The Post-Genomic Era for Cotton

Andrew H. Paterson, ICAC Researcher of the Year 2012
Regents Professor and Head, Plant Genome Mapping Laboratory, University of Georgia, Athens, GA, USA

Background and Rationale

The scientific infrastructure in support of cotton research and improvement took a 'giant leap forward' with the release of the first 'gold-standard' cotton reference genome sequence on 5 January 2012. Later in 2012, two independent publications (Paterson *et al.* 2012; Wang *et al.* 2012) provided initial descriptions of the basic genome of cotton, with one of these also revealing new insights into the genes and processes that have permitted the tetraploid species *Gossypium hirsutum* ('Upland' cotton) and *G. barbadense* (Egyptian, Sea Island, and Pima cotton) to largely supplant the diploids *G. herbaceum* and *G. arboreum* as the providers of the world's leading natural fiber (Paterson *et al.* 2012).

This paper explores some general features of the cotton genome and fundamental messages learned from the sequences, along with new capabilities that the sequence provides to research and development. In particular, the genome sequence provides a means of coalescing many diverse data types, some of which still need to be created for cotton, to gain new understanding from otherwise disparate data. The cotton science infrastructure will quickly grow beyond this 'reference' genome sequence to include much of the diversity among species and genome types in the *Gossypium* genus – and that enriched information has enormous implications for improving the yield and quality of cotton and the sustainability and profitability of its production.

Finally, and perhaps most importantly, the genome sequence is not an ending but a beginning – specifically, a beginning of a new era of research and development using powerful new tools and approaches to identify and manipulate cotton genes of economic importance. While the potential benefits of this era are tangible and large, realizing this potential will require a host of additional enabling tools, technologies, and resources to be developed and creatively deployed, necessitating a new higher level of investment – but offering a new higher level of return on investment.

Which Cotton Genome Should be Sequenced First?

As is widely known, *G. hirsutum* and *G. barbadense*, and the other (wild) tetraploid cotton species, originated 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). A- and D-genome groups are estimated to have diverged from a common ancestor 5-10 million years ago (MYA), then were reunited about 1-2 MYA (Wendel and Cronn 2003) via polyploidization in an A-genome cytoplasm (Wendel 1989; Small and Wendel 1999) following trans-oceanic dispersal to the New World of an A-genome propagule.

Capitalizing on more than a decade of prior research and preparation, in 2005, the worldwide cotton community prioritized the putative D genome progenitor, *G. raimondii* as the first *Gossypium* genotype to be fully sequenced. From first principles, it was preferred to first sequence a homozygous diploid expected to have only two nearly-identical copies of most genes, rather than a tetraploid which would have much more DNA including four copies of most genes comprised of two pairs that were just different enough from one another to be confusing. Although it does not itself produce spinnable fibers, ironically the *G. raimondii*-derived portion of the tetraploid cotton genome (the D_t 'subgenome') accounts for a somewhat larger share of genetic variation in fiber characteristics than the 'A_t' subgenome derived from an ancestor that does produce spinnable fibers (Jiang *et al.* 1998; Rong *et al.* 2007). *Gossypium raimondii* had the important advantages of having only half as much DNA, and in particular much less repetitive 'junk' DNA than the A genome progenitor. A rich history of genetic mapping and molecular analysis had shown *G. raimondii* to have virtually all genes present in the A genome or tetraploid cottons, and that the genes were largely in the same arrangement in the respective genomes. In partial summary, it was clear that information from *G. raimondii* would 'translate' well to cottons of economic importance, while its reduced size and complexity would reduce the cost and time associated with its sequencing and result in an improved outcome.

General Features of Cotton Revealed by the Genome Sequence

(Paterson *et al*. 2012)

Despite having the least-repetitive DNA of the eight *Gossypium* genome types, *G. raimondii* was nonetheless 61% derived from 'transposable elements', often thought of as 'junk DNA' (Paterson *et al.* 2012). One particular class that accounts for the largest share of many flowering plant genomes, long-terminal-repeat retrotransposons (LTRs), likewise account for about 53% of the *G. raimondii* genome.

