

Improved criteria and comparative genomics tool provide new insights into grass paleogenomics

Jerome Salse, Michael Abrouk, Florent Murat, Umar Masood Quraishi and Catherine Feuillet

Submitted: 24th May 2009; Received (in revised form): 31st July 2009

Abstract

In the past decade, a number of bioinformatics tools have been developed to perform comparative genomics studies in plants and animals. However, most of the publicly available and user friendly tools lack common standards for the identification of robust orthologous relationships between genomes leading non-specialists to often over interpret the results of large scale comparative sequence analyses. Recently, we have established a number of improved parameters and tools to define significant relationships between genomes as a basis to develop paleogenomics studies in grasses. Here, we describe our approaches and propose the development of community-based standards that can be used in comparative genomic studies to (i) identify robust sets of orthologous gene pairs, (ii) derive complete sets of chromosome to chromosome relationships within and between genomes and (iii) model common paleo-ancestor genome structures. The rice and sorghum genome sequences are used to exemplify step-by-step a methodology that should allow users to perform accurate comparative genome analyses in their favourite species. Finally, we describe two applications for accurate gene annotation and synteny-based cloning of agronomically important traits.

Keywords: cereals; synteny; paleogenomics; evolution

INTRODUCTION

As more and more mammalian and plant genomes are being sequenced, comparative genomics becomes an increasingly important field of research that has the potential to improve our knowledge into genome function and structure as well as mechanisms of evolution that have shaped the genomes. In the past 5 years, international initiatives led to the development of large sets of genomic resources that allow comparative genomic studies between the grass genomes at a high level of resolution. The International Rice Genome Sequencing Project (IRGSP) recently

completed the sequence of the *Oryza sativa* ssp. *japonica* cultivar Nipponbare [1] and a draft of the sorghum genome sequence was published early this year [2]. In addition, a first release of partial maize B73 genome sequence has been announced in 2008 and large sequenced regions are now available for comparative analyses (<http://www.maizesequence.org/>). Finally, highly saturated gene-based genetic maps are available for most of the agronomically important non-sequenced cereal genomes such as those of the Triticeae tribe, e.g. wheat and barley. These new resources have been used to perform

Corresponding author. Jerome Salse, INRA/Université Blaise Pascal UMR 1095, Génétique, Diversité et Ecophysiologie des Céréales, Domaine de Crouelle, 234 avenue du Brézé, 63100 Clermont-Ferrand, France. Tel: +33-4-73-62-43-80; Fax: +33-4-73-62-44-53; E-mail: jsalse@clermont.inra.fr

Jerome Salse is a junior scientist of the group 'Structure, function and evolution of the wheat genomes' at the INRA Clermont Ferrand. His scientific research focuses on 'Cereal paleogenomics for wheat improvement'.

Florent Murat is a graduate student.

Michael Abrouk and **Umar Masood Quraishi** are two Phd students of the group.

Catherine Feuillet is research director and leader of the group 'Structure, function and evolution of the wheat genomes' at the INRA, Clermont-Ferrand, France. She is one of the co-chairs of the International Wheat Genome Sequencing Consortium (IWGSC), the International Triticeae Mapping Initiative (ITMI) and the European Triticeae Genomics Initiative (ETGI).

Table 1: Bioinformatic tools (algorithms and visualization tools) available online for performing comparative genomics analyses

Algorithm and visualization	
SyntenyAnalyser	http://www2.synteny.net/latest/index.htm
OrthoMCL	http://www.orthomcl.org/cgi-bin/OrthoMcIWeb.cgi?rm=indx
Orthocluster	http://krono.act.uji.es/noticias/orthocluster-a-new-tool-for-mining-syntenic-blocks
PSAT	http://www.nwrce.org/psat/
Syntenyviewer	http://www.yeastgenome.org/fungi/help/yeastsMap.html
DAG chainer	http://dagchainer.sourceforge.net/
Cinteny	http://cinteny.cchmc.org/
vista	http://genome.lbl.gov/vista/index.shtml
osfinder	http://osfinder.dna.bio.keio.ac.jp/
CoGe	http://synteny.cnr.berkeley.edu/CoGe/
synblast	http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-002/tutorial/SynBlast.html#Introduction
Gramene	http://www.gramene.org/
TIGR	http://www.tigr.org/
Narcisse-Cereal	http://www1.clermont.inra.fr/Narcisse-Cereals
GOST (greenphyl)	http://greenphyl.cirad.fr/cgi-bin/greenphyl.cgi
Inparanoid	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
Mscan	http://chibba.agtec.uga.edu/duplication/mcscan/
Sybil	http://sybil.sourceforge.net/index.html
AutoGRAPH	http://autograph.genouest.org/
Visualization only	
GenePalette	http://www.genepalette.org/
GENTle	http://gentle.magnusmanske.de/
pdraw32	http://www.acaclone.com/
SynView	http://eupathdb.org/apps/SynView/index.php
SynBrowse	http://www.synbrowse.org/
Dialign	http://dialign.gobics.de/
Mummer	http://mummer.sourceforge.net/
Circos	http://mkweb.bcgsc.ca/circos/?American.Scientist.cover
Geneious	http://www.geneious.com/
GenomeMatcher	http://www.ige.tohoku.ac.jp/joho/gmProject/gmhome.html

large scale inter-specific sequence comparisons and refine our understanding of colinearity between the grass chromosomes [3, 4].

Numerous tools have been developed and are now publicly available to compare plant genomes and tentatively identify orthologues (Table 1). Some provide an interface to perform genome comparisons based on sequence alignment or sequence phylogeny (SyntenyAnalyser, OrthoMCL [5], Orthocluster [6], PSAT [7], Syntenyviewer [8], DAG chainer [9], Cinteny [10], vista [11], osfinder [12], CoGe [13] synblast [14], Gramene [15], TIGR [16], GOST (greenphyl) [17], Inparanoid [18], Mscan [19], Sybil [20], AutoGRAPH [21]) while others can be used to display comparison data (GenePalette [22], GENTle, pdraw32, SynView [23], SynBrowse [24], Dialign [25], Mummer [26], Circos, Geneious, GenomeMatcher [27]). Several algorithms and/or visualization tools are available

as open source and they can accommodate a number of output file standards to represent comparative analyses derived from genomes alignment. However, the most important and critical step before visualizing alignments is to apply methods that will enable robust assessment of the relationships between the aligned sequences thereby ensuring relevant downstream interpretation of evolutionary mechanisms. We consider that, because it is difficult to infer orthologous (derived from a common ancestor by speciation) and paralogous (derived by duplication within one genome) relationships from sequence comparisons, stringent alignment criteria and statistical validation are essential to evaluate accurately whether the association between two or more genes found in the same order on two chromosomal segments in different genomes occurs by chance or reflects true colinearity. While the web-sites mentioned in Table 1 provide ‘user friendly’

graphical displays of macrocolinearity, they rely on data obtained with low stringency alignment criteria and performed without systematic statistical validation. In addition, they do not take into account the density and location of conserved genes to identify precisely paralogous and orthologous regions and therefore they, generally, overestimate colinearity between different segments of the genomes.

