Funct Integr Genomics
DOI 10.1007/s10142-009-0129-8

ORIGINAL PAPER

Genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection

Umar Masood Quraishi · Michael Abrouk · Stéphanie Bolot · Caroline Pont · Mickael Throude · Nicolas Guilhot · Carole Confolent · Fernanda Bortolini · Sébastien Praud · Alain Murigneux · Gilles Charmet · Jerome Salse

Received: 1 April 2009 / Revised: 26 May 2009 / Accepted: 31 May 2009

Abstract Recent updates in comparative genomics among cereals have provided the opportunity to identify conserved orthologous set (COS) DNA sequences for cross-genome map-based cloning of candidate genes underpinning quantitative traits. New tools are described that are applicable to any cereal genome of interest, namely, alignment criterion for orthologous couples identification, as well as the Intron Spanning Marker software to automatically select intron-spanning primer pairs. In order to test the software, it was applied to the bread wheat genome, and 695 COS markers were assigned to 1,535 wheat loci (on average one marker/2.6 cM) based on 827 robust rice–wheat orthologs. Furthermore, 31 of the 695 COS markers were selected to fine map a pentosan viscosity quantitative trait loci (QTL) on wheat chromosome 7A. Among the 31 COS markers, 14 (45%) were polymorphic between the parental lines and 12 were mapped within the QTL confidence interval with one marker every 0.6 cM defining candidate genes among the rice orthologous region.

Keywords Synteny · SNP · Viscosity · QTL · Candidate genes · Comparative genomics · Genome mapping · Quantitative trait loci

Umar Masood Quraishi and Michael Abrouk contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10142-009-0129-8) contains supplementary material, which is available to authorized users.

U. M. Quraishi · M. Abrouk · S. Bolot · C. Pont · N. Guilhot · C. Confolent · F. Bortolini · G. Charmet · J. Salse (✉)
Génétique, Diversité et Ecophysiologie des Céréales (GDEC),
UMR 1095 INRA/Université Blaise Pascal,
Domaine de Crouelle, 234 avenue du Brézet,
63100 Clermont-Ferrand, France
e-mail: jsalse@clermont.inra.fr

M. Throude · S. Praud
Biogemma,
Site ULICE, ZAC les portes de Riom-BP173,
63204 Riom, France

A. Murigneux
Limagrain Verneuil Holding,
ZAC les portes de Riom-BP 173,
63204 Riom, France

Introduction

We have recently developed a combined strategy devoted to an accurate analysis of the syntenic relationships between cereals requiring stringent consideration of two main criteria: (1) expert sequence alignment, taking into account both alignment quality and length (CIP for cumulative identity percentage and, CALP for cumulative alignment length percentage) and (2) statistical validation criteria (using gene density parameters such as for example in the CloseUp software, Hampson et al. 2003, 2005). We applied this strategy in comparing wheat, rice, maize, and sorghum to provide a detailed view of the synteny between these genomes (Salse et al. 2004, 2008a; for review Bolot et al. 2009). We assessed precise orthologous relationships as well as intra-specific duplications to reveal shared duplicated segments, i.e. conserved at orthologous positions. Thus, increasing the resolution of map-based comparative studies as well as applying statistical tests to the sequence-based comparisons has revealed additional chromosomal rearrangements (such as shared or lineage-specific duplications) between rice and the other grass (i.e. for example maize, Wei et al. 2007; and wheat, Salse et al. 2008a)

Published online: 03 July 2009

genomes compared with those reported previously (that were based on RFLP analysis). Furthermore, this approach provided a more complex picture of the orthologous relationships between these genomes, meaning that marker selection based on synteny analysis requires an accurate investigation of orthologous relationships to derive genetic markers (Bolot et al. 2009). Without such a stringent approach, synteny-based markers could frequently be designed based on artefactual orthologous relationships, leading to loss of time in mapping markers that are in fact selected in non-syntenic regions.

SNPs (i.e. single nucleotide polymorphisms) are the most frequently occurring polymorphism between genotypes in different plant species. As an example, in rice, Feltus et al. (2004) aligned the 12 chromosome pseudomolecules from the two sequenced rice subspecies (indica and japonica) to identify 384,431 SNPs for an overall average of 1.06 SNP/kb. Moreover Yu et al. (2005) clearly established a higher SNP rate in introns (i.e. 6.07 SNP/kb) than in coding regions (i.e. 3.00 SNP/kb) of rice genes. As a consequence, SNPs are clearly the richest source of polymorphism in cereals. As in wheat, or in many unsequenced cereal genomes (such as barley, oat, millet, rye...), the largest sources of SNP markers derived from EST libraries, the development of intron-spanning EST-SNP markers in these genomes would be a valuable resource for genetic experiments such as genetic mapping, linkage disequilibrium, and association studies as well as marker-assisted breeding. In-depth and accurate analysis of the synteny between mapped ESTs with rice annotated genes, considered as reference genome, allows the identification of intron-spanning conserved orthologous set (COS) markers, defining a putative position in rice as well as a precise gene structure (intron/exon boundaries) to design primers. COS markers have been intensively studied in the dicots, taking advantage of the Arabidopsis genome sequence release in 2000 (Fulton et al. 2002; Gupta and Rustgi 2004; Panjabi et al. 2008) but poorly investigated in cereals. Wu et al. (2006) reported the identification of up to 2,869 COS between the euasterid plant species and *Arabidopsis thaliana*. Even if the development of COS markers in bread wheat (*Triticum aestivum*, hexaploid AABBDD genome with $2n=42$) (Salse et al. 2008b) represents a major challenge, it will represent a powerful tool to accelerate either quantitative trait loci (QTL) cloning, genetic diversity analysis or even marker-assisted breeding programs.

Bread-making quality QTL are of major interest in wheat such as for the pentosan viscosity trait. The main physical property of soluble pentosans (mostly arabino-xylans, hereafter AX) lies in their ability to form viscous aqueous solutions (Martinant et al. 1999). These viscous properties of pentosans have a significant influence on the behavior of processed cereal grain. When pentosans are added to dough, its development time increases, as does its viscosity. Pentosan viscosity also has an important digestive and metabolic influence on monogastric animals. Indeed, it reduces the digestibility and water absorption, which leads to economic losses and sanitary problems in animal production, particularly with poultry. On the other hand, soluble pentosans in human diet also prevent lipid absorption, and thus may help to reduce blood lipids such as bad cholesterol. Therefore they are considered as dietary fibers with increasing evidence of protective effects on cardiovascular diseases and even on gut cancer. The trait is of considerable importance both for human and animal needs.

In this article we report an accurate method to identify and use COS markers based on (1) a re-analysis of the syntenic relationships between rice and wheat genomes using a combination of new sequence alignment criteria and statistical analysis (2) the intron spanning markers (ISM) software to automatically select sequence-specific primer pairs to amplify introns as a source of polymorphism. COS markers have been used to fine map a QTL for the grain flour pentosan viscosity trait in a Renan × Recital RIL population. Finally, this approach and tools can be applied for any cereal of interest by using rice as a pivotal sequenced genome.

Materials and methods

Orthologous couples identification: rice–wheat synteny analysis

Nucleic acid sequence alignments

Orthologous (i.e. alignment between genomes) relationships refer to the highest conserved sequence (best nucleotide percentage alignment) on the longest alignment. We have developed in Salse et al. (2008a) two parameters: cumulative identity percentage CIP (i.e. cumulative identity percentage obtained for all the High Scoring Pairs, HSP found), cumulative alignment length percentage CALP (i.e. cumulative of all the HSP length obtained for all the HSP found reported to the query sequence length).