To identify the genes of *G. raimondii*, computational approaches to recognize common features of genes such as 'start' and 'stop' sites were applied in conjunction with massively parallel sequencing of gene-encoded messenger RNA, to reveal 37,505 genes and 77,267 protein-coding transcripts (some genes encoding multiple transcripts). Remarkably, genes comprise only 44.9 Mb (6%) of the

G. raimondii genome and are largely located in distal chromosomal regions.

One surprise from the genome sequence was that shortly after its divergence from an ancestor shared with cacao (*Theobroma cacao*) at least 60 million years ago, the cotton lineage experienced an abrupt 5–6-fold ploidy increase. It was already well known that flowering plants had experienced polyploidy more frequently than other taxa – indeed, the common ancestor of most if not all eudicot (broad-leaf) plants experienced a genome triplication about 125 million years ago (Paterson *et al.* 2010). However, this was the first (and to date the only) discovery of such a large ploidy increase in such a short time.

The abrupt 5–6-fold ploidy increase together with the additional polyploidy that formed the common ancestor of *G. hirsutum* and *G. barbadense* and the wild tetraploid cottons, rendered cotton among the most complex of flowering plant genomes, only known to be matched by members of the *Brassica* genus. However, in modern cottons, this complex history of genome duplications is reflected in different ways for different genes and gene functional groups. For example, paleopolyploidy increased the complexity of a Malvaceae-specific clade of Myb family transcription factors, perhaps contributing to the differentiation of epidermal cells into fibers rather than the mucilages of other Malvaceae such as cacao. However, cottons pest- and disease-resistance genes experienced rapid turnover and evolved largely after the 5–6-fold ploidy increase.

Another surprise has been the extent to which the two 'subgenomes' of tetraploid cotton have exchanged information with one another since being joined in a common nucleus by polyploidy. Indeed, the vast majority of mutations that differentiate tetraploid cotton from its diploid progenitors involved non-reciprocal DNA exchanges between the A_t and D_t subgenomes, with random mutations contributing little. Curiously, these exchanges have been asymmetric, with more than twice as many D_t-genome alleles 'copied' on the A_t genome than the reciprocal. A tantalizing hypothesis is that the nascent polyploid may have gained fitness from D-genome alleles native to its New World habitat – however this offers no intuitive explanation for the evolution of the superior fibers of polyploids relative to A-genome diploids. Further investigation is in progress.

The Genome Sequence as a Means of Coalescing Diverse Data Types into New Understanding

Specific DNA sequences of 16 or more nucleotides in length are generally specific to single locations in higher eukaryotic genomes, and a host of biological information has been attached to 'sequence-tagged sites' that are generally substantially longer than this. For example, hundreds of 'quantitative trait loci' responsible for variation in economically important traits have been associated with DNA markers that have been sequenced and genetically mapped (e.g. (Rong *et al.* 2007)). Recently, massively parallel sequencing of short nucleic acid molecules has become an effective means of quantitating expression levels of vast numbers of genes under diverse conditions.

The contiguity and specificity afforded by a reference genome sequence provides a powerful means to coalesce diverse data types. By aligning different sets of DNA markers to the reference genome sequence, it is routine to align and compare different QTL mapping studies to identify 'QTL hotspots", regions of the genome that contain QTLs affecting multiple fiber traits more frequently than can be accounted for by chance (Rong *et al.* 2007). Likewise, voluminous gene expression data permits one to map sequence tags to the genome to identify concentrations of genes exhibiting coordinated changes in expression of functionally diverse genes under parallel sets of conditions.

Intersections among diverse data types that are revealed by using the reference sequence may suggest relationships of functional importance. For example, among 48 genes for which expression is up-regulated in domesticated *G. hirsutum* fibers at 10 days post-anthesis, 20 (a 10-fold enrichment relative to random genes) are within QTL hotspot D_t09.2 affecting length, uniformity, and short fiber content. Thirteen (a 15-fold enrichment) are in homoeologous hotspot A_t09 affecting fiber elongation and fineness. Of 45 genes down-regulated in domesticated *G. barbadense* at 20 DPA, 16 (35.6%) map to D_t09.2, and 8 (17.7%) to A_t09. In 79% of cultivated *G. barbadense*, this A_t region (then called chr. 5) has been unconsciously introgressed by plant breeders with *G. hirsutum* DNA, suggesting an important contribution to productivity of *G. barbadense* cultivars(Wang *et al.* 1995). Without the genome sequence to discern that these diverse data types each reveal non-random patterns that are concentrated in the same small region of the genome, they would merely represent interesting independent observations. Having discerned their relationships, we are much closer to identifying the causal gene(s).

