Supporting Information

SI Material and Methods

Nucleic acid sequence alignments.

Three new parameters were recently defined in Salse *et al.* (1) to increase the stringency and significance of BLAST sequence alignment by parsing BLASTN results and rebuilding HSPs (High Scoring Pairs) or pairwise sequence alignments. The first parameter, AL (Aligned length), corresponds to the sum of all HSP lengths. The second, CIP (cumulative identity percentage) corresponds to the cumulative percent of sequence identity obtained for all the HSPs (CIP= $\sum$ nb ID by HSP / AL) x 100). The third parameter, CALP is the cumulative alignment length percentage. It represents the sum of the HSP lengths (AL) for all the HSPs divided by the length of the query sequence (CALP = AL / Query length). The CIP and CALP parameters allow the identification of the best alignment, *i.e.* the highest cumulative percentage of identity in the longest cumulative length, taking into account all HSPs obtained for any pairwise alignment. These parameters were applied to all the BLAST alignments that were performed in the present study.

Wheat, barley, rice, maize sorghum sequence databases.

The 6,426 **wheat** ESTs representing 15,569 non redundant loci that were assigned to deletion bins by Qi *et al.* (2) were downloaded from the GrainGene web site (http://wheat.pw.usda.gov/). Since some ESTs may correspond to non overlapping 3' and 5' ends of the same cDNA sequence, a unigene set of 5,003 mapped wheat ESTs was established (1) and used in the current analysis. The **barley** data correspond to 1,015 mapped ESTs (3) the 2,755 mapped ESTs downloaded from the 'HarvEST' web site [http://www.harvest-web.org]. A unigene set of 3,423 non-redundant barley mapped ESTs was constructed from these data set following the same method used for constructing the Wheat unigene set (with 95% CIP, 85% CALP as thresholds criteria). The sequences of the 12 **rice** pseudomolecules (build 5) were downloaded from the 'tigr' web site (www.tigr.org) and the 41,046 genes identified by the annotation were used for the analysis. The sequences of the 10 **sorghum** chromosomes were downloaded from the 'phytozome' web site (www. phytozome.net) and the 34,008 gene models identified by the annotation were used for the analysis. The sequences of the 10 **maize** chromosomes were downloaded from the 'genome arizona' web site (www.genome.arizona.edu/) and the 60,284 gene models (STS, Sequence Tag Sites) identified by the annotation were used for the analysis as described in Wei *et al.* (4).

Identification of duplicated and syntenic regions.

Duplications were identified through intra-genomic sequence comparisons. Paralogous pairs were identified with 70% CIP and 70% CALP cut off values after the alignment of the annotated genes available for each genomes, (1). Syntenic regions were identified through inter-genomic sequence comparisons. In this case, orthologous pairs were identified using 60% CIP and 70% CALP cut off values after the alignment of the annotated genes available for each genome. (1). Paralogous pairs identified during the duplication analyses as well as orthologous gene pairs identified from the colinearity analysis were statistically validated with the CloseUp software (5) as described in Salse *et al.* (1). Two criteria were established to define paralogous or orthologous regions taking into account the physical size (Size), number of annotated genes (Gnumber) and number of orthologous or paralogous couples (Cnumber) in the regions. The two criteria are the Density Ratio (DR) and the Cluster Ratio (CR) (3): DR = [(Size 1+ Size 2) / (2x Cnumber)] x 100 and CR = [(2x Cnumber) / Gnumber 1 + Gnumber 2] x 100. The Density Ratio (DR) considers the number of links between two regions (duplicated or syntenic) as a function of the size of the considered blocks. The Cluster Ratio (CR) considers the number of links between two regions (duplicated or syntenic) as a function of the number of annotated genes available in the considered blocks. Statistically significant collinear or duplicated regions are associated with the highest Density Ratio and Cluster Length values. The remaining collinear or duplicated regions were considered as artificial, *i.e.* obtained at random considering the number of links between two regions (duplicated or syntenic) characterized by a physical size and number of annotated genes available (1).