We recently developed and applied new and stringent alignment criteria and performed statistical tests systematically to redefine interchromosomal duplications in the rice, maize, triticeae, and sorghum genomes and, re-assess colinearity relationships between them. This allowed us to detect and characterize shared duplicated regions and establish a model for the evolution of the grass genomes from a common ancestor with $n=5$ proto-chromosomes [28–30]. The aim of this article is to provide details on the methods developed for these paleogenomics studies and propose a standard procedure that can be applied to compare any genomes of interest and model their common ancestral genome structure. The rice (12 chromosomes—41 046 gene models—372 Mb) and sorghum (10 chromosomes—34 008 gene models—659 Mb) genome sequences are used here to exemplify the different steps of the proposed protocol. We also illustrate the possible applications from such comparative analyses and describe a new web tool called ‘Narcisse-Cereals’ that provides direct access to comparative analyses in the grass genomes.

COMPARATIVE GENOMICS: A NEED FOR INCREASED STANDARDS

Sequence alignments are generally performed using the Basic Local Alignment Search Tool (BLAST) sets of programs [31] developed at the beginning of the 1990s. When two genome nucleotide sequences are aligned, BLASTN [31, 32] produces HSPs (High Scoring Pairs) that consist of two sequence fragments of arbitrary but equal length whose alignment is locally maximal and for which the alignment score meets or exceeds a threshold or cut-off score. HSPs are based on statistical criteria such as the e -value, score and percentage identity. However, the detection of conserved regions is limited when sequence alignments are obtained with the BLAST default parameters. We have shown previously that using

BLAST default parameters can lead to misidentification of orthologous regions because they result in the detection of members of gene families that are not truly orthologous or of homologies that are based only on conserved functional domains [28]. To increase the significance of inter-specific sequence alignments for inferring evolutionary relationships between genomes, we defined two new parameters for BLAST analysis: CIP for Cumulative Identity Percentage and CALP for Cumulative Alignment Length Percentage. $CIP = \sum nb \text{ ID by (HSP/AL)} \times 100$ corresponds to the cumulative percent of sequence identity observed for all the HSPs divided by the cumulative aligned length (AL) which corresponds to the sum of all HSP lengths. $CALP = AL/\text{query length}$ is the sum of the HSP lengths (AL) for all HSPs divided by the length of the query sequence. With these parameters, BLAST produces the highest cumulative percentage identity over the longest cumulative length thereby increasing stringency in defining conservation between two genome sequences. After testing different combinations of CIP and CALP values in the different analyses, stringent values of 60% CIP and 70% CALP were shown to provide the most reliable results and were applied for the analyses of colinearity between grass genomes [28].

Most of the comparative genomics studies performed to date were done without applying statistical validation of the results and therefore, the significance of the relationships established in different studies are difficult to infer. In our studies, we have systematically performed a statistical test after the BLAST comparison with the CIP/CALP parameters to validate non-random associations between groups of sequences. Several recently developed software programs, such as LineUP [33], ADHoRE (Automatic Detection of Homologous Regions [34]), FISH (Fast Identification of Segmental Homology [35]) and CloseUP [36], are suitable to perform such validations. ADHoRE and LineUP are based on complex algorithms that take into account precise gene orders and orientations to define colinearity blocks and therefore, may underestimate colinearity by eliminating small local rearrangements. In contrast, FISH allows for a moderate relaxation in the gene order taking into account local rearrangement in k -colinear gene clusters called ‘clumps’ associated with a P -value for which users have to define

a specific threshold. This results in several alternatives depending on the chosen *P*-value which may complicate the interpretation. CloseUp provides a single representation of the colinearity by looking for less than perfect linear gene correspondence between chromosome segments. It is based on three parameters that relate to the gene density ratio, gene cluster length and match number between orthologues. In order to compare heterogeneous genome sequence data sets (from full sequenced genome to large mapped EST collections), we have derived a new approach based on CloseUp but with only two criteria: the density (DR) and the cluster (CR) ratio that are functions of the physical size (size), number of annotated gene (Gnumber) and number of orthologous sequence pairs (Cnumber) defined in the orthologous regions identified with the CIP/CALP parameters. The Density Ratio, $DR = [(Size_1 + Size_2) / (2 \times Cnumber)] \times 100$, represents the number of links between two orthologous regions as a function of the size of the considered blocks while the cluster ratio (CR) = $[(2 \times Cnumber) / (Gnumber_1 + Gnumber_2)] \times 100$, represents the number of links between two orthologous regions as a function of the number of genes in the considered blocks. Statistically significant chromosome to chromosome collinear relationships between two genomes are associated with the combination of the highest CR and lowest DR values while the remaining collinear regions are considered as artefactual, i.e. obtained at random. Our statistical validation is equivalent to a CloseUp analysis based on a DR of 2, a cluster length of 20, match number of 5 [28–30].