A particularly powerful application of the genome sequence is to align the genes and chromosomes of one organism to those of another – for example, alignment to the botanical model *Arabidopsis thaliana* holds particularly great potential for increasing knowledge of cotton gene functions, albeit by analogy (Rong *et al.* 2005). For example, research into the genetic control of cotton fiber development may benefit from rich 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*.

Capturing the Spectrum of Diversity in the *Gossypium* Genus

The genus *Gossypium* occurs naturally throughout tropical and subtropical regions of the world, with at least 45 diploid species ($2n = 26$) that 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.

The diversity present in the two cultivated species, based on a subset of the diversity present in only two of the eight genome types, provides only a small 'sliver' of the naturally-occurring 'solutions' (adaptations) that *Gossypium* species have devised to survive and flourish in the face of often harsh and always fluctuating conditions. Indeed, while the importance of exotic germplasm is widely understood in terms of providing 'obvious' traits such as resistance to new disease strains, rich knowledge of other crops has shown beyond doubt that many alleles from exotic germplasm have 'cryptic' benefits that only become obvious when placed in elite backgrounds. Thus, a high priority is to clothe the reference genome with knowledge of the spectrum of extant diversity in each gene and indeed, each nucleotide. Only with such data can the intrinsic genetic potential of the genus be truly understood the intrinsic genetic potential of the genus, and craft improvement strategies can be crafted that optimally integrate full utilization of this potential with the need for 'extrinsic' (transgenic) solutions.

Much of the additional information needed to characterize the spectrum of extant *Gossypium* diversity will come not from 'gold-standard' reference sequences, which are costly and time-consuming to assemble rigorously, but from new 'resequencing' technologies that have relatively high per-nucleotide error rates but which can be mitigated by sequencing each nucleotide many times to arrive at a consensus that is often correct (Shendure and Aiden 2012).

The first such 'draft sequence' was conducted in *G. raimondii* itself (Wang *et al.* 2012), and comparison to the gold-standard sequence (Paterson *et al.* 2012) is illustrative (Table 1). The draft sequence is highly fragmented – the gold-standard sequence comprising nearly 80% fewer 'scaffolds' (genomic regions that could be assembled into single tracts of sequence at appropriate quality control standards), that were an average of ~8x longer (18.8 versus 2.3 Mb). The longest such scaffold approached the length of an entire chromosome arm in the reference sequence (52.1 Mb), being only about 26% of this in the draft. Virtually all (98.3% of) scaffolds in the reference sequence contained a sufficient number of DNA-based genetic markers to be aligned and oriented to genetically-defined chromosomes from the rich history of prior research in cotton genetics, versus only about half (52.4%) for the draft sequence. Estimates of the number of cotton genes based on the reference and draft assemblies were similar, indicating an important strength of draft sequencing – to quickly and economically capture the subtle differences in 'spelling' (sequences) of genes in different cotton genotypes.

Table 1: Parameters of Different *G. raimondii* Genome Assemblies

	Draft	Reference
Scaffold number	4715	1084
N50 (Megabases)	2.3	18.8
Longest scaffold	12.8	52.1
Anchored and oriented % genome	52.40%	98.30%

An important early application of draft sequencing has been to reveal clues into the early steps in the evolution of spinnable fibers (Paterson *et al.* 2012). From unremarkable hairs found on all *Gossypium* seeds, 'spinnable' fibers, i.e. with ribbon-like structure which allows spinning into yarn, evolved in the A-genome following divergence from the B, E, and F genomes ~5-10 MYA (Senchina *et al.* 2003). To clarify the evolution of spinnable fibers, we sequenced the *G. herbaceum* A and *G. longicalyx* F genomes, which respectively differ from *G. raimondii* by 2,145,177 single nucleotide variations (SNVs) and 477,309 indels; and 3,732,370 SNVs and 630,292 indels (Paterson *et al.* 2012). Across entire genes, 36 *G. herbaceum* - *G. raimondii* and 11 *G. herbaceum* - *G. longicalyx* ortholog pairs show evidence of diversifying selection. A striking example is Gorai.009G035800, a germin-like protein that is differentially expressed between normal and naked-seed cotton mutants during fiber expansion (Kim and Triplett 2004) and between wild and elite *G. barbadense* at 10 days post-anthesis (Paterson *et al.* 2012). We also identified 'striking mutations' of *G. herbaceum* genes since their divergence from *G. longicalyx* and *G. raimondii* (hence correlated with fiber evolution) including 1,090 non-synonymous mutations in 959 genes comprising the most severe 1% of functional impacts inferred using a modified entropy function (Reva *et al.* 2011); 3,525 frameshifts (3,021 genes); 1077 (987) premature stops; 527 (513) splice site mutations; 102 (102) initiation alterations; and 95 (94) extended reading frames. These striking mutations are enriched ($p=2.6x10^{-18}$) within fiber-related 'quantitative trait locus' (QTL) hotspots in A_tD_t tetraploid cottons (Rong *et al.* 2007), suggesting that post-allopolyploidy elaboration of fiber development (Jiang *et al.* 1998) involved recursive changes in A_t and new changes in D_t genes.