Graphical display

Colinearity was visualized graphically using the Genome Pixelizer software (http://niblrrs.ucdavis.edu/GenomePixelizer/GenomePixelizer_Welcome.html). Duplications were visualized using the Circos software (http://mkweb.bcgsc.ca/circos/). Three data sheets comprising (1) marker information (name, position on chromosome, chromosome number), (2) link information (data obtained for the statistical identification of paralogous or orthologous sequence pairs) and (3) graphical information (number, size of chromosomes) were provided to the two softwares.

CIP/CALP distribution for paralogous and ortholougous gene pairs.

We classically performed the sequence divergence as well as speciation event analysis based on the rate of nonsynonymous (*Ka*) vs. synonymous (*Ks*) substitutions calculated with MEGA-3 (6). The average substitution rate (r) of 6.5×10^{-9} substitutions per synonymous site per year for grasses is classically used to calibrate the ages of the considered gene (7, 8). The time (*T*) since gene insertion is then classically estimated using the formula $T = Ks / r$ (8). We

have recently performed this strategy to study the sequence divergence between paralogs and orthologs for cereals species and subspecies (1, 9, 10). However, as it has been already reported in cereal (1) and vertebrate genome (11) analyses, this strategy is not the most appropriate when looking and comparing sequence divergence for old speciation events such as the 7 shared cereal duplications dating back to more than 50-70 My in the current study. As a consequence, here, sequence similarity has been accurately investigated based on sequence alignment identity [see Jaillon *et al.* (11) for teleost fish *Tetraodon nigroviridis* duplication analysis]. In the current analysis we then investigated the sequence divergence among the paralogous and orthologous gene pairs based on the average CIP/CALP value ranking from 70 to 100%. For example, an average CIP/CALP value of 100% corresponds to two orthologous or paralogous sequences that have 100% identity over 100% of the entire sequence length.

Ancestor genome reconstruction.

The dot plot (12) cluster approach was used to support ancestral genome reconstruction by taking into account the syntenic and duplication relationships that were characterized independently between the 5 cereal genomes. First, a series of dot plots was created for each of the orthologous cereal chromosome segments using rice as a template (12 dot plots). Then, the dot plots were assembled into 5 non-redundant groups of duplicates based on information from the 7 shared duplications. Dot plots were represented using the SpotFire software, TIBCO (http://spotfire.tibco.com). The analysis resulted in the definition of 22 blocks within the 5 ancestral proto-chromosomes delineated by the paleoduplication boundaries. Within each chromosome blocks, the gene number was estimated as the non-redundant number of genes (conserved between at least two genomes) present on the dot plot diagonal. The size of the ancestral genome was then estimated on the basis of the longest shared sequence conserved between at least two genomes.

SI Material and Methods references

1- Salse J, *et al.* (2008) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 20(1):11-24.

2- Qi LL, *et al.* (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* 168:701-712.

3- Stein N, *et al.* (2007) A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor Appl Genet* 114:823-839.

4- Wei F, *et al.* (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. *PLoS Genet* 3(7):e123.

5- Hampson SE, Gaut BS, Baldi P (2005). Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* 21:1339-48.

6- Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 5:150-163.

7- Gaut BS, Morton BR, McCaig BC, Clegg MT (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. *Proc Natl Acad Sci USA* 93(19):10274-9.

8- SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20(1):43-5.

9- Chantret N, *et al.* (2005) Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *Plant Cell* 17(4):1033-45.

10- Salse J, *et al.* (2008) New insights into the origin of the B genome of hexaploid wheat: Evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides. *BMC Genomics* 9(1):555.

11- Jaillon O, *et al.* (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* 431(7011):946-957.

12- Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167:1–10.

Table S1: Orthologous relationships and shared duplications between the rice, wheat, barley, sorghum and maize genomes.