Figure 1A is a dot plot [37] representation of the results obtained after aligning the rice genome (12 chromosomes) and sorghum genome (10 chromosomes) sequences using different alignment methods mentioned above. The alignments were performed either with a BLASTn cut off value of $1 \times e^{-40}$ which is classically used in the literature (grey dots), or with a CIP/CALP threshold of respectively, 60–70% (blue dots) followed by statistical validation of gene pairs (red dots) (the three independent dot plots obtained with the BLASTn cut off value of $1 \times e^{-40}$, CIP/CALP threshold and Density/Cluster statistical validation are provided as Supplementary Data). In such a representation, each dot indicates nucleotide conservation between the two aligned sequences and therefore a diagonal represents the conservation of a gene or

a gene cluster between the two genomes. With the BLASTn alignment based on default parameters (such as expect values), diagonals are visible (see Figure 1A and Supplementary Data), however, the analysis is polluted with high background noise and defining what is significant or not becomes difficult. In contrast, the CIP/CALP alignment followed by the DR/CR validation test provides clear diagonals (red dots in Figure 1A and Supplementary Data) that correspond to robust orthologous gene pairs. In this example, the CIP/CALP analysis identified significant relationships between the following rice (r) and sorghum (s) chromosomes: r1–s3, r2–s4, r3–s1, r4–s6, r5–s9, r6–s10, r7–s2, r8–s7, r9–s2, r10–s1, r11–s5, r12–s8 (Figure 1A). This result provides first indications on the origin of the difference in chromosome number between rice (12 chromosomes) and sorghum (10 chromosomes) with eight single chromosome to chromosome relationships (r1–s3, r2–s4, r4–s6, r5–s9, r6–s10, r8–s7, r11–s5, r12–s8) and two double chromosome to chromosome relationships (s1–r3/r10 and s2–r9/r7, cf. dotted red arrows in Figure 1A) that reflect lineage specific fusion events in the sorghum genome. The level of information provided with such an approach contrasts with the data available online at Gramene that suggest for example orthologous relationships between s1 and the 12 rice chromosomes (http://dev.gramene.org/Sorghum_bicolor/Info/Index) without providing clear indication of those which are the most relevant. The true orthologous relationship between s1–r3 is therefore diluted among artefactual additional chromosome to chromosome relationships. When analysing the number of genes conserved between the rice and sorghum genomes, the BLASTn (cut-off value of $1 \times e^{-40}$) analysis indicates that 80.9% of the genes are conserved whereas <30% (28.9%) of gene conservation is found when the analysis is based on the highest identity over the longest sequence alignments (CIP/CALP parameters) and only 13.9% are defined as true orthologues after statistical validation. Thus, by applying a statistically-based selection of the highest identity over the longest sequence alignment, our rather conservative approach excludes sequences that lead potentially to an overestimation of sequence conservation between genomes. However, because the method is developed in three steps, the user can define which level of the resolution is the most appropriate to his or her own analysis.

