Tropical Plant Biol. (2010) 3:71–74
DOI 10.1007/s12042-010-9051-4

Sequencing and Utilization of the *Gossypium* Genomes

Andrew H. Paterson · Jun-kang Rong · Alan R. Gingle · Peng W. Chee · Elizabeth S. Dennis · Danny Llewellyn · Leon S. Dure III · Candace Haigler · Gerald O. Myers · Daniel G. Peterson · Mehboob ur Rahman · Yusuf Zafar · Umesh Reddy · Yehoshua Saranga · James M. Stewart · Joshua A. Udall · Vijay N. Waghmare · Jonathan F. Wendel · Thea A. Wilkins · Robert J. Wright · Essam Zaki · Elsayed E. Hafez · Jun Zhu

Received: 23 February 2010 / Accepted: 15 March 2010 / Published online: 14 April 2010

Abstract Revealing the genetic underpinnings of cotton productivity will require understanding both the prehistoric evolution of spinnable fibers, and the results of independent domestication processes in both the Old and New Worlds. Progress toward a reference sequence for the smallest *Gossypium* genome is a logical stepping-stone toward revealing diversity in the remaining seven genomes (A, B, C, E, F, G, K) that permitted *Gossypium* species to adapt to a wide range of ecosystems in warmer arid regions of the world, and toward identifying the emergent properties that account for the superior productivity and quality of tetraploid cottons. The greatest challenge facing the cotton community is not genome sequencing *per se* but the conversion of sequence to knowledge.

Keywords Spinnable fiber · Genetic bottleneck · Transgenic · Translational genomics

During the recently ended International Year of Natural Fibers (http://www.naturalfibres2009.org/), it is fitting that progress in sequencing of genomes in the cotton genus

Communicated by: Paul Moore

A. H. Paterson (✉) · J.-k. Rong · A. R. Gingle
Plant Genome Mapping Laboratory, University of Georgia,
111 Riverbend Road Rm 228,
Athens, GA 30602, USA
e-mail: paterson@uga.edu

P. W. Chee
Coastal Plain Experiment Station, University of Georgia,
Tifton, GA 31794, USA

E. S. Dennis · D. Llewellyn
CSIRO Plant Ind,
Canberra ACT 2601, Australia

L. S. Dure III
Department of Biochemistry, University of Georgia,
Athens, GA 30602, USA

C. Haigler
Departments of Crop Science and Botany,
North Carolina State University,
Raleigh, NC 27695, USA

G. O. Myers
LSU AgCenter, Louisiana State University,
Baton Rouge, LA 70803, USA

D. G. Peterson
Mississippi Genome Exploration Laboratory,
Mississippi State University,
Mississippi State, MS 39762, USA

M. ur Rahman · Y. Zafar
Plant Genomics & Molecular Breeding Labs,
National Institute for Biotechnology & Genetic Engineering,
Faisalabad, Pakistan

U. Reddy
Department of Biology, West Virginia State University,
Institute, WV 25112, USA

Y. Saranga
The Robert H. Smith Faculty of Agriculture,
Food and Environment, The Hebrew University of Jerusalem,
Rehovot, Israel

J. M. Stewart
Department Crop, Soil, and Environmental Sciences,
University of Arkansas,
Fayetteville, AR 72701, USA

(*Gossypium*) accelerated rapidly, toward the realization of many novel opportunities to advance knowledge of organic evolution. Of singular importance is dissecting the evolution of the 'lint fiber' that sustains the textile industry, with an aggregate influence estimated at ~$120 billion/yr on US gross domestic product and ~$500 billion/yr worldwide. "*There are only a few cells in the plant kingdom that are as exaggerated in their size or composition as cotton fibers*", and some of these single-celled seed epidermal trichomes "*... reach lengths of over 6 cm, or one-third the height of an Arabidopsis plant* (Kim and Triplett 2001)."

Cotton is unusual among major crops in having been domesticated independently four times at two different ploidy levels. Spinnable fibers evolved in the Old World A genome lineage in the past 5–7 million years (Senchina et al. 2003; Udall et al. 2006). Domestication of A genome cottons *G. herbaceum* and/or *G. arboreum* may have started before 6000 B.C. in Pakistan (Moulherat et al. 2002). In parallel, by 3500–2300 BC (Stephens and Moseley 1974) New World aboriginals were utilizing two tetraploid species that arose from natural hybridization between an A genome species and a New World D genome species. A and D genome taxa diverged ~5–10 million years ago (Senchina et al. 2003; Udall et al. 2006), reuniting by polyploidization ~1–2 million years ago following trans-oceanic dispersal of an A genome propagule to the New World (Wendel 1989). The ancestral allopolyploid spawned two species that were independently domesticated (*G. hirsutum*, or 'Upland' cotton; and *G. barbadense*, including forms referred to as 'Sea Island', Egyptian, and Pima cotton), and three species known only in the wild, native to the Galapagos (*G. darwinii*), Hawaii (*G. tomentosum*), and Brazil (*G. mustelinum*).