Colinear rice, wheat, barley, sorghum and maize chromosomes are displayed on the same lines in the table. The number of orthologs identified between rice and the 4 other genomes is indicated in brackets beside each chromosome. Intra-genomic Duplications (ID) are indicated with black brackets while duplications identified with the Dual Synteny (DS) approach are indicated with red brackets. Dotted black brackets illustrated the chromosome fusions identified in maize and sorghum. Dotted and black squares represent lineage specific events in the Triticeae (translocation between chromosomes 4-5) and maize (tetraploidisation event) genomes, respectively. The 5 ancestral chromosomes (A5, A7, A11, A8, A4) defined by the shared duplications identified in the four genomes are displayed on the left side with the same color code as used in the figures in the text. The 2 duplications shared between the Triticeae and rice chromosomes that do not share a common ancestry but are found on orthologous chromosomes in both species are indicated with black double arrows. The numbers of orthologs, paralogs shared and lineage specific duplications are indicated at the end of the table for the 5 genomes.

Unigene	41046	5003	3423	34008	60284
Species	Rice	Wheat	Barley	Sorghum	Maize
A5	r5(3298) r1(5313)	w1 (67) w3 (176)	b1(23) b3(4) b3(47) b1(3)	s9(446) s1(16) s3(1024) s9(11)	m6(187) m8(133) m10(20) m3(417) m8(296) m1(19)
A7	r3(4559) r7(3270) r10(2404)	w4 (112) w5 (26) w2 (39) w1 (17)	b4(43) b5(20) b7(5) b3(3) b2(14) b7(3) b1(10) b5(4) b4(3)	s1(1049) s6(9) s2(9) s2(448) s6(11) s1(263) s3(8)	m1(515) m9(158) m5(119) m7(13) m2(10) m7(209) m2(113) m5(7) m1(125) m5(48) m9(22) m2(5)
A11	r11(2936) r12(2610)	w4 (15) w5 (30)	(b4(2)) (b1(2)) b5(11) b4(1)	s5(145) s8(18) s8(180) s3(12)	m4(68) m2(32) m1(7) m3(49) m10(45) m1(43) m5(10)
A8	r8(2905) r9(2399)	w7 (45) w5 (11)	b7(11) b6(6) b5(22) b7(1)	s7(379) s1(16) s2(340) s1(8)	m1(99) m4(98) m10(41) m6(24) m2(9) m7(160) m2(87) m5(7)
A4	r2(4319) r4(3613) r6(3420)	w6 (108) w2 (89) w7 (92)	b6(40) b2(3) b2(40) b4(1) b7(26) b5(2)	s4(821) s1(16) s6(595) s10(9) s10(457) s4(15)	m5(346) m4(267) m9(9) m2(274) m10(170) m5(10) m9(141) m6(121) m5(47) m4(18)
Orthologs		827	309	6147	4454
Paralogs	383	102	38	390	3469
Shared Duplications	7	7	5	5	5
Lineage-specific duplications	3-12; 4-8; 4-10	3-5; 1-7; 3-7	1-5; 2-5; 3-7; 5-6	1-3; 1-10; 3-4	15 Tetraploidy + 1-2; 3-4
Total duplications	10	10	9	8	17

Figure S1 : Enlarged resolution of Figure 1A.

Schematic representation of the 11,737 orthologs identified between the rice chromosomes (r1 to r12) used as a reference, and the barley (b1 to b7), wheat (w1 to w7), sorghum (s1 to s10), and maize (m1 to m10) chromosomes. Each line represents an orthologous gene. The 5 different colors used to represent the blocks reflect the origin from the 5 ancestral proto-chromosomes (1).

