A computational pipeline for the development of comparative anchor tagged sequence (CATS) markers

L. Schauser[1], J. Fredslund[1], L. Heegaard Madsen[2], N. Sandal[2] and J. Stougaard[2]
[1]*Bioinformatics Research Center, University of Aarhus, Ny Munkegade, Bldg. 540 8000 Aarhus Denmark*
[2]*Laboratory of Gene expression, Department of Molecular Biology, Aarhus University, Gustav Wieds Vej 10, DK-8000 Aarhus C, Denmark, Email: schauser@daimi.au.dk*

Key points:

1. Molecular markers that allow the transfer of map information from one species to another are vital in comparative genetics.
2. To identify potential anchor marker sequences more efficiently, we have established a bioinformatic pipeline that combines multi-species EST- and genome- sequence data.
3. Taking advantage of information from a few related species, comparative EST sequence analysis identifies evolutionary conserved sequences in less well-characterised species in the same family.
4. Alignment of evolutionary conserved EST sequences with corresponding genomic sequences defines sets of PCR primer sites flanking introns.
5. Markers identified by this procedure will be readily transferable to other species since they are selected on the basis of their common evolutionary origin.
6. We exemplify our procedure on legumes and grasses, where model plant studies and the genome- and EST-sequence data available have a potential impact on breeding crop species.

Keywords: bioinformatics, expressed sequence tags, molecular markers, comparative anchor tagged sequences, polymorphism ascertainment

Introduction

Precise comparison of plant gene maps requires common anchor loci as landmarks for the alignment of conserved chromosomal segments. With the completion of the genomic sequences of Arabidopsis and rice, and large collections of ESTs at hand, comparative genome mapping carries the promise for rapid increase in knowledge about the large and repetitive genomes of many crop species. A common observation is conservation of linkage organization of homologous genes in species from diverse plant lineages. Comparative genome mapping allows the transfer of knowledge from one species to another related species. This information transfer can go two ways: (i) from well characterized model species with detailed genetic maps and / or complete genome sequence information to large genome crop species which are the target of breeding programs, and (ii) from the crop where quantitative trait loci (QTL) have been mapped, to a relevant model species where the gene content of this region is known. Map comparisons are hampered by the fact that genomes are not static in their arrangement, but often undergo chromosomal rearrangements, such as inversions, translocations, duplications, deletions and cycles of polyploidization followed by diploidization. Plants, given their sexual promiscuity and potential for vegetative reproduction are particularly prone to genome rearrangements (Bennetzen, 2000). For example, whole genome duplications have occurred at several occasions during the evolution of modern plant species (Paterson *et al.,* 2004). In the diploid phase, members of a duplicated gene pair are retained or deleted at random in the two duplicated regions, obscuring the common past. This process results in problems with congruency between two genomes that are separated by a polyploidization-diplodization cycle. Hence, in order to succeed, any attempt of comparative genome mapping must carefully choose the species of comparison.

A central step in genome comparisons is the identification of sets of sequences that can readily be identified in the genomes of the species to be compared, and serve as "anchors" of their respective genetic maps. Commonly used markers, such as microsatellite or AFLP markers, can give high resolution genetic maps, but they are of little comparative value because they are not conserved across several species. Anchor sequences should be chosen such that they maximize the potential to serve as markers in several species, and also maximize congruency between the genetic maps of the organisms. Previous comparative maps haves relied on hybridisation of homologous probes and their scoring as RFLP markers (Fulton *et al*., 2002, Draye *et al*., 2001). Hybridization markers are time consuming, labour intensive and often involve the handling of radioactivity. Furthermore, it is not easy to generate specific hybridization markers, as they often cross-hybridize to other genomic regions. PCR based markers are much more efficient, as they are amenable to high throughput automation and, if well designed, of high specificity. Towards this goal we employ a strategy based on the identification of single copy number evolutionary conserved sequences within transcriptomes of representative species of the lineage under study. These sequences are used as PCR primer annealing sites for the amplification of intervening intronic sequences that are subject to subsequent polymorphism discovery. This approach ensures that unique, gene rich regions of the genome are the primary target of this effort.