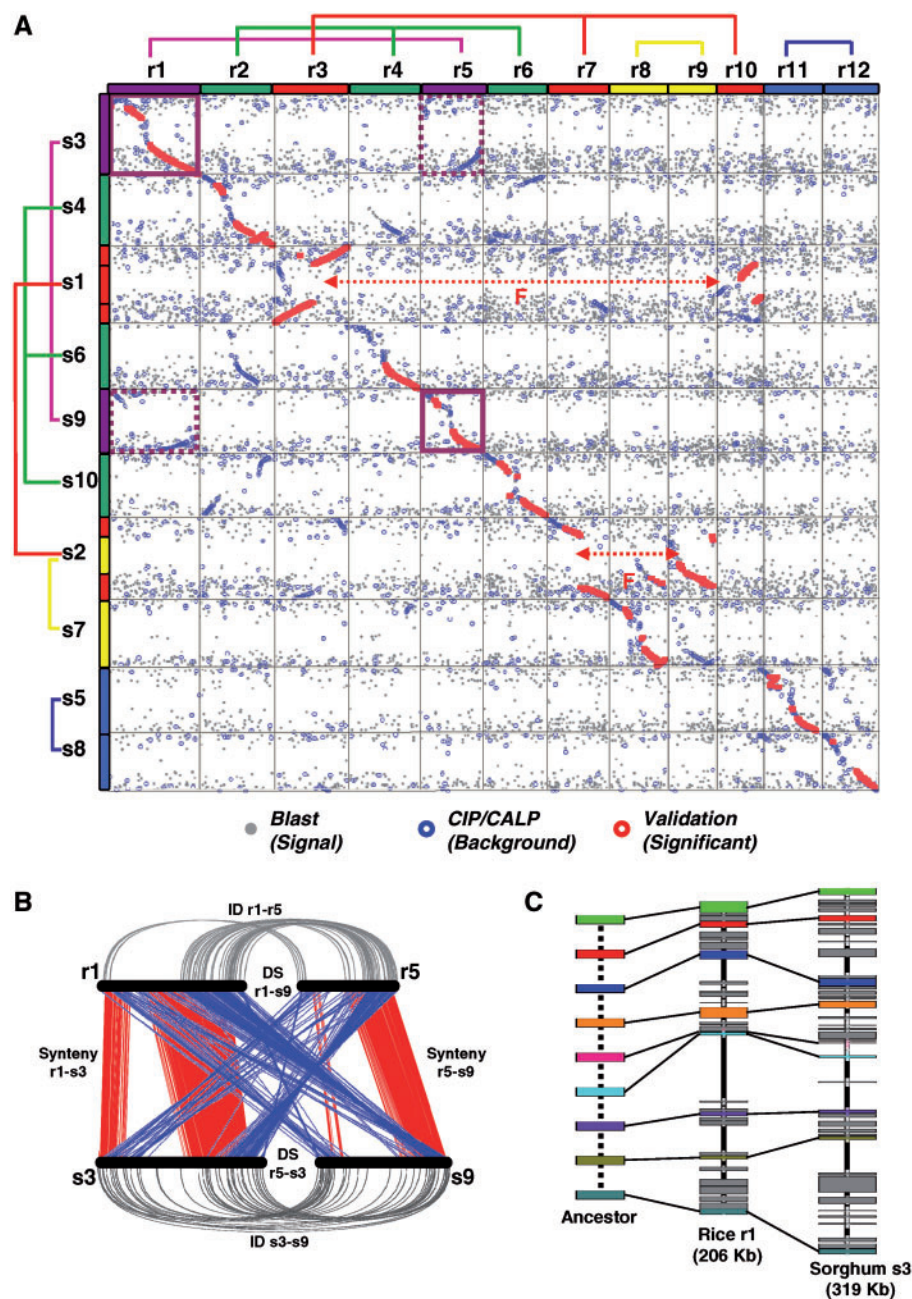


Figure I: Comparison of different alignment methods used for comparative analysis between rice and sorghum. (A) Dot plot alignment of the rice (horizontal) and sorghum genome (vertical) sequences. The dot plot analysis was performed using the DOTTER program [37]. In this representation, diagonals reflect gene conservation between the two genome sequences whereas interruptions indicate loss of colinearity. Gene conservations deduced through (i) BLASTn with a cut off value of $1 \times e^{-40}$ (grey dots) followed by (ii) 60% CIP/70% CALP parameter thresholds (blue dots) followed by (iii) statistical validation (red dots) are illustrated. Ancestral shared duplications identified through dual synteny are indicated with coloured brackets on the rice and sorghum chromosomes. Lineage specific chromosome fusions (F) observed in sorghum (s1, s2) are shown with red arrows. (B) Dual synteny and ID analyses between r1–5 and s3–9 chromosomes—the dual synteny relationships identified between r1–r5 and s3–s9 chromosomes and highlighted with pink squares in (A) is illustrated. The synteny observed between r1–s3 and r5–s9 is indicated with red lines whereas the dual synteny observed between r1–s9 and r5–s3 is represented through the blue lines. The ancestral shared intra-genome duplication (ID) observed between r1–r5 and s3–s9 is shown with grey lines. (C) Ancestral gene content deduced from the synteny between r1 and s3—the synteny relationship identified between the distal ends of the long arms of chromosomes r1 (30 genes on 206 kb) and s3 (40 genes on 319 kb) is illustrated. The coloured boxes indicate orthologous genes (red dot on the panel A) identified between these two chromosome segments. The ancestral gene content (left side) has been deduced from the set of orthologous genes (in different colours) conserved between the rice and sorghum regions.

The need for accurate filtering methods will be increasingly important with the development of comparative genomics studies and community standards need to be established to ensure that different studies can be correlated. The method described here can be directly used to initiate comparative analyses as recently done by Kumar *et al.* [38] and can serve also as a basis for developing further community standards in comparative studies that require stringent filtering such as ancestor genome reconstruction.

GENOME DUPLICATION IDENTIFICATION: DIRECT AND INDIRECT APPROACHES

Two methods have been used for the identification and characterization of genome duplications within the grass genomes. The first and indirect approach called 'Dual Synteny' (DS) has been largely used to identify ancestral intragenomic duplications through the analysis of synteny relationships. It is based on the detection of regions showing a high proportion of gene matches on two different chromosomes within a genome and corresponding to two syntenic regions in another genome. On a dot plot representation, the DS approach allows the identification of regions that are duplicated within a genome and conserved in the other one through the detection of diagonals for two chromosomes in one species corresponding to one chromosome in the other species. In our example, seven regions of dual synteny can be identified on the 12 rice and 10 sorghum chromosomes (coloured brackets in Figure 1A). For example, in addition to the r1/s3 diagonal (red) previously identified with the stringent CIP/CALP and DR/CR analysis, a r5/s3 diagonal (less significant, blue dots) is also detected. Similarly, in addition to the r5/s9 significant relationship, a diagonal is found between the r1 and s9 chromosomes. Another representation of the dual synteny observed between the r1–r5 and s3–s9 chromosomes is provided in Figure 1B. In this representation, orthologous gene pairs between r1–s3 (939 orthologues) and r5–s9 (373 orthologues) are highlighted in red blocks whereas dual synteny relationships between r1–s9 (179 orthologues) and r5–s3 (143 orthologues) are highlighted in blue blocks.

The second and most direct approach called 'Intra-genome Duplication' (ID) consists in aligning a given genome sequence against itself. As the DS

approach, this provides an exhaustive and complete identification of ancestral duplications, but in addition it allows the characterization of lineage specific duplications within genomes. For this reason, we have chosen to use the ID approach with the combination of stringent alignment criteria (CIP/CALP parameters) and statistical validation tests (DR/CR parameters) as a basis to analyse duplications within the Triticeae, rice, maize, sorghum genomes [28–30]. Here, in our example, the ID approach confirms the ancestral duplication shared between r1–r5 and s3–s9 suggested by the dual synteny method but it allows also to identify paralogous gene pairs (47 paralogues between r1 and r5 and 51 paralogues between s3 and s9, grey lines on the Figure 1B) by aligning the rice and sorghum chromosome pairs against themselves,

The conservation of duplications between the rice and sorghum genomes at orthologous positions (e.g. between r1–s3 and r5–s9) indicates that they probably originate from a shared ancestral duplication event that pre-dated the divergence between the two species and occurred in their common ancestor 50–70 million years ago. Under this scenario, the r1 and r5 chromosomes as well as the s3 and s9 chromosomes originate from the duplication of a common ancestral chromosome. Subsequently, the seven dual synteny relationships identified at the whole genome level suggest the presence of seven ancestral duplications (found on the chromosome pair combinations: r11–r12/s5–s8, r5–r1/s9–s3, r10–r3/s1–s1, r7–r3/s2–s1, r4–r2/s6–s4, r9–r8/s2–s7, r2–r6/s4–s10) shared between rice and sorghum. The identification of such ancestral relationships provide the substrate for further ancestral genome structure reconstruction. Using the data obtained from the synteny and orthologous relationships it is first possible to estimate an ancestral or minimum gene content by calculating the number of genes that are conserved between at least two compared genomes. Figure 1C illustrates this process for the rice 1 and sorghum 3 chromosomes. Here, 9 protogenes are identified based on the synteny observed between orthologous regions on r1 (206 kb, 30 genes) and s3 (319 kb, 40 genes) chromosomes. Using this approach at the whole genome level, we have recently estimated the gene content of the grass ancestral genome to 9138 protogenes by combining the orthologous paralogous relationships identified between the rice, sorghum and maize genomes [30].

ANCESTRAL GENOME STRUCTURE MODELING THROUGH THE IDENTIFICATION OF SHARED AND LINEAGE-SPECIFIC REARRANGEMENT EVENTS

Paleogenomics, the study of ancestral genome structures, allows the identification and characterization of mechanisms (e.g. duplications, translocations, and inversions) that have shaped genome species during their evolution. Paleogenomics can be performed through large-scale comparative analyses of actual species and modelling of an ancestral genome structure. Different methods have been developed to compute large genomic scale data for ancestral genome reconstruction. The first step consists in the identification of conserved regions across genomes. In mammals this has been facilitated by a generally moderate reshuffling of chromosomal segments since their divergence from common ancestors ~130 million years ago [39–41]. Different approaches such as (i) cladistic [42], (ii) Genome Rearrangements In Man and Mouse (GRIMM [43]), (iii) Multiple Genome Rearrangement

(MGR [44]) and (iv) Contiguous Ancestral Regions (CAR [45]) have been developed to perform such studies. In contrast to mammals, paleogenomics has been poorly investigated in plants as angiosperm species have undergone a large number of whole genome or segmental duplications, diploidization and small-scale rearrangements (translocations, gene conversions) that make comparative studies between and within the monocotyledon (mainly grasses) and eudicot families very challenging. Here, our conservative approach which allows to define precisely orthologous relationships and shared duplications provides a solid basis to perform reconstruction in the highly rearranged plant genomes.

