Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres.

To cite this version:

HAL Id: hal-01594869
https://hal.archives-ouvertes.fr/hal-01594869
Submitted on 26 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres

Andrew H. Paterson1, Jonathan F. Wendel2, Heidrun Gundlach3, Hui Guo1, Jerry Jenkins4,5, Dianchun Jin6, Danny Llewellyn7, Kurtis C. Showmaker8, Shengqiang Shu4, Joshua Udall9, Mi-jong Yoo10, Robert Byers11, Wei Chen6, Adi Doron-Faigenboim12, Mary V. Duke13, Lei Gong2, Jane Grimwood4,5, Corrinne Grover2, Kara Grupp1, Guanjing Hu2, Tae-Hee Lee1, Jingping Li1, Lifeng Lin1, Frank Bedon14, John E. Bowers1, Curt L. Brubaker7,17, Peng W. Chee18, Sayan Das1, Alan R. Gingle1, Candace H. Haigler19, Mary V. Duke11, Lei Gong2, Jane Grimwood4,5, Corrinne Grover2, Kara Grupp2, Guanjing Hu2, Tae-ho Lee1, Jingping Li1, Lifeng Lin1, Frank Bedon14, John E. Bowers1, Curt L. Brubaker7,17, Peng W. Chee18, Sayan Das1, Alan R. Gingle1, Candace H. Haigler19, David Harker9, Lucia V. Hoffmann20, Ran Hovav9, Donald C. Jones3, Cornelia Lembreck, Shadi Mансoor22, Mehboob ur Rahman22, Lisa N. Rainville1, Aditi Rambani3, Umesh K. Reddy23, Jun-Kang Rong3, Yehoshua Saranga24, brian E. Scheffler25, Jodi A. Scheffler3, David M. Stelly26, Barbara A. Triplett26, Allen Van Deynze25, Maite F. S. Vaslin27, Vijay N. Waghmare28, Sally A. Walford7, Robert J. Wright29, Essam A. Zaki30, Tianzhen Zhang31, Elizabeth S. Dennis32, Klaus F. X. Mayer3, Daniel G. Peterson6, Daniel S. Rokhsar4, Xiyin Wang1,6 & Jeremy Schmutz3,5

Polyploidy often confers emergent properties, such as the higher fibre productivity and quality of tetraploid cottons than diploid cottons bred for the same environments1. Here we show that an abrupt five- to sixfold ploidy increase approximately 60 million years (Myr) ago, and allopolyploidy reuniting divergent Gossypium genomes approximately 1–2 Myr ago, conferred about 30–36-fold duplication of ancestral angiosperm (flowering plant) genes in elite cottons (Gossypium hirsutum and Gossypium barbadense), genetic complexity equalled only by Brassica among sequenced angiosperms. Nascent fibre evolution, before allopolyploidy, is elucidated by comparison of spinnable-fibred Gossypium herbaceum A and non-spinnable Gossypium longicalyx F genomes to one another and the outgroup D genome of non-spinnable Gossypium raimondii. The sequence of a G. hirsutum A Dt (in which ‘t’ indicates tetraploid) cultivar reveals many non-reciprocal DNA exchanges between subgenomes that may have contributed to phenotypic innovation and/or other emergent properties such as ecological adaptation by polyploids. Most DNA-level novelty in G. hirsutum recombines alleles from the D-genome progenitor native to its New World habitat and the Old World A-genome progenitor in which spinnable fibre evolved. Coordinated expression changes in proximal groups of functionally distinct genes, including a nuclear mitochondrial DNA block, may account for clusters of cotton-fibre quantitative trait loci affecting diverse traits. Opportunities abound for dissecting emergent properties of other polyploids, particularly angiosperms, by comparison to diploid progenitors and outgroups.

The Gossypium genus is ideal for investigating emergent consequences of polyploidy. A-genome diploids native to Africa and Mexican D-genome diploids diverged ~5–10 Myr ago4. They were reunited ~1–2 Myr ago by trans-continental dispersal of a maternal A-genome propogae resembling G. herbaceum to the New World2, hybridization with a native D-genome species resembling G. raimondii, and chromosome doubling (Fig. 1). The nascent A D polyploidy spread throughout the American tropics and subtropics, diverging into at least five species; two of these species (G. hirsutum and G. barbadense) were independently domesticated to spawn one of the world’s largest industries (textiles) and become a major oilseed.