Our bioinformatic approach is based on differences in the evolutionary rate of DNA changes in a genome. During evolution, many functional sequences, such as coding regions and regulatory elements, are under strong purifying selection. In contrast, intron sequences are less constrained and will display a higher degree of mutational variation between any two ecotypes / varieties. Although the evolutionary constraints on the exact sequence of the intron are relaxed, the position and approximate length of the intron is usually conserved, even over long evolutionary distances (Roy *et al*., 2003). An automated primer-finding algorithm proposes primers pairs in regions of high conservation for the PCR amplification of intron sequences that have a high probability of capturing polymorphisms between varieties of any species within the clade under study. Subsequent sequencing of intron-spanning PCR products in mapping parents will reveal the presence of any polymorphism that can be used as a molecular marker. Here we present an automated pipeline for the generation of CATS and apply it to two plant lineages of major interest to agriculture: grasses and legumes.

Methods

The algorithm designed to identify conserved anchor tagged sequences (CATS, Lyon *et al.*, 1997) is best illustrated as a succession of three comparative filters and a primer-finding step. *1) Identifying expressed evolutionary conserved sequences (ECS) from different plant species*. Regions of strong homology between collections of ESTs from different species were identified. *2) Counting copy numbers in the Arabidopsis / rice proteome*. In order to avoid gene families and to get a score for the information content of an ECS, we counted the number of highly homologous sequences in the Arabidopsis /rice proteome. *3) ECS-genome alignment.* The presence and length of introns in reference genomes in scored at this step. Introns are highly conserved features, even among distant species. *4) Primer design* Multiple alignments of ECSs with indication of intron position are generated and primers are designed using this alignment as a guide. The order of application of the comparative filters does not influence the results and should hence be organized in a way that minimizes the computational cost.

Sequences: The EST clusters used for this analysis were retrieved from the Institute of Genome Research (TIGR). We downloaded the gene indices (clustered EST collections,

Quackenbush *et al*., 2000, Pertea *et al*., 2003) for legumes (*Lotus japonicus*, *Medicago truncatula* and *Glycine max*) and selected grasses (*Hordeum vulgare*, and *Sorghum bicolor*).

The Arabidopsis and rice (*Oryza sativa*) genome, proteome and coding sequences were downloaded from the TIGR FTP site. The *Lotus japonicus* and *Medicago truncatula* genomic sequences were retrieved using NCBIs ENTREZ.

The Blast package (Altschul *et al.,* 1997) was obtained from the NCBI. For comparison of nucleotide sequences, we used the megablast program with a wordsize of 20 and cutoff e-value of 2e-40. For DNA-protein comparisons, we used the blastx program (e-value 10e10-6). A series of Python scripts were generated to parse the Blast outputs and assemble sequence collections of ECS. Multiple alignments were generated by ClustalW (Chenna *et al*., 2003), and automated primer design was achieved through application of the PriFi program (Fredslund *et al.,* manuscript in preparation).

DNA extraction, PCR amplification and sequencing of amplicons were performed using standard laboratory protocols.

Results

We modified the CATS algorithm (Lyons *et al*., 1997) to reduce the potential pitfalls induced by gene families and paralogous copies. The filtering of sequences is divided into four operational steps and is best illustrated as a pipeline adding consecutive comparative selection criteria (Figure 1). The filters can be applied in any order without affecting the result.