Based on this approach, we recently proposed a method to reconstruct the ancestral proto-chromosomes in plants [30]. Figure 2 illustrates the three steps that can be deployed to model the ancestral grass genome based on the intra- (i.e. paralogues) and inter- (i.e. orthologues) specific comparisons performed between the rice and sorghum genomes. In step 1 (Figure 2), results of the synteny are combined with those of the shared duplication

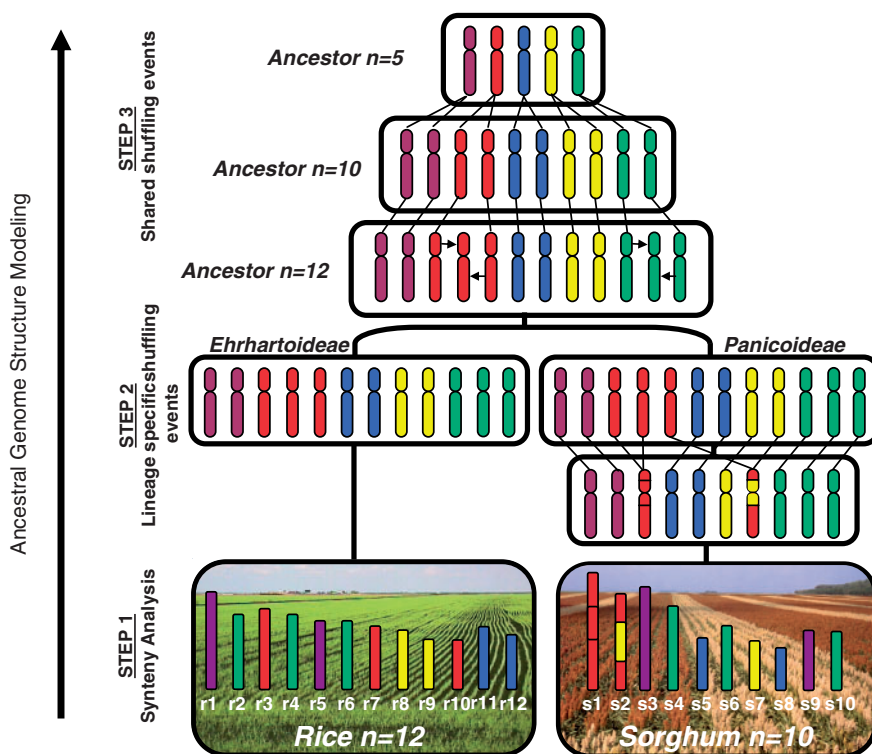


Figure 2: A three-step approach to reconstruct the ancestral genome structure. Chromosomes of the rice and sorghum genomes are represented with a five colour code to illuminate the evolution of segments from a common ancestor with five proto-chromosomes. The number of haploid chromosome (n) is indicated for each species.