AL (i.e. alignment length) = $\sum$ HSP length.
CIP (i.e. cumulative identity percentage) = ($\sum$ nb ID by HSP/AL) × 100.
CALP (i.e. cumulative alignment length percentage) = AL/Query length.

Any BLAST alignment performed in our analysis will refer to CIP and CALP values of respectively 70% and 60%. These two parameters aim at selecting the best alignment in

terms of identity percentage of the longest cumulative alignment.

Wheat and rice sequence databases

Five thousand and three EST-contigs derived from the 6,426 wheat ESTs that were assigned to deletion bins by Qi et al. (2004) identified in Salse et al. (2008a) were considered. Since the EST-contigs are not ordered within deletion bins and because deletions bins size is variable between the homoeologous A, B, and D genomes, comparative analysis with the three genomes would be prone to mapping errors. To simplify the analysis and ensure the correct identification of orthologs between wheat and rice, the analysis was performed on the seven wheat homoeologous chromosome groups. Thus, any EST-contig that was mapped in a deletion bin on two or three of the homoeologous chromosomes of a given chromosome group was considered with a unique position on a single consensus chromosome per group. The consensus position was calculated as follow: [(BSt+((BSi / (NG+1)) × NGR))/CSi] × 100 with BSt=bin start coordinate, BSi=bin size (in percentage of remaining chromosome, Qi et al. 2004), NG=number of gene assigned to a bin, NGR=gene rank within a bin and CSi=total chromosome size (200=100 for the long arm and 100 for the short arm according to Qi et al. (2004)). EST-contigs that mapped on different homoeologous chromosome arms were ignored.

The sequences of the 12 rice pseudomolecules (build 5; 372,077,801 bp) were downloaded from the TIGR web site (ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0/) and the 41,046 genes identified by the annotation were used for the analysis of the rice–wheat synteny.

Statistical validation

Putative orthologous gene pairs identified from the colinearity analysis between rice and wheat were validated through a CloseUp analysis (http://www.igb.uci.edu/servers/cgss.html). The statistical analysis is based on a permutation test (Monte Carlo) performed through a randomization of the initial data to select conserved blocks of genes that are not obtained by chance, i.e. that are not obtained in any random distribution performed. A collinear region was considered as statically significant when the gene density within the region was validated by CloseUp permutation tests with the following parameters: a density ratio of 2, a cluster length of 20 and a match number of 5 were used.

COS markers selection: the ISM software

Wheat EST-contigs exon structures were identified through rice/wheat sequence alignments, as conserved HSPs correspond to exons. Precise exon/intron boundaries identified (i.e. HSP boundaries) for any considered wheat mapped ESTs associated with a rice ortholog was considered in order to define two values, i.e. Ir and Er. Ir (for Included region) and Er (for Excluded region) are associated with any Intron position (Ii) within Wheat EST-contigs aligned with a rice sequence: Er=(I_i−25) to (I_i+25). This region corresponds to 50 nucleotides centered on the predicted intron position within the wheat EST sequence. Ir=(I_{i-1}+10) to (I_{i+1}−10). This region correponds to the two exons spanning the predicted intron position within the wheat EST sequence. The precise sequence region corresponding to Ir-Er is provided to Primer 3 package to select primer pairs on exons for intron amplification with the following parameters suitable for detection on Applied Biosystems (ABI) capillary sequencer: (1) Primer size (20 to 25 mer as default parameters), (2) Amplicon size (between 250–800 bp as default parameters), (3) Tm (between 57–63 as default parameters), (4) GC clamp (equal to 2, i.e. a G or C at the 5' extremity as default parameters), (5) GC percentage (50% as default parameters).

Pentosan viscosity QTL mapping

Material

The plant population studied has been described previously in Groos et al. (2003). It consisted of 194 F_7 RILs obtained by single-seed descent from the cross between 'Renan' and 'Récital'. 'Récital' is more productive while 'Renan' has a higher grain protein content, pentosan viscosity, and thousand-kernel weight. The population was sown in autumn, 1998, and harvested in summer, 1999, in six locations in France, Châlons-en-Champagne, Chartainvilliers, Clermont-Ferrand, Le Moulon, Mons, and Rennes. The experimental field design consisted of a randomized trial with two replications, divided into blocks. The parental lines were replicated in every block to control field heterogeneity. Each plot measured between 5.4 m^2 and 7.5 m^2 depending on location, and plants were grown under normal field conditions.

Phenotyping

Baking quality was measured through two parameters: pentosan viscosity and dough strenght (W) for all plots as described in Groos et al. (2003). Viscosity of aqueous extracts is measured using an automated capillary viscometer (AVS 310 Schott Gerate, Germany).

Genotyping on applied biosystems sequencer

SSCP protocol COS primers were synthesized with 5' extensions in order to facilitate the labeling procedure at

low cost: forward primer with the CACGACGTTG-TAAAACGAC sequence extension and reverse primer with the CAGGAAACAGCTAT GACC sequence extension. Polymerase chain reaction (PCR) fragments were produced in two steps. In a total volume of 15 µl, genomic DNA (30 ng) was first amplified with the following PCR mix: 10 mM Tris–HCL, 3.1 mM MgCl2, 50 mM KCl, 0.001% gelatine pH 8.3, 5% glycerol, 400 µM dNTP, 0.4 µM forward and reverse primers, and 0.2 U Taq polymerase (Qiagen). This PCR product was diluted (1/10) and re-amplified with the same PCR mix including 0.2 µM of each labeled primers (6-FAM and NED, Applied Biosystems) in a final volume of 15 µl. Two microliters of the PCR product was then diluted (1/10) and pooled with 0.2 µl of 900 bp MegaBase ET900-R Size Standard (GE Healthcare), 0.2 µl of 0.3 N NaOH and 9 µl HI-Di Formamide (Applied Biosystems). Fragments were separated by capillary electrophoresis on ABI3100 (Applied Biosystems) in 50 min with a 36 cm capillary. The running polymer consists in 1× of running buffer, 5% Genscan Polymer (Applied Biosystems), and 10% glycerol. Samples were denatured during 2 min at 95 C and 10 min in ice. The sample buffer consists in 1× of running buffer and 10% glycerol. After denaturing, the samples were injected at 2.5 kV during 50 s and separated at 18 C, 25 C, and 35 C at 15 kV. Data were analyzed using GeneMapper 3.7 software.

Sequencing protocol Amplified PCR products were cloned with pGem®-T Easy Vector System II (Promega) according to the manufacturer's instructions and sequenced in both strands with the forward and reverse COS primers. Sequencing reactions were performed using BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) with 10 ng of DNA and 0.25 µl of Big Dye. Fragments migration were performed on 3130XL Genetic Analyser (Applied Biosytems) with polymer POP7 and UltraSeq36_POP7_1 default run module settings.