New insight into Gossypium biology is offered by a genome sequence of G. raimondii Ulbr. (chromosome number, 13) with ~8× longer scaffold N50 (18.8 versus 2.3 megabases (Mb)) compared with a draft5, and oriented to 98.3% (versus 52.4%) of the genome (Supplementary Table 1.3a). Across 13 pseudomolecules totalling 737.8 Mb, ~350 Mb (47%) of euchromatin span a gene-rich 2,059 centimorgan (cM), and ~390 Mb (53%) of heterochromatin span a repeat-rich 186 cM (Supplementary Discussion, sections 1.5 and 2.1). Despite having the least-repetitive DNA of the eight Gossypium genome types, G. raimondii is 61% transposable-element-derived (Supplementary Table 2.1). Long-terminal-repeat retrotransposons (LTRs) account for 53% of G. raimondii, but only 3% of LTR base pairs derive from 2,345 full-length elements. The 37,505 genes and 77,267 protein-coding transcripts annotated (Supplementary Table 2.3 and http://www.phytozome.com) comprise 44.9 Mb (6%) of the genome, largely in distal chromosomal regions (Supplementary Discussion, section 2.1).

Shortly after its divergence from an ancestor shared with Theobroma cacao at least 60 Myr ago, the cotton lineage experienced an abrupt five- to sixfold ploidy increase. Individual grape chromosome segments resembling ancestral eudicot genome structure, or corresponding chromosomal segments, generally have five
(infrequently six) best-matching *G. raimondii* regions and secondary matches resulting from pan-eudicot hexaploidy7,8 (Fig. 2 and Supplementary Table 3.1). Paralogous genes tracing to this five- to sixfold ploidy increase show a single peak of synonymous nucleotide-substitution (*K*<sub>s</sub>) values, suggesting either one, or multiple closely spaced, event(s) (Supplementary Fig. 3.5). Pairwise cytological similarity among A-genome chromosomes9 suggests the most recent event was a duplication.

Paleopolyploidy may have accelerated cotton mutation rates: for 7,021 co-linearity-supported gene triplets, *K*<sub>s</sub> rates and non-synonymous nucleotide-substitution (*K*<sub>a</sub>) rates were, respectively, 19% and 15% larger for cotton–grape than cacao–grape comparisons (Supplementary Table 3.2). Adjusted for this acceleration (Supplementary Fig. 3.5), the cotton ploidy increase occurred about halfway between the pan-eudicot hexaploidy (<125 Myr ago)10 and the present, near the low end of an estimated range of 57–70 Myr ago11.

Paleopolyploidy increased the complexity of a Malvaceae-specific clade of Myb family transcription factors, perhaps contributing to the differentiation of epidermal cells into fibres rather than the mucilages of other Malvaceae. Among 204 R2R3, 8 R1R2R3 and 194 heterogeneous Myb transcription factors in *G. raimondii* (Supplementary Table 3.5), subgroup 9 has six members known only in Malvaceae (Fig. 3a), comprising a possible ‘fibre clade’ distinct from the *Arabidopsis thaliana* GL1-like subgroup 15 involved in trichome and root hair initiation and development12. Expressed predominantly in early fibre development, elite cultivated tetraploid cottons have higher expression of five (50%) of ten subgroup 9 genes compared with wild (undomesticated) tetraploids (Fig. 3a and Supplementary Table 3.3). Some subgroup 9 genes are also active in leaves, hypocotyls and cotyledons (Supplementary Fig. 3.8), consistent with specialization for different types of epidermal cell differentiation such as production of a ‘pulpy layer’ secreted from the teguments surrounding cacao seeds, and mucilages in other Malvaceae fruit (*Abelmoschus* (okra), *Cola* (kola)) and roots (*Althaea* (marshmallow)).

Cotton growers were early adopters of integrated pest management13 strategies to deploy intrinsic defences conferred by pest- and disease-resistance genes that evolved largely after the 5–6-fold ploidy increase.

---

**Figure 1** | Evolution of spinnable cotton fibres. Paleohexaploidy in a eudicot ancestor (red, yellow and blue lines) formed a genome resembling that of grape (bottom right). Shortly after divergence from cacao (bottom left), the *Gossypium* lineage experienced a five- to sixfold ploidy increase. Spinnable fibre evolved in the A genome after its divergence from the F genome, and was further elaborated after the merger of A and D genomes ~1–2 Myr ago, forming the common ancestor of *G. hirsutum* (Upland) and *G. barbadense* (Egyptian, Sea Island and Pima) cottons.