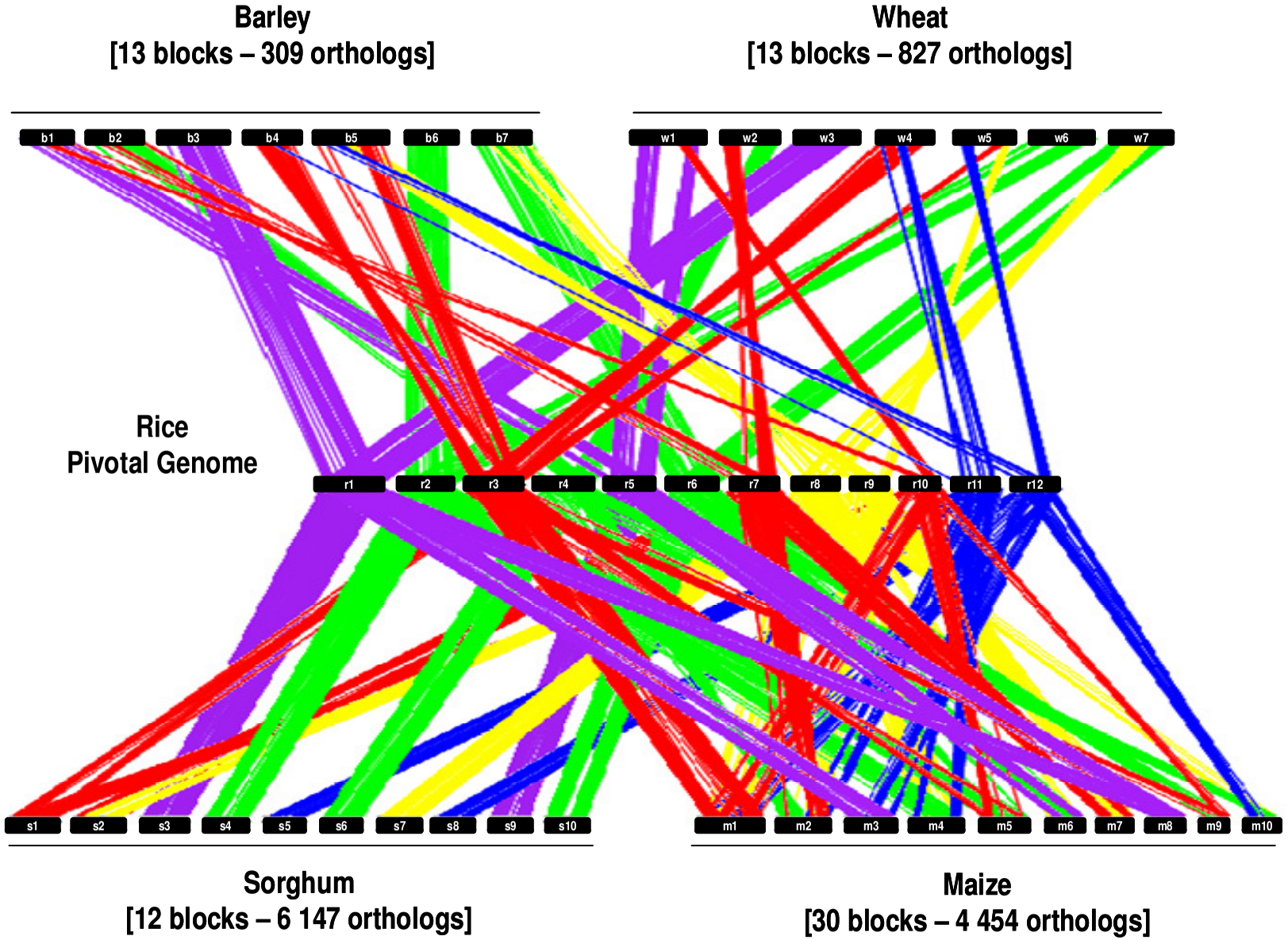

Barley
[13 blocks – 309 orthologs]
Wheat
[13 blocks – 827 orthologs]
b1 b2 b3 b4 b5 b6 b7
w1 w2 w3 w4 w5 w6 w7
Rice
Pivotal Genome
r1 r2 r3 r4 r5 r6 r7 r8 r9 r10 r11 r12
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10
m1 m2 m3 m4 m5 m6 m7 m8 m9 m10
Sorghum
[12 blocks – 6 147 orthologs]
Maize
[30 blocks – 4 454 orthologs]

Figure S2 : Enlarged resolution of the article Figure 1B.

Schematic representation of the 4,382 paralogous pairs identified within the rice (r1 to r12), barley (b1 to b7), wheat (w1 to w7), sorghum (s1 to s10), and maize (m1 to m10) genomes. Each line represents a duplicated gene. The 5 different colors used to represent the blocks reflect the origin from the 5 ancestral proto-chromosomes.