Mapping and QTL analysis

Linkage analyses were performed using Mapmaker/exp 3.06 (Lander and Green 1987). The Kosambi mapping function (Kosambi 1944) was applied to transform recombination frequencies into additive distances in cM. Linkage groups were assigned to chromosomes via comparisons to the ITMI reference maps using microsatellite loci (Röder et al. 1998a, b). QTL analysis was performed using a Splus 'home made' program described by Groos et al. (2003). First, one-way analyses of variance were used to detect significant differences between genotypic classes for each marker. Significant markers were used as candidates in a multiple regression model, in order to select a subset of non-redundant markers for further use as covariates. Then, on every linkage group in which at least one marker was found significant, the marker-regression method (Kearsey and Hyne 1994) was carried out to locate the QTLs more precisely and estimate their additive effects. This program allows the detection of two QTLs on the same chromosome using a two-dimensional scanning of the chromosome (Hyne and Kearsey 1995). The 95% confidence intervals of the QTL locations and additive effects were established by bootstrapping (Visscher et al. 1996) using 1,000 replicates for the one-QTL model and 400 for the two-QTL models.

Results

Identification of 827 rice–wheat orthologs

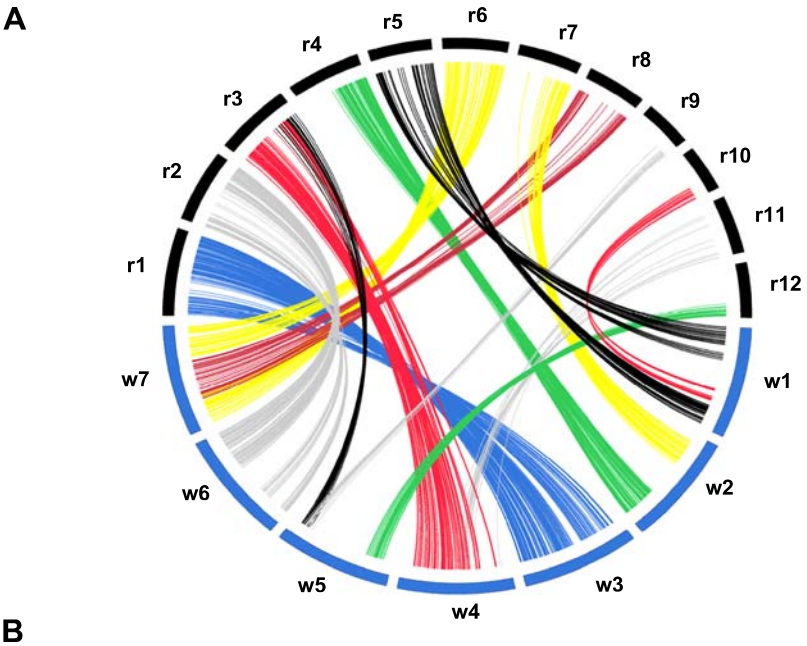
We recently established new alignment as well as statistical criteria to assess the synteny between cereal genomes (rice, wheat, maize, sorghum) including a unigene set of wheat EST-contigs and the fourth version of the rice genome annotation available at TIGR providing 42,654 genes (Salse et al. 2008a). We have now used this strategy on the last version (v5 with 41,046 annoted genes) of the rice genome annotation to re-assess the number of orthologous gene pairs between rice and wheat. To identify orthologous regions between the rice and wheat genomes, we aligned the 5,003 wheat unigene EST-contigs derived from the 6,426 ESTs assigned by the American NSF (National Science Fondation) program (www.gramene.org) to locations on wheat chromosomes, against the 41,046 rice annotated genes. Using 60% CIP and 70% CALP (cf material and method) for the sequence alignment (BLASTN), 34.7% (1,737 compared to 1,055 from Sorrells et al. 2003; and La Rota and Sorrells 2004 and 1,063 from Singh et al. 2007) of the single copy wheat sequences showed similarity (average value of 87.8% CIP and 87.9% CALP) to a unique gene in rice. From the 5,003 mapped wheat unigene EST-contigs, we have identified 1,737 putative orthologs on the basis of 41,046 rice genes compared to 1,805 putative orthologs based on 42,654 rice genes in Salse et al. (2008a). We have lost 68 orthologous relationships. The main difference between the two TIGR versions of the rice genome annotation is the fusion of gene copies in tandem (initially considered as distinct orthologous relationships) into a single-gene model in the last version. From the 1,737 sequence pairs, a subsequent CloseUp analysis (density ratio 2/cluster length 20/match 5) indicated that 827 of them can be considered as true orthologs suitable for COS marker design. This subset was then compared with those that would have been obtained by "classical" BLAST alignments of the 827 rice–wheat

orthologous sequences using a range of cut-off expect values. Nineteen classes containing three to 111 orthologous sequences with E values ranging from $1e^{-01}$ to 0 were obtained (data not shown). The main class contained 111 orthologs with an expect value of $1e^{-41}$ to $1e^{-50}$. The results also showed that with a cut-off value of $1e^{-40}$ which is used generally to define orthologous relationships (Singh et al. 2007), we would have failed to identify 227 (27%) of the 827 ortholog pairs. This demonstrates the power of applying our new combination of alignment criteria and statistical validation to infer orthologous relationships between genomes (Bolot et al. 2009).

The 827 orthologs are present in 13 syntenic regions that cover 76.4 and 91.7% of the rice and wheat genomes, respectively. Figure 1a represents the 13 colinear blocks identified among the 12 rice chromosomes (top semicircle) and the seven wheat chromosome groups (bottom semicircle). They correspond to the following chromosome pairs: w1–r5 (67 genes), w1–r10 (17 genes), w2–r4 (89 genes), w2–r7 (39 genes), w3–r1 (176 genes), w4–r3 (112 genes), w4–r11 (15 genes), w5–r3 (26 genes), w5–r9 (11 genes), w5–r12 (30 genes), w6–r2 (108 genes), w7–r6 (92 genes), and w7–r8 (45 genes; Fig. 1a, Supplementary material S1). We found the largest orthologous regions (176 orthologous genes, #5 in Fig. 1b) between w3 (between coordinates 11.8 to 199.1) and r1 (between 0.55 to 43.34 Mb). The smallest orthologous region (11 orthologous genes, #9 in Fig. 1b) occurred between w5 (between coordinates 160.9 to 180.4) and r9 (between coordinates 16.33 to 22.72 Mb). The 827 true rice–wheat orthologs (cf supplementary material S1) provide the basis for inferring new relationships relative to previous studies that were either based only on small-scale sequence comparisons or lacked statistical validation when carried out at the whole genome scale (Sorrells et al. 2003; La Rota and Sorrells 2004; Singh et al. 2007).

Identification of 695 COS markers

The software package, named ISM, has been developed to automatically select primer pairs that amplify introns for any wheat EST associated with a rice ortholog. The software program is divided into two modules (cf Fig. 2a). The first identifies the exon–intron boundaries