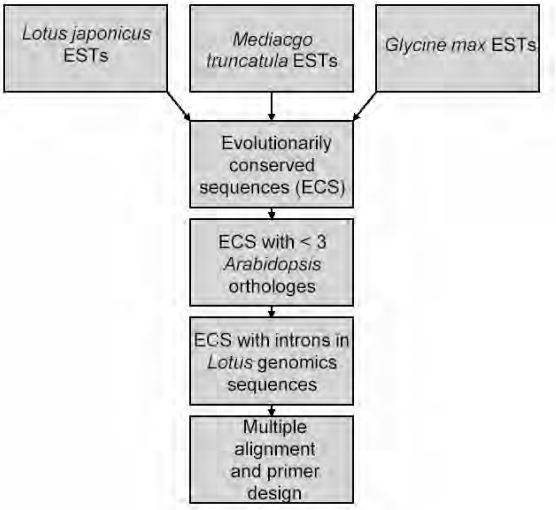

Figure 1 Pipeline of the marker candidate algorithm exemplified in legumes. In the first step, EST collections of selected species are compared. Evolutionary conserved sequences are passed on to the next step. Here the number of sequences with homology to the *Arabidopsis* reference proteome is estimated. Sequences with one or two homologues in the *Arabidopsis* proteome are considered because Arabidopsis has undergone a recent whole genome duplication. ECSs passing this criterion are compared to *Lotus* and *Medicago* genomic sequences and ranked according to overall length of the ECS and optimal length of introns. The ECSs are multiply aligned and primers are designed using this alignment as input.

Selecting species for the comparative approach in legumes and grasses

Our aim is to exploit colinearity between genomes of species with dense genetic maps and crops with important agronomic traits. In plants this colinearity erodes rather fast with phylogenetic distance. It is therefore crucial to choose the resources that allow maximal information transfer between species. Parameters that we considered include the amount of EST information and their phylogenetic relationship. For legumes, the resources originate from *Lotus japonicus*, *Medicago truncatula* and *Glycine max*. Their phylogenetic relationship is depicted in figure 2a. For the grasses we chose the species *Hordeum vulgare*, *Oryza sativa* and *Sorghum bicolor* (figure 2b). The CATS primers developed by our pipeline should amplify PCR products in all species within these clades. They may also be relevant to species outside these clades, such as peanut (*Arachis*) for legumes and banana (*Musa)* for grasses (not shown).

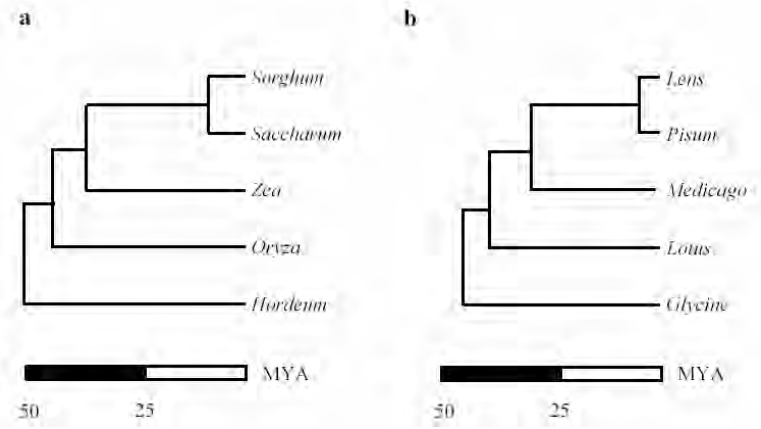

Figure 2 Phylogenetic relationship of the species in a) the grasses and b) the legumes. Species with sequence information used in this study together with selected other species are shown.

Selecting genes and primer design

Filter 1. Identifying expressed ECS from different representative species of the same lineage. We have here used the ready clustered EST collections downloadable from TIGR (gene indices) as input (Table 1), but any assembled collection of ESTs could serve as entry points. Stringent comparisons between different gene indices (using megablast with wordsize 20 and a cutoff e-value of 2e-40) revealed those sequences that display a high degree of conservation. ECSs represent exons that have been under strong purifying selection during evolution i.e. they display a higher-than-average conservation between species.