In partial summary, an important next step beyond the 'gold-standard' reference sequence will be to catalog the spectrum of diversity among *Gossypium* species, toward cataloguing of the true genetic potential of the genus to provide intrinsic low-cost genetic solutions to challenges that affect the yield, quality of cotton and economic and environmental sustainability of its production. The draft sequence of *G. raimondii* (Wang *et al.* 2012) overestimates the degree to which many such subsequent draft sequences might be assembled, as it benefitted from additional measures that are frequently not economical (and for example were not done in *G. herbaceum* or *G. longicalyx*). However, as information about the basal *Gossypium* genome accumulates, additional sequences will need to reveal smaller and smaller changes, for example single nucleotides in specific genes, and the need for high assembly quality will decline. Indeed, as we begin to sequence elite germplasm and learn about patterns of association ('linkage disequilibrium') of alleles at different loci along the chromosome, we will quickly reach a point such that sequencing of only a small subset of loci is a sufficient proxy to impute the probable genotype across the entire genome. Such technology, already in place in leading crops such as maize (Buckler *et al.* 2009; McMullen *et al.* 2009; Tian *et al.* 2011), is expected to be important in the application of genomic tools to mainstream crop improvement.

How do we Identify the Genes of Economic Importance?

"The greatest challenge facing the cotton community is the conversion of sequence to knowledge …." (Paterson 2007)

With the genome and a much improved understanding of cotton's evolutionary history in hand, and a catalog of the spectrum of cotton's natural diversity imminent, how will we convert these new resources into low-cost genetic solutions to challenges that affect the yield, quality of cotton and economic and environmental sustainability of its production?

It was quickly identified that much cotton sequence is repetitive "junk DNA" –this cannot be dismissed as unimportant, but is relatively low in unique information content. 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 (Myers *et al.* 2011), and even some highly-repetitive regions of a genome 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.

Some cotton sequence will quickly be converted to information based on similarity to known sequences (from *Arabidopsis* in particular). As noted above, the relatively close relationship of cotton and *Arabidopsis*, 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.

However, to understand and manipulate the features that make cotton unique will require a host of new enabling tools, technologies, and resources; in particular targeting genes and regulatory features that are substantially different from those of other organisms. Because the basic gene set for flowering plants has largely been revealed (Paterson *et al.* 2010) by the many genomes now sequenced, a natural priority in cotton functional genomics will be to characterize genes that 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 constraints, 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). 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.* 2001a; Swanson *et al.* 2001b)). Examples of such cotton genes have been noted above, by identifying genes experiencing extensive change in the Gossypium A-genome following divergence from the F and D genomes (Paterson *et al*. 2012).

However, recognition of genes that have evolved rapidly does not by itself reveal their functions. More generally, following two episodes of polyploidy, many cotton genes may now have different (or at least partly different) functions than *Arabidopsis* genes with similar sequences. 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. Indeed, several genomes other than *Arabidopsis* are potentially more informative to cotton in terms of understanding gene evolution and function – because *Arabidopsis* itself has experienced two genome duplications since its divergence from cotton (Bowers *et al.* 2003). The genomes of grape (Lin *et al.* 2011), papaya, and cacao have each remained unduplicated since their divergence from cotton – but have received far less attention to understanding gene functions, and accordingly offer far less information to cotton at present. Nonetheless, the cotton community should remain attuned to new information about these genomes as it may 'translate' especially well to cotton.