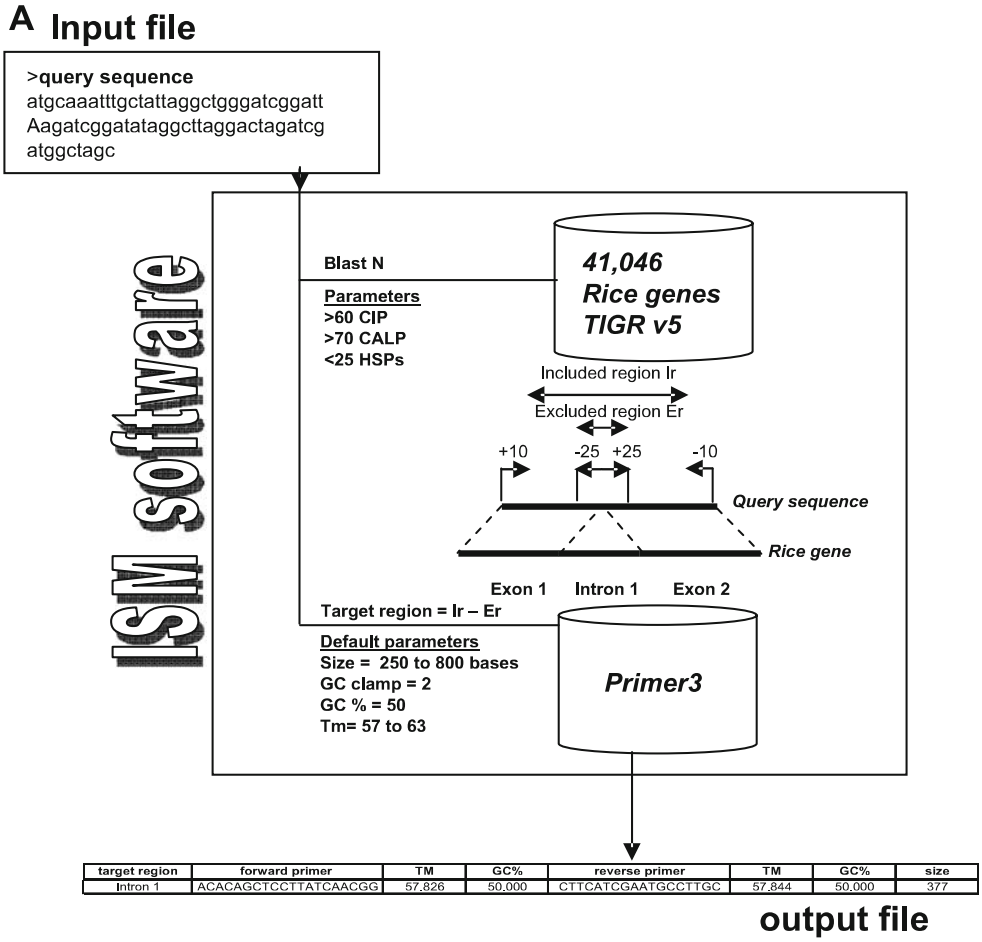
B

Synteny BlocK	Chromosome relationship	Number of orthologs	Rice start (Mb)	Rice end (Mb)	Rice block size (Mb)	Wheat start	Wheat end	Wheat block size	COS markers
1	1W - 5R	**67**	0,27	29,58	29,31	0,0	193,9	193,9	**57**
2	1W - 10R	**17**	16,45	22,86	6,41	122,0	146,4	24,4	**14**
3	2W - 4R	**89**	16,63	35,20	18,57	136,4	199,6	63,2	**74**
4	2W - 7R	**39**	7,79	29,51	21,72	14,9	69,0	54,1	**33**
5	3W - 1R	**176**	0,55	43,34	42,79	11,8	199,1	187,3	**149**
6	4W - 3R	**112**	0,85	30,25	29,40	25,8	178,6	152,8	**93**
7	4W - 11R	**15**	4,86	28,43	23,57	17,9	107,5	89,6	**12**
8	5W - 3R	**26**	23,01	35,59	12,58	165,6	184,3	18,7	**22**
9	5W - 9R	**11**	16,33	22,72	6,39	160,9	180,4	19,5	**8**
10	5W - 12R	**30**	19,98	26,63	6,65	16,9	49,7	32,8	**24**
11	6W - 2R	**108**	0,16	35,81	35,65	0,7	199,2	198,5	**93**
12	7W - 6R	**92**	0,86	31,07	30,21	4,9	196,6	191,7	**77**
13	7W - 8R	**45**	0,11	28,25	28,14	54,2	131,4	77,2	**39**
TOTAL		**827**							**695**

Fig. 1 Thirteen statistically validated syntenic blocks between rice and wheat. **a** Schematic representation using CIRCOS software (http://mkweb.bcgsc.ca/circos/) of 827 pairs of orthologous genes (linked by *thin colored lines*) identified on 13 syntenic blocks between seven wheat chromosomes (*blue bars at the bottom*) and 12 rice chromosomes (*black bars at the top*) through a combination of alignment and statistical criteria. **b** Detailed features for the 13 orthologous regions identified between rice and wheat. For each region, the chromosome carrying the region of orthology in wheat (*w*) and rice (*r*), the number of gene orthologs, the limits of the regions in wheat (coordinates on the consensus chromosomes) and in rice (bp position on the pseudomolecules), the region size are indicated

within the wheat EST sequence from the rice–wheat alignment. The second identifies a list of primers spanning the intron–exon boundaries presenting the following criteria: (1) primer size (20 to 25 mer), (2) amplicon size (between 250–800 bp), (3) Tm (between 57 and 63), (4) GC clamp (equal to 2, i.e. a G or C at the 5' extremity), (5) GC percentage (50%). Users can paste their sequence of interest on the input page and automatically obtain primer pairs for intron amplifications if a unique rice orthologous gene is identified.

Among the 827 orthologous pairs, we have automatically selected 695 COS marker primer pairs. The remaining genes



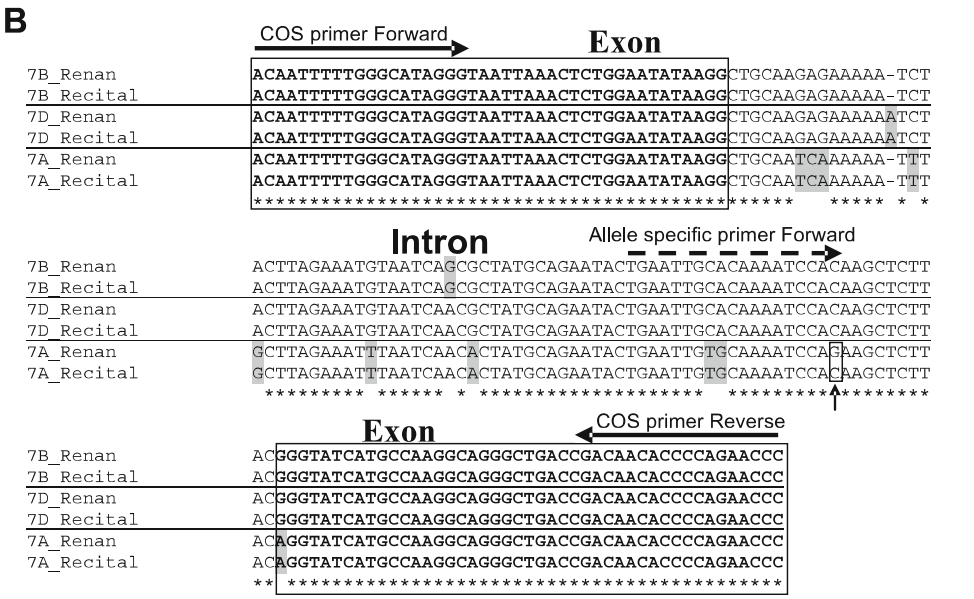


Fig. 2 Graphical representation of the 'ISM' Software pipeline. **a** The figure schematically represents (1) the input sequence file at the top (nucleic sequence in the fasta format) (2) the ISM software principle at the center (alignment against rice genes with appropriate CIP and CALP values, selection of intron-spanning primers with appropriate design criteria such as size, GC%, GC clamp, Tm, and (3) the output file at the bottom (table file with the primer name, sequence, Tm, and GC%). **b** Sequence-based polymorphism between Renan and Recital genotypes for a COS marker derived from the wheat EST bg274300. The six corresponding sequences are presented for the B (*top*), D (*center*), and A (*bottom*) homoeologs for the Renan and Recital genotypes. The primer positions are shown with *horizontal black arrows*. The SNPs identified between the A, B, and D genomes are shown with *gray boxes*. The SNP identified between Renan and Recital is shown with a *vertical black arrow*