Table 1 TIGR Gene indices for the species used in this study

	Hordeum Vulgare	*Sorghum bicolor*	*Medicago truncatula*	*Lotus japonicus*	*Glycine max*
Number of GIs	50,453	39,148	36,976	28,460	63,676

Filter 2. Counting copy numbers in the proteome of a reference species (Arabidopsis for legumes, rice for grasses). In order to get a score for the information content of an ECS, we counted the number of highly homologous sequences encoded by the ECS in the reference species proteome. Repeated sequences are not useful for mapping purposes, since polymorphisms might reflect paralogous origin rather than allelic variation. Furthermore, allelic variation at a candidate marker locus can be partially or completely masked by the presence of paralogues, reducing the information content of this marker. Several rounds of genome duplication and gene family amplification have occurred prior to the split between Leguminosae (Rosid I) and Brassicacae (Rosid II), and also prior to the diversification of the grasses (Paterson *et al*., 2004). The Arabidopsis genome has been subject to at least one round of duplication since the Rosid divide. The diploid legume species *Lotus japonicus* and *Medicago truncatula* do not seem to have undergone a similar duplication (Bowers *et al.,* 2003). Therefore counts of Arabidopsis genes can be taken as an overestimate of the legume count and we allow for two homologues in our pipeline. Some grasses (*Zea mays* and others) have undergone rounds of whole genome duplications, and care must be taken when assessing co-linearity of the genetic maps.

Filter 3. ECS-genome alignment. In order to maximize chances of detecting polymorphisms at later steps, introns interrupting a given ECS are scored by aligning the ECSs with corresponding genomic sequences. For grasses, we assess the presence of introns by comparing rice ECS sequences with the rice genome. For legumes we make use of the *Lotus japonicus* and *Medicago truncatula* genome sequences, but this could easily be extended to the Arabidopsis genomic sequence. This would still be informative, since the presence, location and approximate length of introns are highly conserved features, even among distantly related species (Roy *et al*., 2003). We also score the length of introns. This quantity is of interest for two reasons: (i) short introns are less likely to be polymorphic than longer ones, and thus longer introns are of interest and (ii) the final PCR reaction using degenerate primers is limited to the maximum amplicon size of ~3 kb using standard polymerases.

Filter 4. Primer design. Multiple alignments of ECSs with indication of intron position are generated and forward and reverse primers are designed using this alignment as a guide. A number of criteria are scored which have to do with the number of species in the alignment the melting temperature and GC content of the proposed primer and the length of the intron(s) which separates two primers. A conservation score reflects the degree of similarity between the most evolutionarily divergent species in the alignment. Finally a score is given for the distance from primer site to the exon-intron junction. This score is introduced as a means of selecting primers that allow the identification of the PCR product as being derived from a homologous locus in a subsequent sequencing step. A combined score for each primer pair allows their comparison and ranking within and between candidate regions.

Legumes

The collections of gene indices (preclustered EST collections) for Legumes were downloaded from TIGR (Table 2). In order to estimate the number of homologues in Arabidopsis, these sequences were compared to the proteome of Arabidopsis. Since Arabidopsis has undergone a recent whole genome duplication, we considered sequences with both one and two hits in the Arabidopsis proteome (Table 2). Most gene indices have several (>2) Arabidopsis homologues. Next, we tested for the presence of introns in the corresponding genomic regions of *Lotus* (122 Mbp of genomic sequence) and *Medicago* (143 Mbp of genomic sequence). If no corresponding genomic sequence was identified, we ignored the gene index. The fractions

of gene indices which have genomic regions sequenced is slightly higher in *Lotus* than in *Medicago*, indicating that the *Lotus* genome project covers more genes than the *Medicago* genome project (18% vs.13%)

Table 2 Information content of the collections of gene indices. The species name is followed by the Release version (in parenthesis), and the number of gene indices (clustered ESTs and singleton ESTs) are indicated for each species. These sequences are binned according to the count of homologous genes in Arabidopsis. The numbers in parenthesis indicate the size of the subset with an intron in the respective genomic sequence: *Lotus* GIs were compared to *Lotus* genomic sequences, whereas *Medicago* GIs were compared to *Medicago* genomic sequences.

	Lotus japonicus (v. 3.0)	*Medicago truncatula* (v. 7.0)	*Glycine max* (v. 12.0)
Number of gene indices	28,460	36,976	63,676
One Arabidopsis homolologe	2,282 (394)	3,088 (397)	4,281
Two Arabidopsis homologes	1,606 (306)	2,151 (265)	3,349

Next, we compared the gene indices with intron information to the other EST collections. It is striking, that if a sequence is found in both *Lotus* and *Medicago*, it has a very high probability of being present in Glycine as well. There was about 25% redundancy between the sequences identified through *Lotus* and *Medicago* genomic information (introns), reflecting the unfinished state of the two genomes. These collections of three sequences were the basis for a ClustalW multiple sequence alignment followed by our CATS primer finding algorithm.