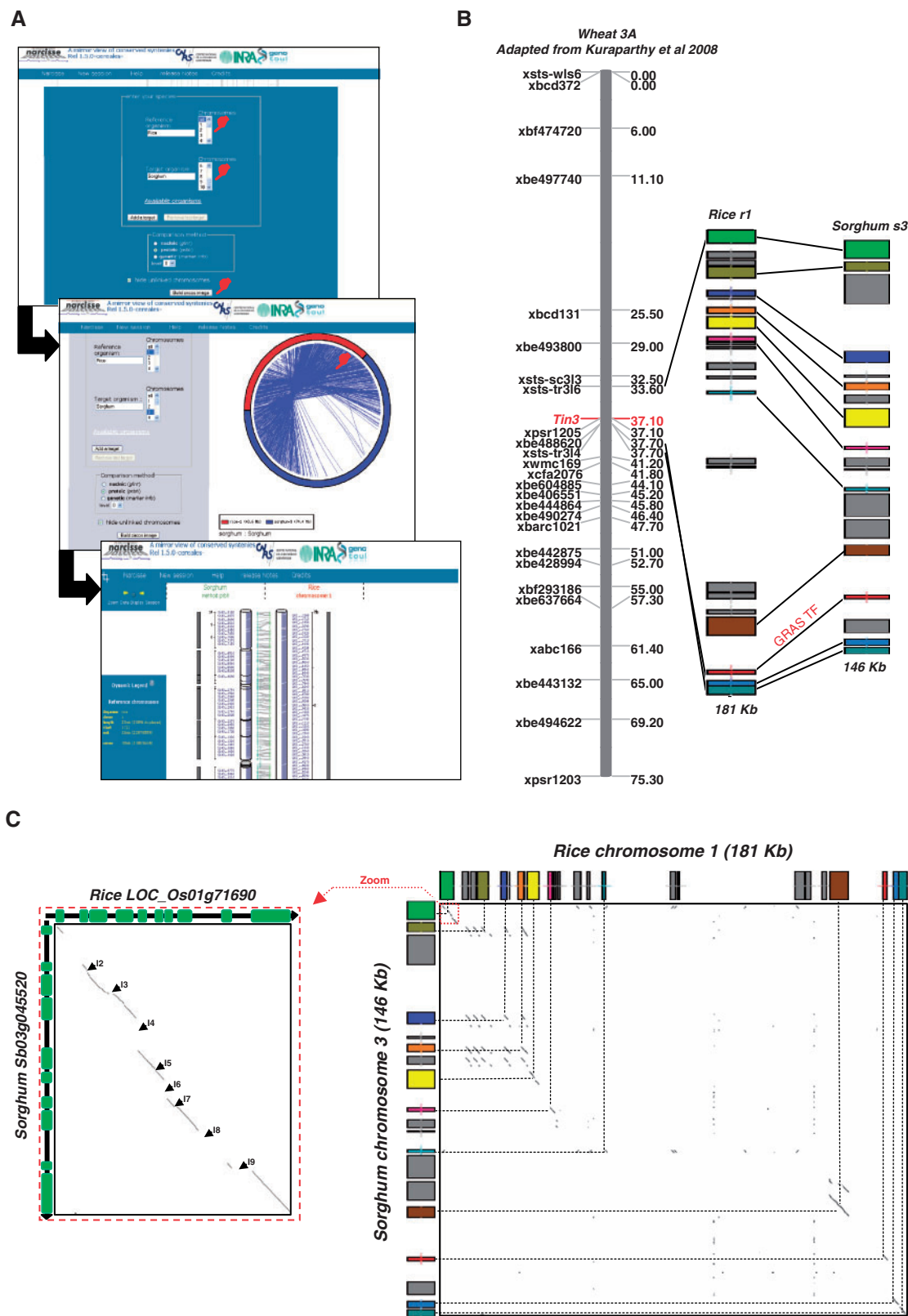


Figure 3: Comparative genomics as a tool for trait dissection and gene annotation. **(A)** Narcisse-Cereal web site output layers—the top level layer displays the entry page in which genomes (Triticeae, rice, maize, sorghum) and specific chromosome can be selected. The intermediate layer represents orthologous gene pairs (link with coloured lines) identified between the selected chromosomes (vertical bars). The bottom layer illustrates the detail information that can be obtained after selecting orthologous relationships on the previous layer. **(B)** Synteny-based

relationships to identify in each genome the chromosomes that share a common ancestral relationship (same colour code in Figure 2 that derive from the dot plot analysis in Figure 1A). In step 2, lineage-specific rearrangements are taken into account to identify an ancestral chromosome structure. In this example, the lineage specific rearrangements observed within the sorghum genome consist of two chromosome fusions that resulted in chromosomes s1 and s2 (red dotted arrows in Figure 1A). In Step 3, shared rearrangement events are considered to model the ancestral genome structure. Here, the seven shared duplications identified between the rice and sorghum genomes lead to an ancestral chromosome set of five proto-chromosomes (Figure 2). Thus, taking into account the five ancestral chromosome groups and the actual rice and sorghum genome structure, our analysis suggests an evolutionary scenario starting with a whole genome duplication (from $n=5$ to 10) followed by conserved segmental duplications (contributing to the construction of two new chromosomes) resulted into a $n=12$ intermediate ancestor with a structure similar to the current rice genome. Rice would have then retained this original chromosome number whereas it has been reduced in the sorghum genome through two chromosomal fusions that resulted in $n=10$ chromosomes Panicoideae ancestor.

COMPARATIVE GENOMICS AS A TOOL FOR IMPROVING TRAIT DISSECTION AND GENOME ANNOTATION

In addition to bringing new insight into genome evolution, the knowledge about the extent of conservation between the cereal genomes and the tools generated through the comparative genomics studies can be used to (i) define efficient strategies for

genetic studies and gene isolation through the design of conserved orthologous markers sets and (ii) improve the accuracy of gene annotation through the alignment of conserved orthologous genes. This is particularly useful for genomes for which no physical map and genome sequences are available yet such as those of the Triticeae.

To support these applications, we have developed an online user friendly interface to access our comparative analyses, called 'Narcisse-Cereals' (Figure 3A), that allows to visualize orthologues as well as paralogues among cereal genomes (<http://www1.clermont.inra.fr/Narcisse-Cereals>, described in details in Salse *et al.* [30]). The web site provide access to the raw data (gene name, sequence, position and alignment criteria) obtained from the synteny and duplications analyses of the rice, maize, sorghum, wheat, and barley genomes. 'Narcisse-Cereals' also provide information about the non-redundant grass genome ancestral gene set that can be used as a platform for the development of Conserved Orthologous Set (COS) markers [46] to support cross genome map-based cloning strategies. This information can greatly increase the success rate of COS marker design because the selection of markers (genes) is not done from one genome only and applied to another with the risk that the locus of interest may have been prone to lineage-specific rearrangements not shared by definition with the target species. 'Narcisse-Cereals' can greatly simplify and accelerate the identification of candidate genes using the synteny information. This is illustrated in Figure 3B with an example taken from the recent literature for a locus affecting the Tiller number in wheat [47]. In this study, the authors characterized a locus involved in tillering inhibition (*Tin3*) in a *Triticum monococcum* mutant, within a 4.1 cM interval flanked by COS markers (STS-TR3L6 and STS-TR3L4) (Figure 3B). Using this information,

identification of a tillering candidate gene using Narcisse-Cereal. Genetic mapping data obtained by Kuraparthi *et al.* [47] and locating the *Tin3* gene within a 4.1 cM interval flanked by two COS markers STS-TR3L6 and STS-TR3L4 are shown on the left-hand side. Orthologous regions that were extracted from the Narcisse-Cereal web site are displayed at the right-hand side of the figure. The *Tin3* candidate gene, (GRAS family transcription factor), conserved between rice (LOC.Os01g71790) and sorghum (Sb03g045660) is indicated in red. (C) Comparative analyses support gene annotation—dot plot alignment between 181 kb of rice chromosome 1 and the 146 kb from sorghum chromosome 3 (right-hand side). The dot plot was carried out using the Dotter programme [37]. Diagonals on the dot plot output represent conserved nucleotide sequences. Horizontal and vertical boxes represent annotated genes respectively on the chromosome 1 and sorghum 3 regions. The zoom (right-hand side inset) represents a dot plot alignment between a rice gene (LOC.Os01g71690) and its orthologous counterpart (Sb03g045520) on the sorghum chromosome 3 region. Green boxes show the conservation of the gene exons structure between the orthologous gene pair.

'Narcisse-Cereals' allows the immediate identification and visualization of the orthologous *Tin3* region in rice on chromosome 1 and sorghum on chromosome 3 and subsequently the definition of a non-redundant list of candidate genes that can be considered as a source of new COS markers. Here, the GRAS family transcription factor conserved between rice (LOC_Os01g71790) and sorghum (Sb03g045660) represents a good candidate for *Tin3* since it is homologous to *LS* a lateral suppressor gene of tomato [48] and *MOC1* [49] of rice, two recently cloned genes involved in lateral branching or tillering in plants.

Finally, the possibility to identify and align genes within and between genomes provides strong support for genome annotation. This has been recently shown by Bossolini *et al.* [50] in a comparative analysis at the *Lr34* locus in wheat, rice and Brachypodium where the comparison between orthologous regions improved the annotation of the rice genome. Figure 3C (right-hand side) represents a dot plot alignment between the 181 kb of rice chromosome 1 and the 146 kb from sorghum chromosome 3, in the orthologous wheat *Tin3* region described above (Figure 3B). Diagonals on the dot plot output represent genes that are conserved between the two chromosome fragments and the magnified dot plot shows an example of the conservation at the gene structure level between the 10 exons of a gene present in the two orthologous regions.