were not suitable for primer design as (1) no intron or 5'/3' UTRs (UnTranslated Region) was identified; (2) valid primer design criteria were not satisfied (i.e. primer size; amplicon size, Tm, GC Clamp, GC%). The primer list is provided in supplementary material S2. We selected a unique primer pair for a single intron (or UTR) of each rice–wheat ortholog pair from this list. As other primer pairs may be required for the same genes or for other rice–wheat relationships, the software is publicly available for users at http://www.clermont.inra.fr/umr1095/ism_software/. Among the 695 COS markers available, 71, 107, 149, 105, 54, 93, and 116 are distributed on wheat chromosome groups 1, 2, 3, 4, 5, 6 and 7, respectively (cf Fig. 1b, supplementary material S2). Moreover, the 695 COS markers defined 470, 511, and 554 loci on chromosome groups A, B, and D, respectively; for a total of 1,535 mapped loci.

Figure 2b illustrates the SNP origin nature of the polymorphism based on COS marker sequencing (cf Materials and methods section). In a polyploid context such as wheat, PCR products derived from homoeologous genes are often amplified simultaneously. The three homoeoloogus sequences (gA-B-D) derived from the COS marker bg274300 clearly demonstrates that a single SNP could be identified between Renan and Recital genotypes for the gA homoeolog corresponding to a G to C substitution (cf black arrow on Fig. 2b). From the identification of the SNP, it is then possible to design an allele specific primer to derive a simple dominant marker that can be genotyped on agarose gel (cf black dotted arrow on Fig. 2b).

Fine mapping of wheat QTL on chromosome 7A for baking quality trait with COS markers

Among the 827 orthologous gene pairs identified between rice and wheat, 137 (92+45) define the synteny between rice chromosome 6/8 and wheat chromosome 7A (cf Fig. 1b blocs # 12 and 13). A subset of 31 COS markers was identified for fine mapping of the QTL for pentosan viscosity previously identified on chromosome 7A within an 81 cM confidence interval using a Renan × Recital progeny (Groos et al. 2003).

In wheat, PCR products derived from homoeologous genes are amplified simultaneously. Although high degrees of similarity are found among wheat homoeologous genes, intron regions have a greater degree of polymorphism in terms of insertions/deletions and base substitution than exons. An efficient system requires COS markers spanning introns, associated with a high-throughput method of SNP detection such as single-strand conformational polymorphism (SSCP) that allows identification of homoeologous amplicons. Although various genotyping methods for SNP detection are available (for review Kim and Misra 2007), we were interested in using a co-dominant, robust, and high-throughput (multiplexing) technique allowing to decipher between amplified homoeologous gene copies. Single-strand conformation polymorphism has been widely used, based especially on acrylamide gels. This concept has recently been applied in pearl millet by Bertin et al. (2005), where millet ESTs were aligned against the rice gene sequences to manually predict the location of introns and amplify products across the sequences followed by detection by SSCP on acrylamide gels. Here we have developed and applied an efficient SSCP protocol on capillary sequencers (Applied Biosystem) to be able to detect (1) homoeologous copies of a gene and (2) SNP and size polymorphism (cf Materials and methods section for details). Other genotyping strategies may be used depending on the facilities available such as direct sequenced as referenced previously or cleaved amplified polymorphic sequences (CAPS) derived PCR markers on agarose gel (Ishikawa et al. 2007).

From the 31 selected COS markers, homogeneously distributed on rice chromosomes 6 and 8, for fine QTL mapping (cf Fig. 3a, supplementary material S3), 14 (45%) COS markers were identified as polymorphic between parents of a mapping population (Renan and Recital, cf "Materials and methods"). Among the 14 COS markers selected from the 31 initially available for wheat chromosome 7A, 12 were mapped within the QTL interval (cf Fig. 3b, supplementary material S3, see also Groos et al. 2003). In other words 86% (12 out of 14) of the markers selected on a synteny basis that are polymorphic between the population parents successfully map in the targeted region. The two remaining COS markers were mapped on chromosome 7B and on other chromosome groups, and thus define rearrangements between rice and wheat (cf supplementary material S3). The initial QTL was mapped on a 81 cM interval through a SSR genotyping approach (Groos et al. 2003). By mapping 12 COS markers distributed within the rice orthologous region, the confidence interval was reduced from 81 cM to 4.6 cM (R^2 of 32.9% between positions 158.3 and 162.9 cM) by the mean of 12 COS markers derived from the synteny between rice and wheat (cf Fig. 3b). The eight COS markers mapped within a small confidence interval provide a fine mapping resolution of one marker every 0.6 cM. Bootstrap analysis (i.e. *p*-value interval calculation on 1,000 permutations of 1,000 random samples) and refine a 2.9 cM interval flanked by COS markers c744394 and bg274300 with an associated *p*-value above 30%. The use of COS markers derived from the synteny with the rice genome sequence offers the opportunity to identify a list of candidate genes, discussed in the next section, within the rice orthologous region flanked by COS markers surrounding the locus (or QTL) of interest in wheat (or other cereal genome). In our case the two flanking COS markers define two rice regions on