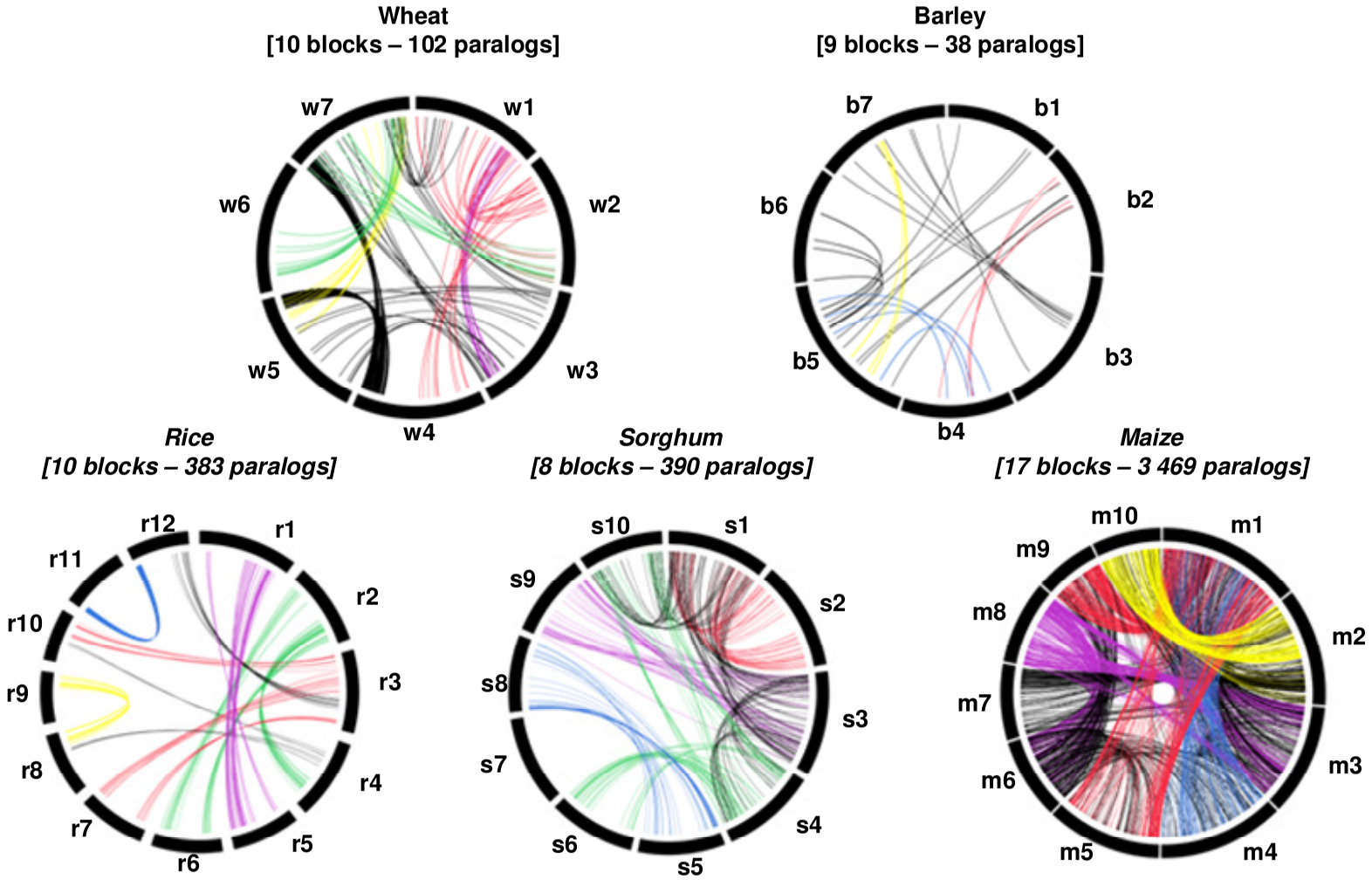

Wheat
[10 blocks – 102 paralogs]
w7
w1
w6
w2
w5
w3
w4
Barley
[9 blocks – 38 paralogs]
b7
b1
b6
b2
b5
b3
b4
Rice
[10 blocks – 383 paralogs]
r12
r1
r11
r2
r10
r3
r9
r8
r4
r7
r5
r6
Sorghum
[8 blocks – 390 paralogs]
s10
s1
s9
s2
s8
s3
s7
s4
s6
s5
Maize
[17 blocks – 3 469 paralogs]
m10
m1
m9
m8
m2
m7
m3
m6
m5
m4

Figure S3 : Enlarged resolution of the article Figure 2A.

