajab, Noyer and Hau 2007; Liu, Cantrell, McCarty and Stewart 2000; Liu, Guo, Lin, Nie and Zhang 2005; Pillay and Myers 1999; Rungis, Llewellyn, Dennis and Lyon 2005), with paucity of genetic diversity among the modern Upland cultivars as a recurring theme (Lacape, Dessauw, Rajab, Noyer, and Hau 2007). One practical consequence of this, exacerbated by the presence of a genome that contains 26 chromosomes, is that marker-assisted selection in Upland cotton breeding is less advanced than it is for other crops. Accordingly, geneticists continue to exploit the more accessible interspecific genetic differences between *G. barbadense* and *G. hirsutum* and between *G. arboreum* and *G. herbaceum* (Ulloa, Brubaker and Chee 2007).

8 Diversity and Crop Improvement

One of the opportunities that arises from an appreciation of the evolutionary history and diversity of *Gossypium* is that of deliberate introgression of useful genes from various wild sources of germplasm. It is a truism that the morphological and ecological breadth encompassed by the wild species must have parallels in the underlying genes that control, for example, physiology, chemistry, disease resistance traits, and fiber characteristics. The wild species of cotton, consequently, represent an ample genetic repository for potential exploitation by cotton breeders. Although these wild species remain a largely untapped genetic resource, there are many examples of their inclusion in breeding programs (reviewed by McCarty and Percy 2001). Cotton improvement