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, to permit deductions about gene function based on coordinated expression patterns. Such information is rapidly accumulating thanks to the ability of next-generation sequencing technologies (Shendure and Aiden 2012) to economically and quickly capture information about messenger RNA, as well as DNA.

2) Large-scale sampling of patterns of between-species divergence and within-species diversity of the full set of cotton genes (indeed, preferably entire genomes), as detailed above providing the means to distinguish among genes that show evolutionary patterns such as:

- 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;
- 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;
- 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; Tsai *et al.* 2011) that are likely to become still faster and less costly using next-generation sequencing 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 tetraploid species (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 A and D genome progenitors of the tetraploid.

4) Well-characterized populations of diverse genotypes that are carefully selected to broadly and deeply sample allelic variation within particular gene pools. Such 'diversity panels' (Morris *et al.* 2013) comprised of a few hundred individuals, including careful phenotyping of these individuals, offer the means to utilize historical accumulations of recombination events to associate relatively abundant alleles with phenotypes, providing more precise 'mapping' than can generally be accomplished using conventional QTL mapping (Paterson *et al.* 1988).

5) New genetic populations of two types:

- 'Tiling paths' of NIILs that collectively cover the genome of a target genotype, toward genome wide application (Eshed and Zamir 1995) of the 'substitution mapping' strategy (Paterson *et al.* 1990) providing for fine-scale (1-3 cM) dissection of complex variation into individual components. This approach reveals both predictable alleles and 'cryptic' variation (Gibson and Dworkin 2004) not expected based on the parental phenotypes but often of practical value (Eshed and Zamir 1995; Tanksley *et al.* 1996; Fulton *et al.* 1997; Bernacchi *et al.* 1998a; Bernacchi *et al.* 1998b; Bernacchi *et al.* 1998c; Fridman *et al.* 2004; Chee *et al.* 2005a; Chee *et al.* 2005b; Draye *et al.* 2005; Schauer *et al.* 2006). The precision afforded by the NIILs provides a foundation for establishing causality between phenotypes and specific mutations.
- Nested association mapping populations, that combine the ability to search much diverse germplasm for novel variation with the ability to precisely map the novel variation, guiding the breeder to the specific recombinants needed to separate desirable from undesirable alleles/effects (Yu *et al.* 2005; Yu *et al.* 2008).

Synthesis

In closing, 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* 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, may provide a means to harness this variability in a manner that is minimally subject to public concerns.

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 sequence(s) are laying 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 $200 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 much benefit from Arabidopsis (detailed above), the greater complexity of cotton will require a similar level of investment in its unique genes and features in order to fully realize the potential benefits of its sequencing.

While some ongoing investments in cotton genomics may be in intellectual property of potential commercial value that are appropriately made in the private sector, *many will be in pre-competitive enabling tools that might most efficiently be produced in the public domain or by public-private consortia*. In an industrial crop such as cotton, public-private consortia are particularly attractive, engaging core competencies of public researchers as a 'virtual research and development network' that offers new opportunities for small and medium-sized businesses while also enhancing opportunity for large businesses, by providing new tools, information, and young scientists with the expertise to put these resources to work. Many of the challenges, particularly regarding the spectrum of adaptations that permit cotton to adapt to such a wide range of ecosystems, may best be met by international collaborations.

Acknowledgements

The author thanks many colleagues and collaborators with whom he has partnered in research mentioned herein, and alphabetically Dr. Peng Chee (University of Georgia), Dr. Don Jones (Cotton Incorporated), and Dr. Robert Wright (Texas Tech University) for helpful comments on this manuscript.

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.

Benton, M. J., 1993 *The Fossil Record* 2. Chapman and Hall, New York.

Bernacchi, D., T. Beck-Bunn, D. Emmatty, Y. Eshed, S. Inai *et al.*, 1998a Advanced back-cross QTL analysis of tomato. II. Evaluation of near-isogenic lines carrying single-donor introgressions for desirable wild QTL-alleles derived from Lycopersicon hirsutum and L-pimpinellifolium (vol 97, pg 170, 1998). T*heoretical and Applied Genetics* **97:** 1191-1196.