**Figure 2** | Syntenic relationships among grape, cacao and cotton. a, Macro-synteny connecting blocks of >30 genes (grey lines). Highlighted regions (pink and red) trace to a common ancestor before the pan-eudicot hexaploidy, with the *Gossypium* lineage five- to sixfold ploidy increase forming multiple derived regions. Inferred duplication depth in cotton varies (top). b, Micro-synteny of grape chromosome (Chr) 3, cacao chromosome 2 and five cotton chromosomes. Rectangles represent predicted genes, with connecting grey lines showing co-linear relationships. An example (1 grape, 1 cocoa, 5 cotton) is highlighted in red.
NBS-encoding genes. The most NBS-rich (21%) region of *T. cacao*, on chromosome 7, corresponds to parts of *G. raimondii* chromosome triplets 2/10/13 and 7/9/4. In total, 27% and 25% of 294 mapped *G. raimondii* NBS genes are on these parts of chromosomes 7 and 9, often clustered in otherwise gene-poor surroundings (Supplementary Fig. 2.2). Most NBS clusters are species and chromosome specific (Fig. 3b and Supplementary Table 3.7), indicating rapid turnover and/or concerted evolution after cotton paleopolyploidy. In total, 230 (76.7%) NBS-encoding genes have experienced striking mutations (as detailed below) in the A genome since A–F divergence, reflecting an ongoing plant–pathogen ‘arms race’ (Supplementary Table 3.8).

Changes in gene expression during domestication have contributed to the deposition of >90% cellulose in cotton fibres, single-celled models for studying cell wall and cellulose biogenesis15. *G. raimondii* has at least 15 cellulose synthase (CESA) sequences required for cellulose synthesis16 (Supplementary Table 3.3), with four single-gene *Arabidopsis* clades having three (CESA3, required in expanding primary walls) or two (CESA4, CESA7 and CESA8, each required in the thickening of secondary walls) clade members in *G. raimondii*16. *G. raimondii* has at least 35 cellulose-synthesise-like (CSL) genes required for synthesis of cell wall matrix polysaccharides that surround cellulose microfibrils16 (Supplementary Table 3.4), including one family (CSL) absent in *Arabidopsis*16. Elite tetraploids have higher expression than wild cottons in 6 (40%) of 15 CESA genes and 12 (34%) of 35 CSL genes (Supplementary Table 5.3).

A total of 364 *G. raimondii* microRNA precursors from 28 conserved and 181 novel families (Supplementary Table 3.12), are predicted regulators of 859 genes enriched for molecule binding factors, catalytic enzymes, transporters and transcription factors (Supplementary Fig. 3.11, 12). Four conserved and 35 novel mRNAs were specifically expressed in *G. hirsutum* fibres, respectively targeting 53 and 318 genes, most with homology to proteins involved in fibre development (Supplementary Table 3.14, 15). Among 183,690 short interfering RNAs (siRNAs) found, 33,348 (18.15%) were on chromosome 13 (Supplementary Fig. 3.12), a vast enrichment. Small RNA17–19 biogenesis proteins include 13 argonaute, 6 dicer-like (DCL) and 5 RNA-dependent RNA polymerase orthologues (Supplementary Table 3.16). *G. raimondii* seems to be the first eudicot with two DCL3 genes and two genes encoding RNA polymerase IVa (Supplementary Table 3.16), perhaps relating to control of its abundant retrotransposons.

From unremarkable hairs found on all *Gossypium* seeds, ‘spinnable’ fibres (fibres with a ribbon-like structure that allows for spinning into yarn) evolved in the A genome after divergence from the B, E and F genomes ~5–10 Myr ago4 (Fig. 1). To clarify the evolution of spinnable fibres, 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.

Specific genes are implicated in initial fibre evolution by both whole-genome and individual-nucleotide analyses. Across entire genomes, 36 *G. herbaceum*–*G. raimondii* and 11 *G. herbaceum*–*G. longicalyx* orthologue pairs show evidence of diversifying selection (ω > 1, P < 0.05) (Supplementary Table 4.1). A notable example, with *G. herbaceum*–*G. raimondii* ω > 9, is Gorai.009G035800, a germin-like protein that is differentially expressed between normal and naked-seed cottons (Supplementary Fig. 5.1, 2), indicating rapid turnover and/or concerted evolution after cotton paleopolyploidy. In total, 230 (76.7%) NBS-encoding genes have experienced striking mutations (as detailed below) in the A genome since A–F divergence, reflecting an ongoing plant–pathogen ‘arms race’ (Supplementary Table 3.8).