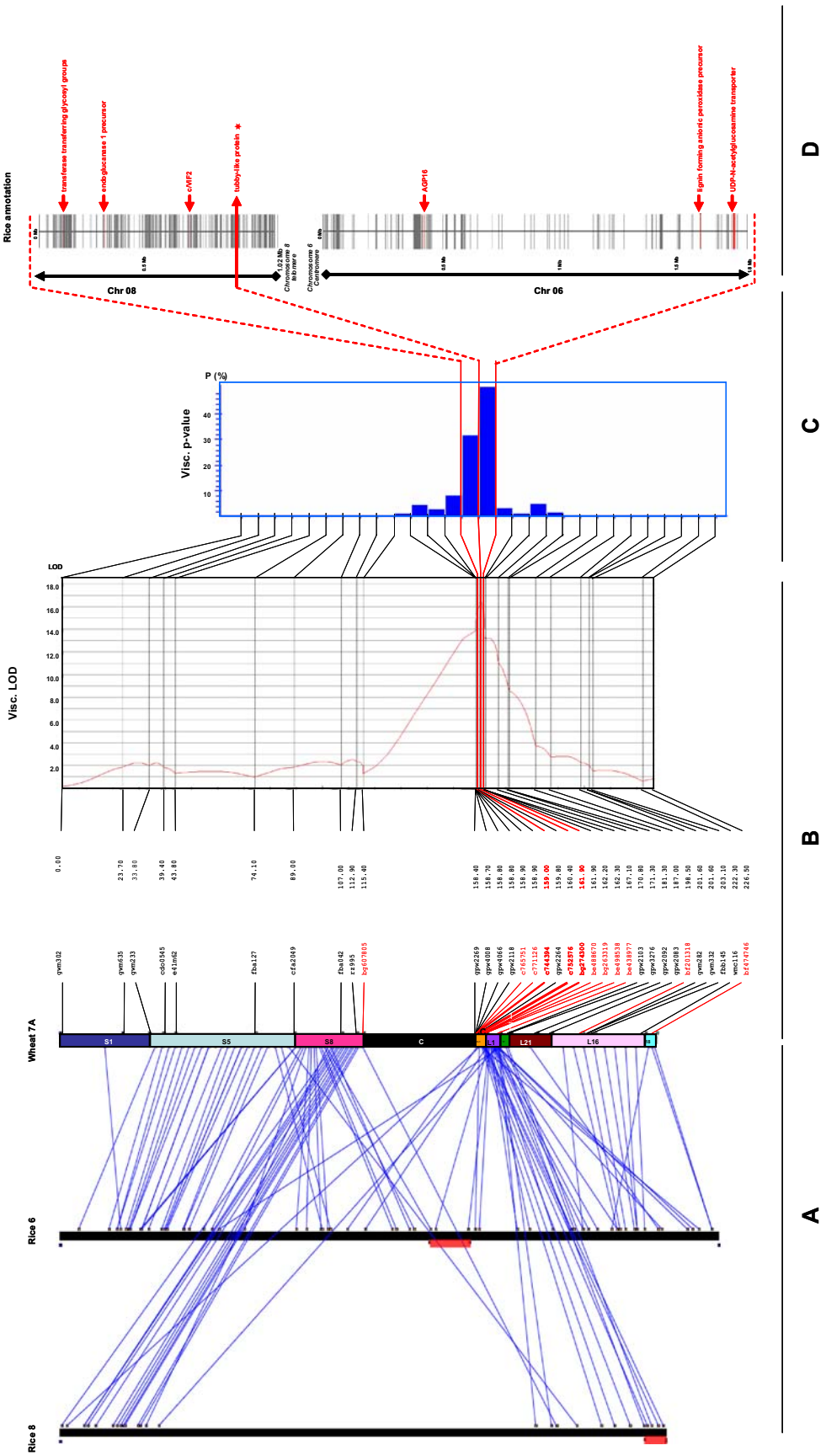


Fig. 3 Fine mapping of a wheat flour viscosity QTL on chromosome 7A in wheat using COS markers. **a** Graphical representation of the synteny between rice (r6 and r8) and wheat (w7A). The wheat and rice orthologs are linked with *blue lines*. **b** The wheat genetic map including SSR (*black*) and COS (*red*) markers within deletion bins (*vertical colored blocks*) and the PV QTL associated with the calculated LOD score are graphically represented respectively at the left and the right end side. **c** Graphical representation of *p*-value (*y*-axis) calculation on genetic intervals (*x*-axis) from a bootstrap analysis on 1,000 permutations of 1,000 random samples. **d** Graphical representation of the rice annotated region that lies under the QTL confidence interval. Candidate genes are highlighted with *red arrows* and associated with their corresponding function delivered from the fifth version of the rice genome annotation available at TIGR

chromosome 8 (c744394 to the telomere) and chromosome 6 (from bg274300 to the centromere), cf Fig. 3d. Such a clear orthologous region in rice can be considered as a source of new COS markers (such as the COS c752576 derived from a conserved tubby-like gene highlighted with a red star in the Fig. 3d) for a second round of mapping if necessary on an enlarged recombinant population. In other words, any following series of mapping will be targeted through the use of the rice orthologous region. Finally, the identification of a clear orthologous region in rice flanked by two COS markers surrounding the QTL in wheat provides a list of candidate genes for the trait of interest (cf supplementary material S4). Figure 3 represents the overall fine mapping approach on chromosome 7A based on COS markers that can be applied to any region of interest with the same resolution. At the left hand side of the figure (cf Fig. 3a), the 92 and 45 rice–wheat orthologs are linked by blue lines as representative of the synteny between wheat chromosome 7A (on the basis of wheat ESTs assigned to deletion bins) and r6 and r8 genes respectively (cf Fig. 1b). In the center of the figure (cf Fig. 3b) the Renan × Recital genetic map and its associated QTL for PV (LOD, *p*-value) are shown with a mapping resolution with more than one marker every CentiMorgan. Among the 12 COS markers (highlighted in red) only two markers do not map to the QTL interval. This result is in agreement with the 27.2% of wheat orthologs previously identified at the genome-wide level of synteny analysis that were not located in the orthologous wheat deletion bin thereby indicating rearrangements within orthologous regions. Finally, at the right of the figure (cf Fig. 3d) the gene annotation of the rice orthologous region of the PV QTL is shown and provided as supplementary material S4.

Discussion

COS marker identification for cross-genome map-based QTL cloning in wheat

The 827 orthologous gene pairs conserved in rice and wheat identified, define the synteny between the two genomes and were used to implement a software program (http://www.clermont.inra.fr/umr1095/ism_software/) to automatically design sequence-specific primers that amplify gene introns (as a source of SNP polymorphism). We were able to develop primer pairs for 695 COS markers mapped on 1,535 loci (one primer pair for each rice–wheat alignment available) suitable for SNP detection (rice–wheat orthologs and COS markers are respectively provided in supplementary material S1 and S2). Considering the ITMI map (International Triticeae Mapping Initiatives, Röder et al. 1998a, b) made of 2,293 different molecular markers (maily SSRs) and covering 3,980.4 cM, we can consider that our panel of 1,535 loci-based COS markers offers a fine mapping resolution of one COS marker, on average, every 2.6 cM. These results are consistent with previous data obtained in sorghum by Feltus et al. (2006) with the identification of 384 conserved intron-spanning primers using rice as a template. Moreover, the 695 COS markers derived from an accurate characterisation of 827 rice and wheat orthologous pairs are complementary to recent data obtained by Ishikawa et al. (2007) providing a list of 4,312 intron-spanning TaEST-LUG for *T. aestivum* (expressed sequence tags landmark unique gene loci) derived from 5,665 rice–wheat homologous pairs. These markers, derived from the sequence alignment of non-mapped wheat ESTs to the rice genes, cannot be considered as COS markers since it is of primary importance to deliver a chromosome position for the COS markers on the wheat genome as is the case with our marker data set (through the identification of 1,535 loci).

Early comparative genetic studies demonstrated that a number of genes and quantitative trait loci for developmental and domestication traits, such as shattering, plant height, vernalisation, flowering time, row number, and kernels per row, were at orthologous positions in cereal genomes (Lin et al. 1995; Paterson et al. 1995; Bailey et al. 1999; Peng et al. 1999; Monna et al. 2002). In recent years, the isolation of genes by map-based cloning in barley, wheat, and maize has revealed other examples of conservation between genes at orthologous positions in cereals (*Vrn1*, Yan et al. 2003; *PPD-H1*, Turner et al. 2005). In contrast, other types of genes do not show colinearity between the cereal genomes (for review Salse and Feuillet 2007). Indeed, there is no example of colinearity for disease resistance (R) genes in grasses and, so far, map-based cloning of R genes in cereals has not benefited significantly from the rice genome information (Leister et al. 1998). However, even if the gene is not present at its orthologous position in rice, the flanking genes are often sufficiently well conserved to provide a collection of markers (such as COS) that can be used to saturate the target region in the other cereal genomes (for example, *Rpg1*,Brueggeman et al. 2002; *Rph7*, Brunner et al. 2003; and *Bru1*, Asnaghi et al. 2004).