Bernacchi, D., T. Beck-Bunn, Y. Eshed, S. Inai, J. Lopez *et al.*, 1998b Advanced backcross QTL analysis of tomato. II. Evaluation of near-isogenic lines carrying single-donor introgressions for desirable wild QTL-alleles derived from Lycopersicon hirsutum and L-pimpinellifolium. *Theoretical and Applied Genetics* **97:** 170-180.

Bernacchi, D., T. Beck-Bunn, Y. Eshed, J. Lopez, V. Petiard *et al.*, 1998c Advanced backcross QTL analysis in tomato. I. Identification of QTLs for traits of agronomic importance from Lycopersicon hirsutum. *Theoretical and Applied Genetics* **97:** 381-397.

Bowers, J. E., B. A. Chapman, J. Rong and A. H. Paterson, 2003 Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422:** 433-438.

Buckler, E. S., J. B. Holland, P. J. Bradbury, C. B. Acharya, and P. J. Brown *et al.*, 2009 The Genetic Architecture of Maize Flowering Time. *Science* **325:** 714-718.

Chee, P., X. Draye, C. X. Jiang, L. Decanini, T. A. Delmonte *et al.*, 2005a Molecular dissection of interspecific variation between *Gossypium hirsutum* and *Gossypium barbadense* (cotton) by a backcross-self approach: I. Fiber elongation. *Theoretical and Applied Genetics* **111:** 757-763.

Chee, P. W., X. Draye, C. X. Jiang, L. Decanini, T. A. Delmonte *et al.*, 2005b Molecular dissection of phenotypic variation between *Gossypium hirsutum* and *Gossypium barbadense* (cotton) by a backcross-self approach: III. Fiber length. *Theoretical and Applied Genetics* **111:** 772-781.

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

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

Draye, X., P. Chee, C. X. Jiang, L. Decanini, T. A. Delmonte *et al.*, 2005 Molecular dissection of interspecific variation between *Gossypium hirsutum* and *G-barbadense* (cotton) by a backcross-self approach: II. Fiber fineness. *Theoretical and Applied Genetics* **111:** 764-771.

Eshed, Y., and D. Zamir, 1995 An Introgression Line Population of Lycopersicon Pennellii in the Cultivated Tomato Enables the Identification and Fine Mapping of Yield-Associated OTL. *Genetics* **141:** 1147-1162.

Fridman, E., F. Carrari, Y. S. Liu, A. R. Fernie and D. Zamir, 2004 Zooming in on a quantitative trait for tomato yield using interspecific introgressions. *Science* **305:** 1786-1789.

Fulton, T. M., T. BeckBunn, D. Emmatty, Y. Eshed, J. Lopez *et al.*, 1997 QTL analysis of an advanced backcross of Lycopersicon peruvianum to the cultivated tomato and comparisons with QTLs found in other wild species. *Theoretical and Applied Genetics* **95:** 881-894.

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

Gibson, G., and I. Dworkin, 2004 Uncovering cryptic genetic variation. *Nature Reviews Genetics* **5:** 681-U611.

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

Jiang, C. X., R. J. Wright, K. M. El-Zik and A. H. Paterson, 1998 Polyploid formation created unique avenues for response to selection in *Gossypium* (cotton). Proceedings of the National Academy of Sciences of the United States of America **95:** 4419-4424.

Kim, H. J., and B. A. Triplett, 2004 Cotton fiber germin-like protein. I. Molecular cloning and gene expression. *Planta* **218:** 516-524.

Lin, L., H. Tang, R. O. Compton, C. Lemke, L. K. Rainville *et al.*, 2011 Comparative analysis of *Gossypium* and Vitis genomes indicates genome duplication specific to the *Gossypium* lineage. *Genomics* **97:** 313-320.

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

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.

McMullen, M. D., S. Kresovich, H. S. Villeda, P. Bradbury, H. H. Li *et al.*, 2009 Genetic Properties of the Maize Nested Association Mapping Population. *Science* **325:** 737-740.

Morris, G. P., P. Ramu, S. P. Deshpande, C. T. Hash, T. Shah *et al.*, 2013 Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proceedings of the National Academy of Sciences of the United States of America **110:** 453-458.

Myers, R. M., J. Stamatoyannopoulos, M. Snyder, I. Dunham, R. C. Hardison *et al.*, 2011 A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9:** e1001046.