Figure 3 | Paleo-evolution of cotton gene families. a. Myb subgroup 9 (ref. 12) originated from a gene on the progenitor of cacao chromosome 2 that formed two adjacent copies after Malvales–Brassicales divergence and then triplicated in cotton, with subsequent loss of one chromosome 8 and two chromosomes 12 paralogues. One extant parologue traces to pan-eudicot hexaploidy, Tc04 g009420, and reduplicated in cotton (Gorai.012G052500.1 and Gorai.011G122800.1) and *Arabidopsis* (At5g15310 and At2g12000). The other, Tc01 g009420, has reduplicated in cotton (Gorai.004G157600.1 and Gorai.003G156700.1). Asterisk indicates increased gene expression in elite cottons, indicating lineage-specific expansion. *G. raimondii* chromosome triplets 2/10/13 and 7/9/4. Cotton chromosome 7 NBSs form a single branch, indicating lineage-specific expansion. *G. raimondii* chromosome 7 and 13 NBSs form distinct branches, indicating cluster/tandem duplication (gene numbers also reflect physical proximity of genes to one another).

A total of 300 (0.8%) *G. raimondii* genes encode nucleotide-binding site (NBS) domains (Supplementary Table 3.6), largely of coiled-coil (CC)-NBS and CC-NBS-leucine rich repeat subgroups (165, 55%). Like cereals14, after paleopolyploidy *G. raimondii* evolved clusters of new
distribution (Supplementary Fig. 2.2) but are over-represented in genes coding for cell-wall-associated, kinase or nucleotide-binding proteins (Supplementary Table 4.5).

Striking mutations in the A-genome lineage are enriched ($P = 2.6 \times 10^{-15}$; Supplementary Discussion, section 4.4) within fibre-related quantitative trait locus (QTL) hotspots in A0D, tetraploid cottons22, suggesting that post-polyploidy elaboration of fibre development involved recursive changes in A and new changes in D genes. Striking A-genome mutations have orthologues in 1,051 D1 and 951 A1 fibre QTL hotspots. Likewise, sequencing of G. hirsutum cultivar Acala Maxxa revealed 495 striking mutations in 391 genes, with 83 (21.2%) in D1 fibre QTL hotspots and 73 (18.7%) in A1 hotspots (Supplementary Table 4.6).

QTL hotspots affecting multiple fibre traits22 may reflect coordinated changes in expression of functionally diverse cotton genes. A total of 671 (1.79%) genes with >100 reads per million reads were differentially expressed in fibres from wild versus domesticated G. hirsutum (mostly at 10 DPA) and/or G. barbadense (mostly at 20 DPA) (Supplementary Table 5.3). Among 48 genes upregulated in domesticated G. hirsutum at 10 DPA, 20 (42%) are among 1,582 (4.2%) genes within QTL hotspot D09.2 (ref. 22) affecting length, uniformity, and short-fibre content, with 13 (27%) out of 677 (1.8%) genes in homoelogous hotspot A09 affecting fibre elongation and fineness. Out of 45 genes downregulated in domesticated G. barbadense at 20 DPA, 16 (35.6%) map to Dt09.2, and 8 (17.7%) to At09. In 79% of cultivated G. barbadense, this A region (which was then thought to be on chromosome 5, and is now known to be on chromosome 9) has been unconsciously introgressed by plant breeders with G. hirsutum DNA, suggesting an important contribution to productivity of G. barbadense cultivars25.

A putative nuclear mitochondrial DNA (NUMT) sequence block24 has an intriguing relationship with fibre improvement. A G. raimondii chromosome 1 region includes many genes closely resembling mitochondrial homologues ($K_2$ $P$ < 0.22; Supplementary Table 4.7a). NUMT genes experienced a coordinated change in expression associated with G. barbadense domestication. The 105 (0.2%) genes upregulated in 10 DPA fibre of wild (versus elite) tetraploid G. barbadense (Supplementary Table 5.3) include 30 (37%; $P < 0.001$) of the 81 NUMT genes, including 8 NADH dehydrogenase and 4 cytochrome-c-related genes. All are within the QTL hotspot D01 that affects fibre fineness, length, and uniformity27, suggesting a fibre-specific change in electron transfer in G. barbadense domestication.