Revealing the genetic underpinnings of cotton productivity will require understanding both the prehistoric evolution of spinnable fibers, and the results of independent domestication processes in both the Old and New Worlds. In particular, the New World D genome (similar to extant G. *raimondii*) played a surprising role in cotton improvement. Although no D genome species produce spinnable fiber, more than half of genetic differences in fiber traits between the two domesticated tetraploid species map to D-genome chromosomes (Jiang et al. 1998; Rong et al. 2007). Moreover, gene expression in tetraploid cotton fiber shows a like bias in favor of D-genome alleles (Hovav et al. 2008). These data support the hypothesis that the superior fiber yield and quality of tetraploids may be an emergent property of combining two genomes (Jiang et al. 1998). Indeed, cotton has gone 'full circle'—evolution of spinnable fibers may have unwittingly provided the Old World A genome a dispersal mechanism by which to transiently colonize the New World and permit the tetraploid to form. In turn, in the post-Columbian era, more productive and finer-quality New World tetraploids have largely supplanted cultivated diploids in the Old World.

Cotton enjoys many opportunities to participate in a bio-based products revolution that may reduce dependence on petrochemicals (Council 2000). Cotton fiber with increased uniformity, durability, and strength might replace synthetic fibers that require ~230 million barrels of petroleum per year to produce in the USA alone. Cotton seed oil, and byproducts of fiber processing, are raw materials for biofuel production (Holt et al. 2003).

Discovery and utilization of new *Gossypium* diversity may be especially important for sustainable cotton production because of its narrow gene pool (Chee et al. 2004; Lubbers et al. 2004). The natural 'genetic bottleneck' imposed by polyploid formation has been exacerbated by repeatedly crossing relatively few closely-related genotypes to one another to breed new cultivars (May et al. 1995) and using only a few cultivars to deploy transgenes (Helms 2000). For example, a looming worldwide water crisis (UNESCO 2002) makes it important to identify adaptations that permitted wild cottons to endure periodic drought and temperature extremes (Kohel et al. 1974), restoring such valuable alleles that may have been "left behind" during domestication (Gur and Zamir 2004) to create cultivars that produce more with less (water).

DNA sequencing promises to reveal the spectrum of diversity in the *Gossypium* genus. A high degree of

J. A. Udall
Department of Plant & Wildlife Sciences,
Brigham Young University,
Provo, UT 84602, USA
e-mail: jaudall@byu.edu

V. N. Waghmare
Central Institute for Cotton Research,
Nagpur, Maharashtra 440010, India

J. F. Wendel
Department of Ecology, Evolution and Organismal Biology,
Iowa State University,
Ames, IA 50011, USA

T. A. Wilkins · R. J. Wright
Department of Plant and Soil Science, Texas Tech University,
Lubbock, TX 79409, USA

E. Zaki
Nucleic Acids Research Department,
Genetic Engineering & Biotechnology Research Institute,
Borg El Arab Post Code 21934 Alexandria, Egypt

E. E. Hafez
Mubarak City for Scientific Research
and Technology Applications,
New Borg El Arab City 21934 Alexandria, Egypt

J. Zhu
Institute of Bioinformatics, Zhejiang University,
Hangzhou, Peoples Republic of China

conservation of gene order and sequence suggests that the vast majority of data from diploids will extrapolate to tetraploids (Rong et al. 2004). Accordingly, obtaining a reference sequence of the smallest *Gossypium* genome (D, ~900 Mb) is a logical stepping-stone toward characterizing the larger A diploid (~1700 Mb) and AD tetraploid genomes (~2500 Mb) (Paterson 2007; Chen et al. 2007). Rapid low cost re-sequencing might then be sufficient to reveal diversity in the remaining six genomes (B, C, E, F, G, K) that permitted *Gossypium* species to adapt to a wide range of ecosystems in warmer, arid regions of the world. The US Department of Energy Joint Genome Institutehas completed a 0.4x genome-equivalent 'pilot study' of *G. raimondii* that strongly supports the feasibility of assembling a whole-genome shotgun (WGS) sequence (A.H.P. and X. Wang, unpubl. data), and has begun further sequencing (www.jgi.doe.gov/sequencing/cspseqplans2009.html). Early explorations of the A and AD genomes are also in progress.

As a leading crop in the implementation of transgenes in agriculture, a reference genome sequence may expedite ongoing development and stewardship of genetically-modified (GM) cotton. It will become easy to determine whether each transgene insertion site is in euchromatin or heterochromatin, and identify any genes inadvertently disrupted. Identification of genomic characteristics associated with favorable expression of transgenic traits might reduce the need for costly empirical testing of numerous transgenic insertions to commercialize one. Unifying principles of useful transgene insertions might be found by comparison to the only transgenic plant sequenced to date, papaya, in which five of six insertions were in nuclear-encoded DNA fragments of chloroplast origin, with four matching topoisomerase I recognition sites (Ming et al. 2008). Using the sequence to identify DNA markers closely linked to transgenes may reduce the undesirable chromatin (and traits) transmitted to elite genotypes from the otherwise-obsolete cottons that are most efficiently transformed.