Nagaki, K., Z. K. Cheng, S. Ouyang, P. B. Talbert, M. Kim *et al.*, 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., 2007 Sequencing the cotton genomes, pp. in *World Cotton Research Conference*. International Cotton Advisory Committee, Lubbock TX.

Paterson, A. H., J. W. Deverna, B. Lanini and S. D. Tanksley, 1990 Fine Mapping of Quantitative Trait Loci Using Selected Overlapping Recombinant Chromosomes, in an Interspecies Cross of Tomato. *Genetics* **124:** 735-742.

Paterson, A. H., M. Freeling, H. Tang and X. Wang, 2010 Insights from the Comparison of Plant Genome Sequences. *Annual Review of Plant Biology* **61:** 349-372.

Paterson, A. H., E. S. Lander, J. D. Hewitt, S. Peterson, S. E. Lincoln *et al.*, 1988 Resolution of Quantitative Traits Into Mendelian Factors By Using a Complete Linkage Map of Restriction Fragment Length Polymorphisms. *Nature* **335:** 721-726.

Paterson, A. H., J. Wendel, F., H. Gundlach, H. Guo, J. Jenkins *et al.*, 2012 Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492:** 423-427.

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

Reva, B., Y. Antipin and C. Sander, 2011 Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* **39:** e118.

Rong, J., J. E. Bowers, S. R. Schulze, V. N. Waghmare, C. J. Rogers *et al.*, 2005 Comparative genomics of *Gossypium* and *Arabidopsis*: Unraveling the consequences of both ancient and recent polyploidy. *Genome Research* **15:** 1198-1210.

Rong, J., E. A. Feltus, V. N. Waghmare, G. J. Pierce, P. W. Chee *et al.*, 2007 Meta-analysis of polyploid cotton QTL shows unequal contributions of subgenomes to a complex network of genes and gene clusters implicated in lint fiber development. *Genetics* **176:** 2577-2588.

Schauer, N., Y. Semel, U. Roessner, A. Gur, I. Balbo *et al.*, 2006 Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology* **24:** 447-454.

Senchina, D. S., I. Alvarez, R. C. Cronn, B. Liu, J. K. Rong *et al.*, 2003 Rate variation among nuclear genes and the age of polyploidy in *Gossypium*. *Molecular Biology and Evolution* **20:** 633-643.

Shendure, J., and E. L. Aiden, 2012 The expanding scope of DNA sequencing. *Nature Biotechnology* **30:** 1084-1094.

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.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner and C. F. Aquadro, 2001a 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.

Swanson, W. J., Z. H. Zhang, M. F. Wolfner and C. F. Aquadro, 2001b 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.

Tanksley, S. D., S. Grandillo, T. M. Fulton, D. Zamir, Y. Eshed *et al.*, 1996 Advanced backcross QTL analysis in a cross between an elite processing line of tomato and its wild relative *L-pimpinellifolium*. *Theoretical and Applied Genetics* **92:** 213-224.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun *et al.*, 2011 Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature Genetics* **43:** 159-162.

Till, B. J., S. H. Reynolds, E. A. Greene, C. A. Codomo, L. C. Enns *et al.*, 2003 Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Researc*h **13:** 524-530.

Tsai, H., T. Howell, R. Nitcher, V. Missirian, B. Watson *et al.*, 2011 Discovery of Rare Mutations in Populations: TILLING by Sequencing. *Plant Physiology* **156:** 1257-1268.

Wang, G. L., J. M. Dong and A. H. Paterson, 1995 The Distribution of *Gossypium-hirsutum* Chromatin in *Gossypium-barbadense* Germ Plasm - Molecular Analysis of Introgressive Plant-Breeding. *Theoretical and Applied Genetics* **91:** 1153-1161.

Wang, K., Z. Wang, F. Li, W. Ye, J. Wang *et al.*, 2012 The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genetics* **44:** 1098-1103.

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, pp. 139-186 in *Advances in Agronomy, Vol 78*.

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.

Yu, J., G. Pressoir, W. Briggs, I. V. Bi, M. Yamasaki *et al.*, 2005 A Unified Mixed-Model Method for Association Mapping that Accounts for Multiple Levels of Relatedness. *Nature Genetics* **38:** 203-208.

Yu, J. M., J. B. Holland, M. D. McMullen and E. S. Buckler, 2008 Genetic design and statistical power of nested association mapping in maize. *Genetics* **178:** 539-551.