As a proof of concept 31 COS markers located on wheat chromosome 7A, defining the synteny with rice chromosomes 6 and 8, were studied in detail. Almost 45% were polymorphic between two parent lines of interest. Eighty-six percent of the polymorphic markers were positioned within a QTL interval for pentosan viscosity. Only 14% failed to map in the appropriate orthologous position in wheat due to rice/wheat rearrangements. Twelve COS markers provided a fine mapping resolution of one marker every 0.6 cM. A 2.9 cM interval flanked by COS markers

c744394 and bg274300 with a significant PV *p*-value define two rice orthologous regions on chromosome 8 (c744394 to the telomere) and chromosome 6 (from bg274300 to the centromere), cf Fig. 3. The gene annotation of the rice orthologous region of the PV QTL allows the establishment of a list of 124 genes of known function which can be considered as (1) a source of new COS markers and (2) candidate genes for the PV traits suitable for functional validation. To test this last hypothesis we compared the candidate gene list with the genes known to be involved in the Arabinoxylan biosynthesis pathway (Saulnier et al. 2007). Six genes of major interest are associated with a biological function involved in this pathway (highlighted in red in Fig. 3d): A cell wall vacuolar inhibitor of fructosidase C/VIF2 (LOC_Os08g01670), an Arabinogalactan protein AGP16 (LOC_Os06g30920), a lignin forming anionic peroxidase precursor (LOC_Os06g32960-90), a transferase, transferring glycosyl groups (LOC_Os08g02370), a UDP-N-acetylglucosamine transporter (LOC_Os06g33210), and an endoglucanase precursor (LOC_Os08g02220). Genes that are located in the rice orthologous region of the QTL of interest can thus be prioritized for cloning, functional analyses, and the development of new COS markers for crop improvement.

Tools for COS identification for any cereal genome using rice as a pivotal sequence

We have presented in this article an accurate and automatic approach to develop COS markers for fine mapping in wheat. The same procedure can be performed for any other cereal by using rice as a pivotal genome through three deliverables provided in this article: (1) statistically validated COS marker identification based on a combination of alignment and statistical validation criterion (cf supplementary material S1) and (2) ISM software for automatic selection of sequence-specific primers for intron amplification as a source of SNP polymorphism (http://www.clermont.inra.fr/umr1095/ism_software/; cf "Materials and methods"). These COS markers offer three main advantages: (1) portability (as they are selected within exons of genes conserved between rice and wheat), (2) highly polymorphic (as they exploit the highest source of polymorphism, SNP), and (3) informativeness for series of genetic mapping (as they define clear orthologous regions in rice, a source of new markers for successive round of mapping).

The value of intron-spanning COS markers depends on the ortholog criteria used to define them (as artefactual syntenic relationships may increase the percentage of COS markers not mapped in the region of interest) and the designed primers (as exons lack polymorphism and only intron polymorphism exploitation will increase the polymorphism rate). Several genotyping approach can be used depending on the facilities available such as SSCP (on capillary sequencer, current article or on acrylamide gel, Bertin et al. 2005), CAPS on agarose gel (Ishikawa et al. 2007), Sequence-specific PCR on agarose gel (based on homeolog sequencing, current article). Moreover, Recent high-throughput sequencing technologies, such as 454 (Roche), may offer rapidly the opportunity to sequence the 695 COS and more precisely the associated 1,535 loci in several wheat genotypes, as well as a subset of loci of interest.

Acknowledgements Data included in the current article including COS genotyping data as well as QTL analysis in the Renan × Recital INRA genetic population were generated within programs partially funded by GENOPLANTE and HEALTH GRAIN (FP6 EU). We would like also to thank Richard Cooke (Laboratoire Génome et Développement des Plantes, Perpignan University, France) for fruitful discussions during the article preparation.

References

Asnaghi C, Roques D, Ruffel S, Kaye C, Hoarau JY, Telismart H, Girard JC, Raboin LM, Risterucci AM, Grivet L, D'Hont A (2004) Targeted mapping of a sugarcane rust resistance gene (*Bru1*) using bulked segregant analysis and AFLP markers. Theor Appl Genet 108:759–764

Bailey PC, McKibbin RS, Lenton JR, Holdsworth MJ, Flintham JE, Gale MD (1999) Genetic map locations for orthologous *Vp1* genes in wheat and rice. Theor Appl Genet 98:281–284

Bertin I, Zhu JH, Gale MD (2005) SSCP-SNP in pearl millet-a new marker system for comparative genetics. Theor Appl Genet 110:1467–1472

Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J (2009) Curr Opin Plant Biol 12(2):1–7

Brueggeman R, Rostoks N, Kudrna D, Kilian A, Han F, Chen J, Druka A, Steffenson B, Kleinhofs A (2002) The barley stem rust-resistance gene *Rpg1* is a novel disease-resistance gene with homology to receptor kinases. Proc Natl Acad Sci U S A 99:9328–9333

Brunner S, Keller B, Feuillet C (2003) A large rearrangement involving genes and low copy DNA interrupts the micro-colinearity between rice and barley at the *Rph7* locus. Genetics 164:673–683

Feltus FA, Wan J, Schulze SR, Estill JC, Jiang N, Paterson AH (2004) An SNP resource for rice genetics and breeding based on subspecies *Indica* and *Japonica* genome alignments. Genome Res 14:1812–1819

Feltus FA, Singh HP, Lohithaswa HC, Schulze SR, Silva TD, Paterson AH (2006) A comparative genomics strategy for targeted discovery of single-nucleotide polymorphisms and conserved-noncoding sequences in orphan crops. Plant Physiol 140 (4):1183–1191

Fulton TM, Van der Hoeven R, Eannetta NT, Tanksley SD (2002) Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14 (7):1457–1467

Gupta PK, Rustgi S (2004) Molecular markers from the transcribed/expressed region of the genome in higher plants. Funct Integr Genomics 4(3):139–162

Groos C, Robert N, Bervas E, Charmet G (2003) Genetic analysis of grain protein content, grain yield and thousand kernel weight in bread wheat. Theor Appl Genet 106:1032–1040

Hampson S, McLysaght A, Gaut B, Baldi P (2003) LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. Genome Res 13:1–12

Hampson SE, Gaut BS, Baldi P (2005) Statistical detection of chromosomal homology using shared-gene density alone. Bioinformatics. 21:1339–1348

Hyne V, Kearsey MJ (1995) QTL analysis: further uses of 'marker' regression. Theor Appl Genet 91:471–476

Ishikawa G, Yonemaru J, Saito M, Nakamura T (2007) PCR-based landmark unique gene (PLUG) markers effectively assign homoeologous wheat genes to A, B and D genomes. BMC Genomics 30(8):135

Kearsey MJ, Hyne V (1994) QTL analysis: a simple 'marker-regression' approach. Theor Appl Genet 89:698–702

Kim S, Misra A (2007) SNP genotyping: technologies and biomedical applications. Annu Rev Biomed Eng 9:289–320

Kosambi DD (1944) The estimation of map distances from recombination values. Ann Eugen 12:172–175

Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84:2363–2367

La Rota M, Sorrells ME (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. Funct Integr Genomics 4:34–46

Leister D, Kurth J, Laurie DA, Yano M, Sasaki T, Devos K, Graner A, Schulze-Lefert P (1998) Rapid reorganization of resistance gene homologues in cereal genomes. Proc Natl Acad Sci U S A 95:370–375

Lin YR, Schertz KF, Paterson AH (1995) Comparative analysis of QTLs affecting plant height and maturity across the poaceae, in reference to an interspecific sorghum population. Genetics 141:391–411