The synteny between rice (r), considered as the reference sequence (vertical), and maize (m) or sorghum (s) (horizontal) is shown as 12 dot-plots. The rice/sorghum synteny is depicted with 12 red dot-plots. The synteny between rice and maize is displayed as 12 blue and green dot-plots (reflecting the tetraploid nature of the maize genome). The seven paleo-duplications are indicated by grey blocks within the dot plots. 22 ancestral proto-chromosome blocks (from A5S3 to A7L2) were identified with respect to paleo-duplication boundaries shown with grey blocks on A5, A11, A8, A4, A7 harboring different colored blocks reflecting the origin from the 5 ancestral proto-chromosomes. The number of conserved genes (cumulative diagonal dot-plots) as well as the physical size (cumulative coding sequence length) of each proto-chromosome block is shown in parenthesis on the right end side of the figure.

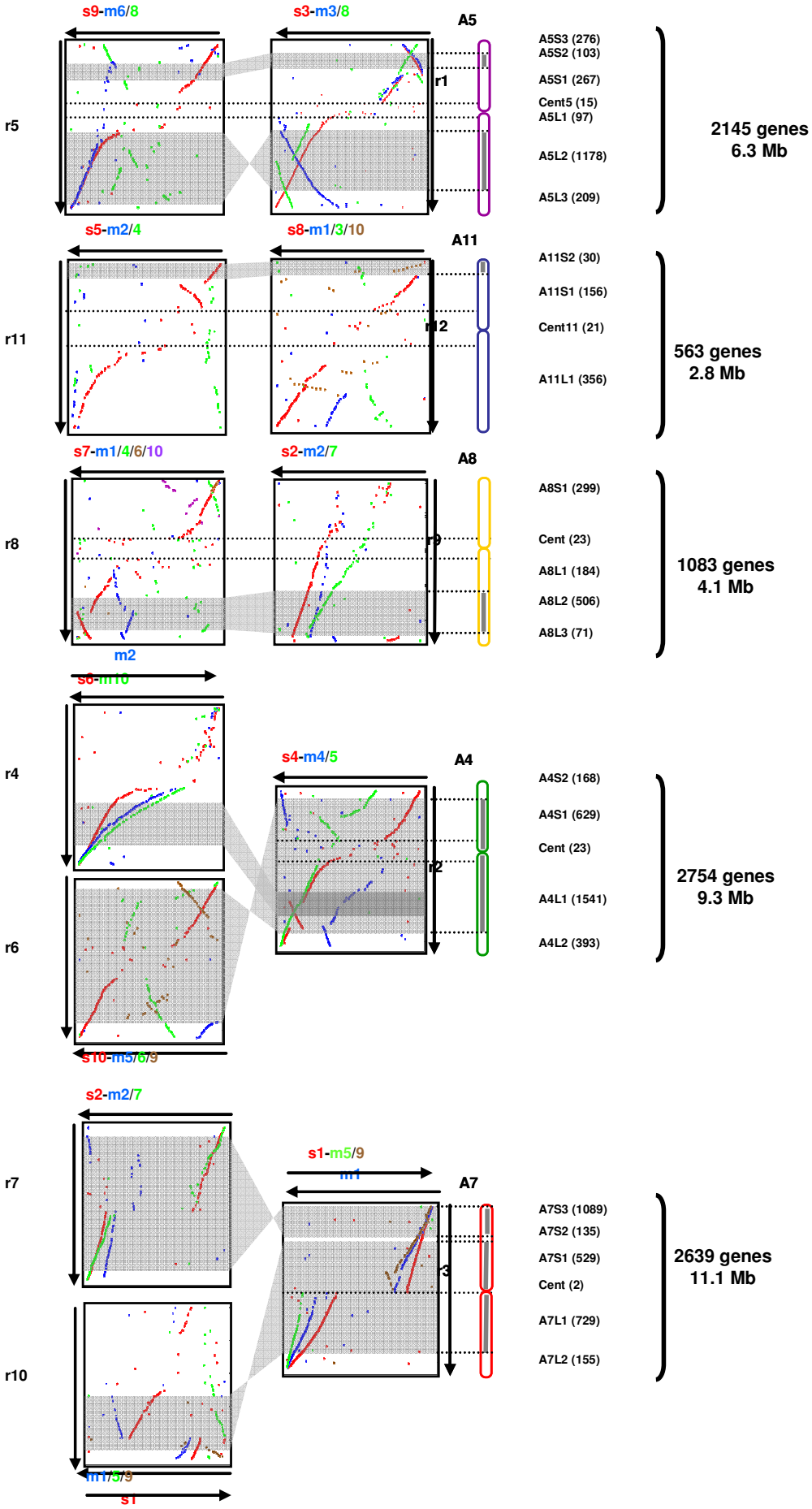
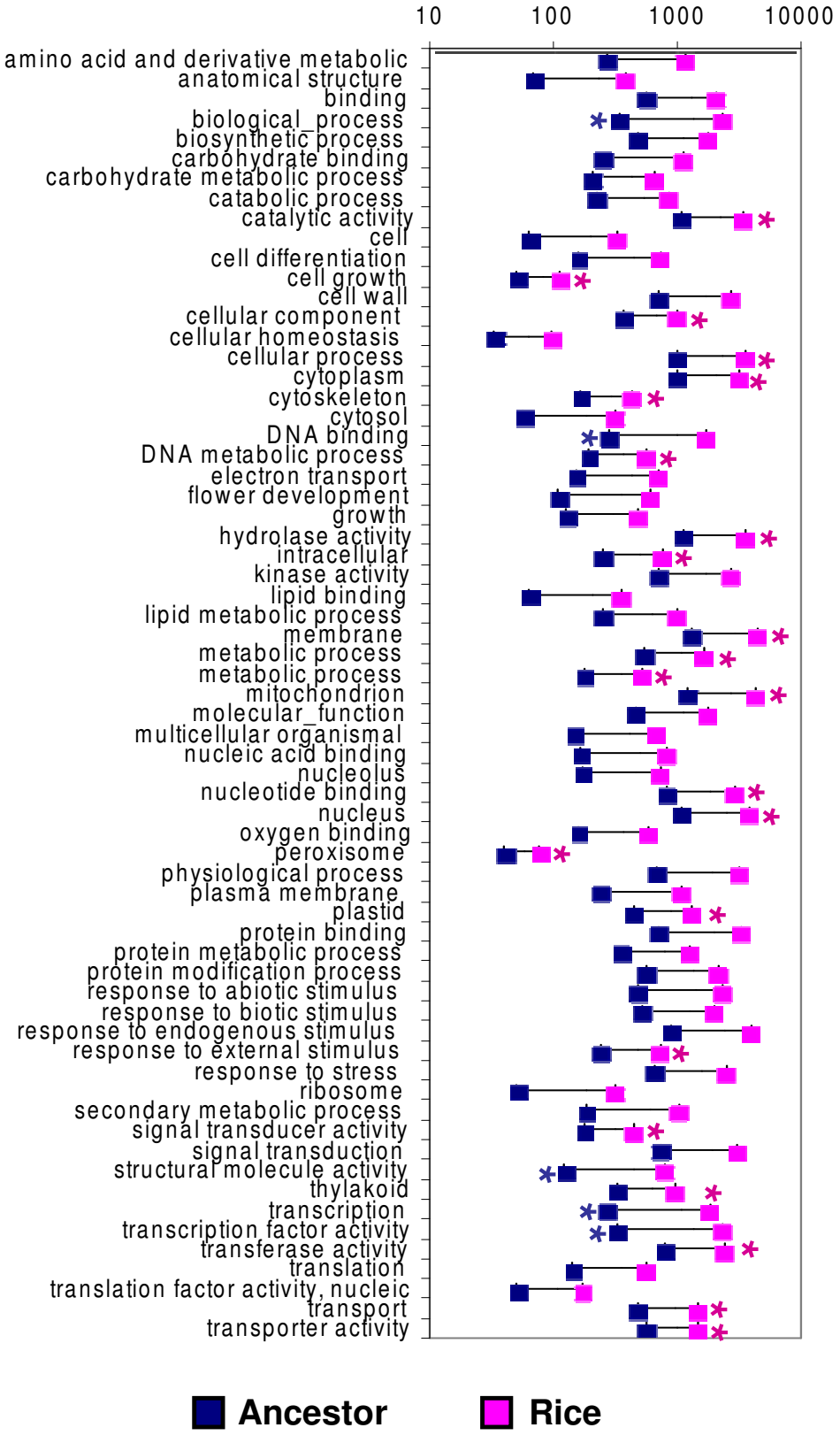