CONCLUSIONS

Many online bioinformatics tools are now available for performing comparative studies and a growing number of genomes are becoming available to increase our knowledge of genome evolution through these analyses. However, it is important to use these tools with caution and be aware that many of them do not use stringent alignment criteria and do not perform statistical validation thereby lowering their reliability in describing evolutionary relationships and mechanisms. As done for other areas of genomics such as functional genomics, standards need to be set up and discussed within the comparative genomics community to ensure homogeneity and the possibility to compare different analyses on the same genomes. Here, we have presented and illustrated with the rice/sorghum genome comparisons the approach that was recently developed

in our group to redefine precisely the orthologous relationships and the shared duplications between the grass genomes and that enabled us to propose a model for the reconstruction of the ancestral genome structure. We believe that the combination of new parameters and new standards for stringent alignment (CIP/CALP parameters) and statistical validation (DR/CR parameters) has improved our capacity to develop comparative genomics studies and provided new insights into the evolution of the grass genomes. The availability of the 'Narcisse-Cereals' web tool should allow other researchers to apply efficiently the same method for additional comparative genomics in the cereals. New sequenced genomes such as the one of Brachypodium which should be released by summer 2009 will be implemented in 'Narcisse-Cereals' providing additional information along the way. Similar analyses will soon be feasible for eudicot genomes (Arabidopsis, grape, poplar, medicago) through a new version of the web site ('Synteny-Plants') currently under development.

Key Points

- Improved alignment parameters were defined to establish significant evolutionary relationships between (grass) genomes.
- New bioinformatics tools have been developed to perform and display comparative genomics analyses and examples are provided using the rice and sorghum genomes as templates.
- The manuscript describes and discusses our approaches and proposes standards that can be used for comparative genomic studies to (i) identify robust sets of orthologous gene pairs, (ii) derive complete sets of chromosome to chromosome relationships within and between genomes and (iii) model common paleo-ancestor genome structures.
- Two applications for accurate gene annotation and synteny-based cloning of agronomically important traits in wheat are described.

SUPPLEMENTARY DATA

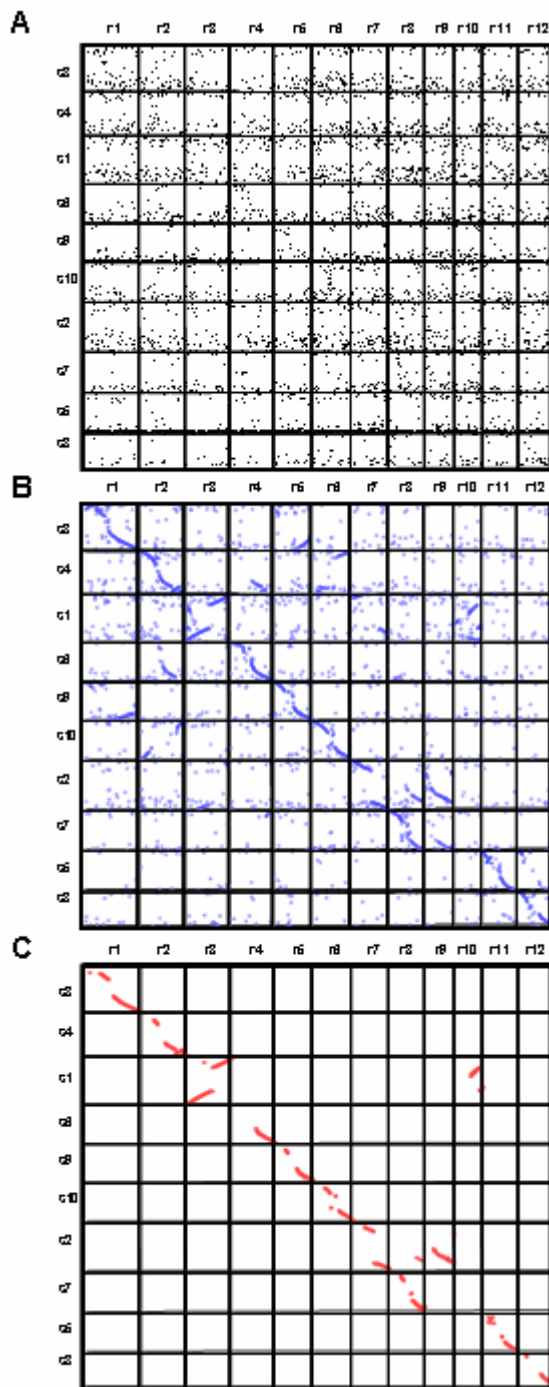
Supplementary data are available online at <http://bib.oxfordjournals.org/>.

References

1. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* 2005;**436**: 793–800.
2. Paterson AH, Bowers JE, Bruggmann R, *et al.* The Sorghum bicolor genome and the diversification of grasses. *Nature* 2009;**457**:551–6.