Emergent features of polyembryos may be related to processes that render them no longer the sum of their progenitors and permit them to explore transgressive phenotypic innovations. Despite the A-genome origin of spinnable fibres, after ~2 Myr of co-habitation in tetraploid nuclei most A and D homeologues are now expressed in fibres at similar levels (Supplementary Table 5.4). Such convergence is not ubiquitous: gene families involved in the synthesis of seed oil show strong A bias in wild G. hirsutum and its sister G. tomentosum, but strong D bias in an improved G. hirsutum (Supplementary Table 5.6).

Recruitment of D1 genome genes into tetraploid fibre development may have involved non-reciprocal DNA exchanges from A1 genes. In the ~40% of Acala Maxxa A1 and D1 genes that differ in sequence from their diploid progenitors (Fig. 4), most mutations are convergent, with A1 genes converted to the D1 state at more than twice the rate (25%) as the reciprocal (10.6%). Known to occur between cereal paralogous genes27–29, Dt-biased allele conversion may have contributed to the origin of Gossypium gossypioides, sister to G. raimondii and the only D-genome cotton containing many otherwise A-genome-specific repetitive DNAs27–29. Dt-biased allele conversion may have contributed to slightly greater protein-coding nucleotide diversity in the A1 compared with the D1-genome (Supplementary Table 5.7).

Whereas the G. raimondii reference sequence and G. hirsutum short-read sequences reveal much about tetraploid cotton genome structure and polyploid evolution, high-contiguity sequencing of polyploids may elucidate still-cryptic features. Tetraploid cotton sequencing appears feasible: among six pairs of A1 and D1 BAC clones, the most similar pair shows 99.1% shared D1-D and 97.6% A1-D content (Supplementary Table 6.1), sufficient divergence to de-convolute shotgun sequence to the correct subgenome. Increased knowledge of molecular diversity is a foundation for integrating genomics with ecological and field-level knowledge of Gossypium species and their diverse adaptations to warm arid ecosystems on six continents.

**METHODS SUMMARY**

**Sequencing.** Reads were collected from Applied Biosystems 3730xl, Roche 454 XLR and Illumina Genome Analyzer IIx machines at the Joint Genome Institute

---

**Figure 4** Allesic changes between A- and D-genome diploid progenitors and the A1 and D1 subgenomes of G. hirsutum cultivar Acala Maxxa.
Annotation. PERTRAN software was used to construct transcript assemblies from ~1.1 billion pairs of G. raimondii paired-end Illumina RNA-seq reads, 250 million G. raimondii single end reads, and 150 million G. hirsutum single end reads. PASA™ was used to build transcript assemblies from 454 and Sanger resources (Supplementary Table 2.3). Loci were determined by transcript assembly and amino acid sequences (Fig. 3) were used to guide coding sequence alignments. Phylogenetic trees were constructed by bootstrap neighbour-joining with a Kimura 2-parameter model using ClustalW2, assessing internal nodes with 1,000 replicates.

Full Methods and any associated references are available in the online version of the paper.

Received 17 August; accepted 21 November 2012.


Supplementary Information is available in the online version of the paper.

Acknowledgements The work, conducted by the US Department of Energy’s Genomic Institute, is supported by the Office of Science of the US Department of Energy under Contract no. DE-AC02-05CH11231. The authors appreciate financial support from the US National Science Foundation (DBI 08-27680 to A.H.P., J.F.W., R.J.W., D.G.P.; DE-AC02-05CH11231. The authors appreciate financial support from the US National Science Foundation (DBI 08-27680 to A.H.P., J.F.W., R.J.W., D.G.P.; DE-AC02-05CH11231. The authors appreciate financial support from the US National Science Foundation (DBI 08-27680 to A.H.P., J.F.W., R.J.W., D.G.P.; DE-AC02-05CH11231. The authors appreciate financial support from the US National Science Foundation (DBI 08-27680 to A.H.P., J.F.W., R.J.W., D.G.P.; DE-AC02-05CH11231.