Table 3 Comparative CATS identification. Sequences with introns and one or two Arabidopsis homologues were successively compared to the EST collections of the other legumes, generating sets of three sequences. These sets of sequences were the basis for ClustalW multiple sequence alignment and an automated CATS primer finding algorithm.

Query sequences:	Compared to:		Number of CATS primer pairs identified
Lotus japonicus	*Medicago* only	*Medicago* & *Glycine*	
GIs with one Arabidopsis homologue and intron information	288	269	48
GIs with two Arabidopsis homologues and intron information	186	166	22
Medicago truncatula	*Lotus* only	*Lotus* and *Glycine*	
GIs with one Arabidopsis homologue and intron information	220	207 (57)	22
GIs with two Arabidopsis homologues and intron information	128	118 (27)	18

Grasses

When the pipeline was applied to the grasses by simply exchanging all the relevant data files, we were able to identify 1335 CATS primer pairs. The selected gene indices originated from *Hordeum vulgare* (Release 9.0) and *Sorghum bicolor* (Release 8.0). These were compared to the rice (*Oryza sativa*) genome and annotated CDS.

Testing CATS primers

We tested the potential of our pipeline to generate CATS markers by randomly choosing 36 of the legume CATS primer sets and attempting to develop them as markers in the legumes *Phaseolus vulgare* (common bean) and *Arachis* (Peanut spp.). 70 % of these primer sets amplified the correct product in the relatively closely related *P. vulgare*, whereas this figure dropped to 62 % in the outgroup species *Arachis.* Of these, up to 90% were polymorphic, depending on the mapping parents used.

Discussion

In this presentation we exploit evolutionary conserved sequences for developing molecular markers useful as anchors when comparing genetic maps of different species. Our goal is to use these sequences as conserved and unique sites for primer annealing. A pair of such sites can then be used for amplifying intervening intronic sequences that subsequently can be scanned for polymorphisms distinguishing breeding varieties or ecotypes. We have shown that our automated bioinformatic algorithm is a versatile tool allowing the quick generation of marker candidates useful for map construction projects in legumes and grasses and by extension, to any phylogenetic clade with appropriate comparative sequence information.

Since only unique sequences are useful as markers, we are interested in the number of paralogous sequences in the genome. An approximation to this number is obtained by counting homologous sequences in the proteome of a reference species. Strictly speaking, we are not able to discern between orthologous and paralogous origin of homologous sequences. However, for those sequences with only one homologue in the Arabidopsis / rice genome, we can reasonably assume orthology. Although this criterion maximises congruency when comparing maps, it by no means guarantees it. Both clades studied here have a common ancestor that at some point has undergone a whole-genome duplication (Paterson *et al.,* 2004), potentially obstructing colinearity through differential gene loss. The degree of microsynteny depends on the timing of the duplication event relative to the most recent ancestor of the clade. In any case, our selection filters out genes that are prone to duplication and hence are members of gene families.

A main application of this algorithm is the transfer of genome information between model species and closely related large genome crops. In plant breeding programs traits of economic importance are screened out of large populations. Breeders introduce variation through crosses between varieties and in some cases also wild relatives but there is rarely a simple method for following the segregation of the trait or allele of interest. Instead of screening for the traits per se, which can be difficult to score due to environmental conditions, late onset or small contributions to the phenotype, breeders often use linked markers as indicators of inheritance. For this purpose, molecular markers are best suited, since they can be co-dominant, cheap and readily scored. Dense genetic maps spanning all linkage groups are of invaluable help for breeding purposes. If dense genetic maps are not available for the species