Martinant JP, Billot A, Bouguennec A, Charmet G, Saulnier L, Branlard G (1999) Genetic and environmental variations in water-extractable arabinoxylans content and flour extract viscosity. J Cereal Sci 30:45–48

Monna L, Kitazawa N, Yoshino R, Suzuki J, Masuda H, Maehara Y, Tanji M, Sato M, Nasu S, Minobe Y (2002) Positional cloning of rice semi-dwarfing gene, *sd-1*: rice 'green revolution gene' encodes a mutant enzyme involved in gibberellin synthesis. DNA Res 9:11–17

Panjabi P, Jagannath A, Bisht NC, Padmaja KL, Sharma S, Gupta V, Pradhan AK, Pental D (2008) Comparative mapping of Brassica juncea and Arabidopsis thaliana using intron polymorphism (IP) markers: homoeologous relationships, diversification and evolution of the A, B and C Brassica genome. BMC Genomics 9:113

Paterson AH, Lin YR, Li ZK, Schertz KF, Doebley JF, Pinson SRM, Liu SC, Stansel JW, Irvine JE (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. Science 269:1714–1718

Peng JR, Richards DE, Hartley NM, Murphy GP, Devos KM, Flintham JE, Beales J, Fish LJ, Worland AJ, Pelica F, Sudhakar D, Christou P, Snape JW, Gale MD, Harberd NP (1999) 'Green revolution' genes encode mutant gibberellin response modulators. Nature 400:256–261

Qi LL, Echalier B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorák J, Linkiewicz AM, Ratnasiri A, Dubcovsky J, Bermudez-Kandianis CE, Greene RA, Kantety R, La Rota CM, Munkvold JD, Sorrells SF, Sorrells ME, Dilbirligi M, Sidhu D, Erayman M, Randhawa HS, Sandhu D, Bondareva SN, Gill KS, Mahmoud AA, Ma XF, Miftahudin GJP, Conley EJ, Nduati V, Gonzalez-Hernandez JL, Anderson JA, Peng JH, Lapitan NL, Hossain KG, Kalavacharla V, Kianian SF, Pathan MS, Zhang DS, Nguyen HT, Choi DW, Fenton RD, Close TJ, McGuire PE, Qualset CO, Gill BS (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

Röder MS, Korsun V, Gill BS, Ganal MW (1998a) The physical mapping of microsatellite markers in wheat. Genome 41:278–283

Röder MS, Korsun V, Wendehake K, Plaschke J, Tixier MH, Leroy P, Ganal MW (1998b) A microsatellite map of the wheat genome. Genetics 149:2007–2023

Salse J, Feuillet C (2007) Comparative genomics of cereals. Chapter 18 in Genomics-assisted crop improvement; published by Springer, pp 177–205

Salse J, Piegu B, Cooke R, Delseny M (2004) New *in silico* insight into the synteny between rice (*Oryza sativa* L.) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplications in the rice genome. Plant J 38:396–409

Salse J, Bolot S, Throude T, Jouffe V, Piegu B, Masood Quraishi U, Calcagno C, Cooke C, Delseny M, Feuillet C (2008a) Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. Plant Cell 20(1):11–24

Salse J, Chagué V, Bolot S, Magdelenat G, Huneau C, Pont C, Belcram H, Couloux A, Gardais S, Evrard A, Segurens B, Charles M, Ravel C, Samain S, Charmet G, Boudet N, Chalhoub B (2008b) 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 9:555

Saulnier L, Sado PE, Branlard G, Charmet G, Guillon F (2007) Wheat arabinoxylans: exploiting variation in amount and composition to develop enhanced varieties. J Cereal Sci 46:261–281

Singh NK, Dalal V, Batra K, Singh BK, Chitra G, Singh A, Ghazi IA, Yadav M, Pandit A, Dixit R, Singh PK, Singh H, Koundal KR, Gaikwad K, Mohapatra T, Sharma TR (2007) Single-copy genes define a conserved order between rice and wheat for understanding differences caused by duplication, deletion, and transposition of genes. Funct Integr Genomics 7(1):17–35

Sorrells ME, La Rota M, Bermudez-Kandianis CE, Greene RA, Kantety R, Munkvold JD, Miftahudin MA, Ma X, Gustafson PJ, Qi LL, Echalier B, Gill BS, Matthews DE, Lazo GR, Chao S, Anderson OD, Edwards H, Linkiewicz AM, Dubcovsky J, Akhunov ED, Dvorak J, Zhang D, Nguyen HT, Peng J, Lapitan NL, Gonzalez-Hernandez JL, Anderson JA, Hossain K, Kalavacharla V, Kianian SF, Choi DW, Close TJ, Dilbirligi M, Gill KS, Steber C, Walker-Simmons MK, McGuire PE, Qualset CO (2003) Comparative DNA sequence analysis of wheat and rice genomes. Genome Res 13:1818–1827

Turner A, Beales J, Faure S, Dunford RP, Laurie DA (2005) The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. Science 310:1031–1034

Visscher PM, Haley CS, Thompson R (1996) Marker-assisted introgression in backcross breeding programs. Genetics 144:1923–1932

Wei F, Coe E, Nelson W, Bharti AK, Engler F, Butler E, Kim H, Goicoechea JL, Chen M, Lee S, Fuks G, Sanchez-Villeda H, Schroeder S, Fang Z, McMullen M, Davis G, Bowers JE, Paterson AH, Schaeffer M, Gardiner J, Cone K, Messing J, Soderlund C, Wing RA (2007) Physical and genetic structure of the maize genome reflects its complex evolutionary history. PLoS Genet 3(7):e123

Wu F, Mueller LA, Crouzillat D, Pétiard V, Tanksley SD (2006) Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics 174(3):1407–1420

Yan L, Loukoiannov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene *VRN1*. Proc Natl Acad Sci U S A 100:6263–6268

Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C, Zhang J, Zhang Y, Li R, Xu Z, Li S, Li X, Zheng H, Cong L, Lin L, Yin J, Geng J, Li G, Shi J, Liu J, Lv H, Li J, Wang J, Deng Y, Ran L, Shi X, Wang X, Wu Q, Li C, Ren X, Wang J, Wang X, Li D, Liu D, Zhang X, Ji Z, Zhao W, Sun Y, Zhang Z, Bao J, Han Y, Dong L, Ji J, Chen P, Wu S, Liu J, Xiao Y, Bu D, Tan J, Yang L, Ye C, Zhang J, Xu J, Zhou Y, Yu Y, Zhang B, Zhuang S, Wei H, Liu B, Lei M, Yu H, Li Y, Xu H, Wei S, He X, Fang L, Zhang Z, Zhang Y, Huang X, Su Z, Tong W, Li J, Tong Z, Li S, Ye J, Wang L, Fang L, Lei T, Chen C, Chen H, Xu Z, Li H, Huang H, Zhang F, Xu H, Li N, Zhao C, Li S, Dong L, Huang Y, Li L, Xi Y, Qi Q, Li W, Zhang B, Hu W, Zhang Y, Tian X, Jiao Y, Liang X, Jin J, Gao L, Zheng W, Hao B, Liu S, Wang W, Yuan L, Cao M, McDermott J, Samudrala R, Wang J, Wong GK, Yang H (2005) The genomes of *Oryza sativa*: a history of duplications. PLoS Biology 3(2):e38