The greatest challenge facing the cotton community is not genome sequencing *per se* but the conversion of sequence to knowledge. Completion of the *Arabidopsis thaliana* sequence was quickly followed by inception of the NSF 2010 project, which has greatly increased knowledge about the functions of *Arabidopsis* genes at a cost approaching $200 million. While the functions of perhaps half of the cotton genes might be deduced by analogy to those of *Arabidopsis* (Rong et al. 2005), *de novo* functional analysis of the remaining cotton genes faces the disadvantages of ~20 times as much DNA, the necessity of completing its longer life cycle to see effects on the primary organ of commerce (seedborne lint fiber), and a larger body that cannot complete its life cycle in a test tube.

To realize the potential economic benefits of sequencing the cotton genomes will require investments of at least the same order-of-magnitude made in *Arabidopsis*. Had *Arabidopsis* not gone first the cost of cotton functional genomics would be much higher. Much of the required investment will need to come from the private sector, but few single enterprises have the critical mass of knowledge, skills, and resources needed to accomplish such innovation alone. Cotton is an attractive target for public-private partnership to develop enabling tools that will nurture rapid accumulation of fundamental information necessary to empower development and commercialization of products and applications across the value chain.

References

Chee P, Lubbers E, May O, Gannaway J, Paterson AH (2004) Changes in genetic diversity of the U.S. Upland cotton. Beltwide Cotton Conference. National Cotton Council, San Antonio

Chen ZJ et al (2007) Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol 145:1303–1310

Council NR (2000) Biobased industrial products: priorities for research and commercialization

Gur A, Zamir D (2004) Unused natural variation can lift yield barriers in plant breeding. Plos Biology 2:1610–1615

Helms AB (2000) Yield study report. In: Dugger P, Richter D (eds) Proc Beltwide Cotton Prod Conf. Natl. Cotton Council, San Antonio

Holt G, Simonton J, Beruvides M, Canto AM (2003) Engineering economic analysis of a cotton by-product fuel pellet operation. J Cotton Sci 7:205–216

Hovav R, Udall JA, Hovav E, Rapp RA, Flagel L, Wendel JF (2008) Gene expression during cellular differentiation of the single-celled cotton trichome (fiber). Planta 227:319–329

Jiang CX, Wright RJ, El-Zik KM, Paterson AH (1998) Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proc Natl Acad Sci USA 95:4419–4424

Kim JK, Triplett BA (2001) Cotton fiber growth in planta and in vitro. Models for plant cell elongation and cell wall biogenesis. Plant Physiol 127:1361–1366

Kohel RJ, Richmond TR, Lewis CF (1974) Genetics of flowering response in cotton. VI. Flowering behavior of Gossypium hirsutum L. and G. barbadense L. hybrids. Crop Sci 14:696–699

Lubbers E, Chee P, Gannaway J, Wright R, El-Zik K, Paterson AH (2004) Levels and patterns of genetic diversity in upland cotton. Plant and Animal Genome XII Conference, San Diego

May OL, Bowman DT, Calhoun DS (1995) Genetic diversity of U.S. upland cotton cultivars released between 1980 and 1990. Crop Sci 35:1570–1574

Ming R et al (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452:991–997

Moulherat C, Tengberg M, Haquet J-F, Mille B (2002) First evidence of cotton at Neolithic Mehrgarh, Pakistan: analysis of mineralized fibres from a copper bead. J Archaeol Sci 29:1393–1401

Paterson AH (2007) Sequencing the cotton genomes. World Cotton Research Conference. International Cotton Advisory Committee, Lubbock

Rong J-K et al (2004) A 3347-locus genetic recombination map of sequence-tagged sites reveals features of genome organization, transmission and evolution of cotton (Gossypium). Genetics 166:389–417

Rong J, Bowers JE, Schulze SR, Waghmare VN, Rogers CJ, Pierce GJ, Zhang H, Estill JC, Paterson AH (2005) Comparative genomics of Gossypium and Arabidopsis: unraveling the con-

sequences of both ancient and recent polyploidy. Genome Res 15:1198–1210

Rong J-K, Feltus FA, Waghmare VN, Pierce GJ, Chee PW, Draye X, Saranga Y, Wright RJ, Wilkins TA, May OL, Smith CW, Gannaway JR, Wendel JF, Paterson AH (2007) Meta-analysis of polyploid cotton QTLs shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. Genetics 176:2577– 2588

Senchina DS, Alvarez I, Cronn RC, Liu B, Rong JK, Noyes RD, Paterson AH, Wing RA, Wilkins TA, Wendel JF (2003) Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. Mol Biol Evol 20:633–643

Stephens SG, Moseley ME (1974) Early domesticated cottons from archaeological sites in central coastal Peru. Am Antiquity 39:109–122

Udall JA et al (2006) A global assembly of cotton ESTs. Genome Res 16:441–450

UNESCO (2002) Vital water graphics, water use and management. United Nations Education Scientific and Cultural Organization, Paris

Wendel JF (1989) New world tetraploid cottons contain old-world cytoplasm. Proc Natl Acad Sci USA 86:4132–4136