3. Salse J, Feuillet C. Genomics-Assisted Crop Improvement. In: Varshney RK, Tuberosa R (eds). *Comparative Genomics of Cereals*. Springer: Verlag, 2007, 177–205.
4. Feuillet C, Salse J. Genetics and Genomics of the Triticeae. In: Feuillet C (ed). *Comparative Genomics in the Triticeae*. Springer: Verlag, 2009, 451–77.
5. Li L, Stoeckert CJ, Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 2003;**13**:2178–89.
6. Zeng X, Nesbitt MJ, Pei J, *et al*. OrthoCluster: a new tool for mining synteny blocks and applications in comparative genomics EDBT 2008. 25–30 March 2008, Nantes, France.
7. Fong C, Rohmer L, Radey M, *et al*. PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. *BMC Bioinformatics* 2008;**9**:170.
8. Halling-Brown M, Sansom C, Moss DS, *et al*. A Fugu-Human Genome Synteny Viewer: web software for graphical display and annotation reports of synteny between Fugu genomic sequence and human genes. *Nucleic Acids Res* 2004;**32**:2618–22.
9. Haas BJ, Delcher AL, Wortman JR, *et al*. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 2004;**20**:3643–6.
10. Sinha AU, Meller J. Cinteny: flexible analysis and visualization of synteny and genome rearrangements in multiple organisms. *BMC Bioinformatics* 2007;**8**:82.
11. Brudno M, Poliakov A, Salamov A, *et al*. Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res* 2004;**14**:685–92.
12. Hachiya T, Osana Y, Pependorf K, *et al*. Accurate identification of orthologous segments among multiple genomes. *Bioinformatics* 2009;**25**:853–60.
13. Lyons E, Pedersen B, Kane J, *et al*. Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar, and grape: CoGe with rosids. *Plant Physiol* 2008;**48**:1772–81.
14. Lehmann J, Stadler PF, Prohaska SJ. SynBlast: assisting the analysis of conserved synteny information. *BMC Bioinformatics* 2008;**9**:351.
15. Liang C, Jaiswal P, Hebbard C, *et al*. Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res* 2008;**36**:D947–53.
16. Yuan Q, Ouyang S, Liu J, *et al*. The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res* 2003;**31**:229–33.
17. Conte MG, Gaillard S, Lanau N, *et al*. GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res* 2008;**36**:D991–8.
18. O'Brien KP, Remm M, Sonnhammer EL. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* 2005;**33**:D476–80.
19. Tang H, Wang X, Bowers JE, *et al*. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res* 2008;**18**:1944–54.
20. Crabtree J, Angiuoli SV, Wortman JR, *et al*. Sybil: methods and software for multiple genome comparison and visualization. *Methods Mol Biol* 2007;**408**:93–108.
21. Derrien T, André C, Galibert F, *et al*. AutoGRAPH: an interactive web server for automating and visualizing comparative genome maps. *Bioinformatics* 2007;**23**:498–9.
22. Rebeiz M, Posakony JW. GenePalette: a universal software tool for genome sequence visualization and analysis. *Dev Biol* 2004;**271**:431–8.
23. Wang H, Su Y, Mackey AJ, *et al*. SynView: a GBrowse-compatible approach to visualizing comparative genome data. *Bioinformatics* 2006;**22**:2308–9.
24. Pan X, Stein L, Brendel V. SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics* 2005;**21**:3461–8.
25. Morgenstern B. Alignment of genomic sequences using DIALIGN. *Methods Mol Biol* 2007;**395**:195–204.
26. Delcher AL, Salzberg SL, Phillippy AM. Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* 2003;**10**:10–3.
27. Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, *et al*. GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics* 2008;**9**:376.
28. Salse J, Bolot S, Throude M, *et al*. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* 2008;**20**:11–24.
29. Bolot S, Abrouk M, Masood Quraishi U, *et al*. The 'inner circle' of the cereal genomes. *Curr Opin Plant Biol* 2008;**12**:1–7.
30. Salse J, Abrouk M, Bolot S, *et al*. Reconstruction of monocotyledonous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci USA* 2009, doi:10.1073/pnas.0902350106.
31. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
32. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
33. Hampson S, McLysaght A, Gaut B, *et al*. LineUp: statistical detection of chromosomal homology with application to plant comparative genomics. *Genome Res* 2003;**13**:1–12.
34. Vandepoele K, Saeys Y, Simillion C, *et al*. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between Arabidopsis and rice. *Genome Res* 2002;**12**:1792–801.
35. Calabrese PP, Chakravarty S, Vision TJ. Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* 2003;**19**:74–80.
36. Hampson SE, Gaut BS, Baldi P. Statistical detection of chromosomal homology using shared-gene density alone. *Bioinformatics* 2005;**21**:1339–48.
37. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 1995;**167**:1–10.
38. Kumar S, Mohan A, Balyan HS, Gupta PK. Orthology between genomes of Brachypodium, wheat and rice. *BMC Res Notes* 2009;**2**:93.
39. Murphy WJ, Pevzner PA, O'Brien SJ. Mammalian phylogenomics comes of age. *Trends Genet* 2004;**20**:631–9.
40. Ferguson-Smith MA, Trifonov V. Mammalian karyotype evolution. *Nat Rev* 2007;**8**:950–62.
41. Lopez Rascol V, Pontarotti P, Levasseur A. Ancestral animal genomes reconstruction. *Curr Opin Immunol* 2007;**19**:542–6.
42. Dobigny G, Ducroz JF, Robinson TJ, *et al*. Cytogenetics and cladistics. *Syst Biol* 2004;**53**:470–84.

43. Tesler G. GRIMM: genome rearrangements web server. *Bioinformatics* 2002;**18**:492–3.
44. Bourque G, Pevzner PA, Tesler G. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. *Genome Res* 2004;**14**: 507–16.
45. Ma J, Zhang L, Suh BB, *et al.* Reconstructing contiguous regions of an ancestral genome. *Genome Res* 2006;**16**: 1557–65.
46. Masood Quraishi U, Abrouk M, Bolot S, *et al.* Comparative genomics in cereals: from genome-wide conserved orthologous set (COS) sequences to candidate genes for trait dissection. *Funct Integr Genomics* 2009, doi:10.1007/s10142-009-0129-8.
47. Kuraparthi V, Sood S, Gill BS. Genomic targeting and mapping of tiller inhibition gene (*tin3*) of wheat using ESTs and synteny with rice. *Funct Integr Genomics* 2008;**8**:33–42.
48. Schumacher K, Schmitt T, Rossberg M, *et al.* The Lateral suppressor (*Ls*) gene of tomato encodes a new member of the VHIID protein family. *Proc Natl Acad Sci USA* 1999; **96**:290–5.
49. Lu F, Ammiraju JS, Sanyal A, *et al.* Comparative sequence analysis of MONOCULM1-orthologous regions in 14 *Oryza* genomes. *Proc Natl Acad Sci USA* 2009;**106**:2071–6.
50. Bossolini E, Wicker T, Knobel PA, *et al.* Comparison of orthologous loci from small grass genomes *Brachypodium* and rice: implications for wheat genomics and grass genome annotation. *Plant J* 2007;**49**:704–17.



Supplementary Information: Comparative genomics analyses between rice and sorghum genomes to illustrate the importance of applying stringent alignment parameters and statistical validation. A dot plot alignment of the rice (r1-r12, horizontal) and sorghum genome (s1-s10, vertical) sequences was performed using the DOTTER program. In this representation, diagonals reflect gene conservation between the two genome sequences whereas interruptions indicate loss of colinearity. Gene conservations deduced through (A) BLASTn with a cut off value of $1e-40$ (grey dots) followed by (B) 60% CIP/70% CALP parameter thresholds (blue dots) followed by (C) DR/CR statistical validation (red dots) are illustrated.