©2012 Macmillan Publishers Limited. All rights reserved
METHODS

Sequencing. Reads were collected with standard protocols (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html) on Applied Biosystems 3730XL, Roche 454 XLR and Illumina Genome Analyzer (GA)IIx machines at the US Department of Energy Joint Genome Institute. Linear 454 data included standard XLR (47 runs, 16,868 Gb) and pre-release FLX+ data (5 runs, 3,262 Gb). Eight paired 454, 3–4 kilobase (kb) average insert size and one paired 12-kb average insert size were sequenced on standard XLR (23 runs, 5,931 Gb). One standard 400-base pair (bp) fragment library was sequenced at 2 × 150 (7 channels, 41.9 Gb) on an Illumina GAIIx. One 2.5-kb average insert size (405,024 reads, 286.1 Mb), one 6.5-kb average insert size library (374,125 reads, 263.0 Mb), six fosmid libraries (1,222,643 reads, 702.1 Mb) of 34–39-kb insert size, and two BAC libraries (107,520 reads, 77.5 Mb) of 98-kb and 115-kb (73,728 reads, 48.8 Mb) average insert size were sequenced on both ends for a total of 2,183,240 Sanger reads of 1.38 Gb of high-quality bases. FLX+ data were collected at the Roche Service Center. BAC end sequence (BES) was collected using standard protocols at the HudsonAlpha Institute.

Genome assembly and construction of pseudomolecule chromosomes. Organellar reads were removed by screening against mitochondria, chloroplast and ribosomal DNA. Any Roche 454 linear read <200 bp was discarded. Roche 454 paired reads in which either was shorter than 50 bp were discarded. An additional de-duplication step was applied to the 454 paired libraries that identifies and retains only one copy of each PCR duplicate. All remaining 454 reads were compared against a full Illumina GA2x run and any insertion/deletions in the 454 end sequence (BES) was collected using standard protocols at the HudsonAlpha Institute. The aim of the completeness analysis was to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. cDNAs were aligned to the assembly using BLAT31 (parameters: -t = dnum -q = rna –extendThroughN). The best scored predictions were improved by PASA. Improvement includes adding untranslated regions, splicing correction, and adding alternative transcripts. PASA-improved gene model peptides were subject to peptide-homology analysis to above-mentioned proteomes in order to obtain Cacore and peptide coverage. Cacore is a peptide BLASTP score ratio mutual best hit BLASTP score and peptide coverage is highest percentage of peptide aligned to the best of homologues. PASA-improved transcripts were selected on the basis of Cacore, peptide coverage, EST coverage and its coding sequence (CDS) overlapping with repeats. The transcripts were selected if their Cacore was larger than or equal to 0.5 and peptide coverage larger than or equal to 0.5, or if it had EST coverage, but its CDS overlapping with repeats was less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cacore needed to be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose peptide was more than 30% in Pfam transposable element domains were removed. The final gene set had 37,505 protein-coding genes and 77,267 protein-coding transcripts.

85% EST coverage were retained. The screened alignments and/or EXONERATE alignments of peptides from A. thaliana, cacao, rice, soybean, grape and poplar peptides to repeat-soft-masked D5 genome using RepeatMasker (http://www.repeatmasker.org) with up to 2,000-bp extensions on both ends, unless extending into another locus on the same strand. Gene models were predicted by homology-based predictors, FGENSEH +41, FGENSEH_EST (similar to FGENSEH+; EST as splice site and intron input instead of peptide/translated open-reading frames) and GenomeScan42. The best scored predictions for each locus are selected using multiple positive factors including EST and peptide support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding untranslated regions, splicing correction, and adding alternative transcripts. PASA-improved gene model peptides were subject to peptide-homology analysis to above-mentioned proteomes in order to obtain Cacore and peptide coverage. Cacore is a peptide BLASTP score ratio mutual best hit BLASTP score and peptide coverage is highest percentage of peptide aligned to the best of homologues. PASA-improved transcripts were selected on the basis of Cacore, peptide coverage, EST coverage and its coding sequence (CDS) overlapping with repeats. The transcripts were selected if their Cacore was larger than or equal to 0.5 and peptide coverage larger than or equal to 0.5, or if it had EST coverage, but its CDS overlapping with repeats was less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cacore needed to be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam analysis and gene models whose peptide was more than 30% in Pfam transposable element domains were removed. The final gene set had 37,505 protein-coding genes and 77,267 protein-coding transcripts.


©2012 Macmillan Publishers Limited. All rights reserved