s9-m6/8
s3-m3/8
A5
r5
r1
A5S3 (276)
A5S2 (103)
A5S1 (267)
Cent5 (15)
A5L1 (97)
A5L2 (1178)
A5L3 (209)
2145 genes
6.3 Mb
s5-m2/4
s8-m1/3/10
A11
r11
r12
A11S2 (30)
A11S1 (156)
Cent11 (21)
A11L1 (356)
563 genes
2.8 Mb
s7-m1/4/6/10
s2-m2/7
A8
r8
r9
A8S1 (299)
Cent (23)
A8L1 (184)
A8L2 (506)
A8L3 (71)
1083 genes
4.1 Mb
m2
s6-m10
s4-m4/5
A4
r4
r2
A4S2 (168)
A4S1 (629)
Cent (23)
A4L1 (1541)
A4L2 (393)
r6
s10-m5/6/9
2754 genes
9.3 Mb
s2-m2/7
s1-m5/9
m1
A7
r7
r3
A7S3 (1089)
A7S2 (135)
A7S1 (529)
Cent (2)
A7L1 (729)
A7L2 (155)
r10
m1/3/9
s1
2639 genes
11.1 Mb

Figure S4 : Evolution of the cereal proto-chromosome gene functions.

Graphical representation of the number of genes (in log scale) identified for the 65 Gene Ontology (GO provided on the left end side) classes established for the 41,046 rice genes (http://gnn.tigr.org/tdb/e2k1/osa1/GO.retrieval.shtml). The pink distribution represents the GO distribution for the rice genome and the blue distribution the number of genes for the corresponding GO classes observed for the ancestral genome gene content, *i.e.* genes that are conserved between at least two of the cereal genomes analysed in this work. Blue and pink stars indicate the GO classes that are statistically underrepresented in the ancestor and in the actual cereal genomes, respectively.


10
100
1000
10000
amino acid and derivative metabolic
anatomical structure
binding
biological_process
biosynthetic process
carbohydrate binding
carbohydrate metabolic process
catabolic process
catalytic activity
cell
cell differentiation
cell growth
cell wall
cellular component
cellular homeostasis
cellular process
cytoplasm
cytoskeleton
cytosol
DNA binding
DNA metabolic process
electron transport
flower development
growth
hydrolase activity
intracellular
kinase activity
lipid binding
lipid metabolic process
membrane
metabolic process
metabolic process
mitochondrion
molecular_function
multicellular organismal
nucleic acid binding
nucleolus
nucleotide binding
nucleus
oxygen binding
peroxisome
physiological process
plasma membrane
plastid
protein binding
protein metabolic process
protein modification process
response to abiotic stimulus
response to biotic stimulus
response to endogenous stimulus
response to external stimulus
response to stress
ribosome
secondary metabolic process
signal transducer activity
signal transduction
structural molecule activity
thylakoid
transcription
transcription factor activity
transferase activity
translation
translation factor activity, nucleic
transport
transporter activity
Ancestor
Rice