at hand, comparison with other species can help in the development of new markers and qualified guesses at candidate genes in the region under investigation. It is therefore generally advisable to initially map markers that can serve as anchors, connecting genetic maps of as many species as possible. The phylogenetic distances that limit such an approach should be considered. Most macrosyntenic information is lost between evolutionary diverse lineages. This information loss is dramatically enhanced when a whole genome duplication event has occurred in one of the lineages. For example, no macrosynteny is recognizable between *Medicago truncatula* and Arabidopsis (Zhu *et al*., 2003). On the other hand, microsynteny is generally much better conserved (Zhu *et al.,* 2003, see also Krusell *et al*., 2002). Within a given clade, such as legumes or grasses, both micro- and macrosynteny are well-conserved (Choi *et al.,* 2004, Bennetzen, 2000, Draye *et al*., 2001).

As for any marker, the success of applying our pipeline depends on the variation between any two varieties used for mapping. We have shown that the pipline produces valid marker candidates when applied to outgroups of the clade under consideration, as in the *Arachis* example. For related legumes the pipeline should be of value as a tool to bridge the genetic maps of model and crop legumes. The density of CATS is not very high in legumes. This could be changed by lowering the requirements to consider pairwise comparisons, instead of comparisons between three species. Another improvement could be gained when looking for introns. Here, we have here only exploited the incomplete genome sequence information of *Lotus japonicus* and *Medicago truncatula*. Given the observed conservation of intron positions it should even be possible to use the Arabidopsis genome as a reference, a strategy which has recently been employed by Choi *et al.* (2004). Thus we should be able to generate more than a thousand CATS marker candidates for any legume cross in the near future. The comparative approach described here is broadly applicable to all EST resources collected from species with appropriate phylogenetic distance and reference genome information. The phylogenetic distance and the amount of sequence information for the species chosen will determine success. For grasses, we found 1335 CATS primer pairs, illustrating this point. When developed as markers and mapped in several species, these could add considerable density to existing comparative mapping databases such as Gramene (Ware *et al*., 2002).

Conclusion

Our automated bioinformatic pipeline for the generation of CATS is an efficient approach to generating anchor markers, allowing rapid information transfer between traits of interest to the breeders and the dense genetic maps of model plants.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-402.

Bennetzen JL. (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell. 12:1021-9.

Bowers JE, Chapman BA, Rong J, Paterson AH. (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 422(6930):433-8.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31:3497-500.

Choi HK, Mun JH, Kim DJ, Zhu H, Baek JM, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, Young ND, Cook DR. (2004) Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci U S A. 101:15289-94.

Draye X, Lin YR, Qian XY, Bowers JE, Burow GB, Morrell PL, Peterson DG, Presting GG, Ren SX, Wing RA,
Fedorov A, Roy S, Fedorova L, Gilbert W. (2003) Mystery of intron gain. Genome Res. 13:2236-41

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 14:1457-67.
Krusell L, Madsen LH, Sato S, Aubert G, Genua A, Szczyglowski K, Duc G, Kaneko T, Tabata S, de Bruijn F, Pajuelo E, Sandal N, Stougaard J. (2002) Shoot control of root development and nodulation is mediated by a receptor-like kinase. Nature 420:422-6.
Lyons, L. A. T. F., Laughlin, N. G. Copeland, N. A. Jenkins, J. E. Womack, S. J. O'Brien (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. Nature Genetics 15, 47 - 56.
Paterson AH. (2001) Toward integration of comparative genetic, physical, diversity, and cytomolecular maps for grasses and grains, using the sorghum genome as a foundation.
Plant Physiol. 125:1325-41.
Paterson, AH., J. E. Bowers and B. A. Chapman (2004) Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. PNAS 101, 9903-9908
Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics. 19:651-2.
Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. 29:159-64.
Zhu H, Kim DJ, Baek JM, Choi HK, Ellis LC, Kuester H, McCombie WR, Peng HM, Cook DR. (2003) Syntenic relationships between *Medicago truncatula* and Arabidopsis reveal extensive divergence of genome organization. Plant Physiol. 131:1018-26.
Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, Teytelman L, Schmidt SC, Zhao W, Chang K, Cartinhour S, Stein LD, McCouch SR. (2002) Gramene, a tool for grass genomics. Plant Physiol. 130:1606-13.