Comparative mapping

Andrea Seres[*], Gábor Deák[*], Gábor Tóth[*], Grégoire Aubert[+], Judith Burstin[+], Noel Ellis[#], György B. Kiss[*]

[*]Agricultural Biotechnology Center, Gödöllő, Hungary
[#]John Innes Center, Norwich, United Kingdom
[+]URLEG INRA, 21110 Bretenières, France

Table of contents:

Abstract

Comparative genetic mapping studies reveal similarities and differences in gene content and gene order between genera belonging to different taxa. A pressing need in legume genomics is to integrate knowledge gained from the study of model legume genomes with the biological and agronomic questions of importance in the crop species. It is important to know whether similar features are prevalent in other plant families, in particular because the extent of such differences may define the limits of comparative structural genomics as a strategy for applied agriculture. We provide in the following section protocols for marker development in genes, genotyping and mapping in *M. tr*. and other legume species.

1. Selection of parental plants and development of mapping populations

As a first step for the establishment of a linkage map the parental individuals of a mapping population have to be selected. During selection of the parental plants one must consider that plants which are more distantly related are more likely to result in a highly polymorphic mapping population (e. g. interspecific mapping population), facilitating the later positioning of the markers on the genetic map (in contradiction with the intraspecific mapping populations). It is also convenient if the candidate parents present morphological differences (e.g. different flower color, difference in height), which can provide easy means of ensuring that the F_1 progenies resulted from cross pollination.

The selection of diploid parental plants will foster the genetic studies performed later on the mapping population (e.g. diploid *Medicago truncatula*, diploid *Medicago sativa* ssp. *quasifalcata* and ssp. *cerulea*). After selecting the parental plants, different kinds of mapping populations can be generated depending on the number of generations created after F_1. They can be F_2 mapping populations, RIL (**R**ecombinant **I**nbreed **L**ine) populations, back-cross populations and NIL (**N**early **I**sogenic **L**ine) populations each having particularities which are useful in different situations for different purposes.

2. **Genomic DNA isolation**:

2.1. Homogenizing plant tissue

- Harvest different organs from the plant: roots, leaves, floral organs (to obtain large amount of DNA it is preferable to use young leaves).
- Harvest the organs, put them in labeled paper bags and place on regular ice during harvest to make sure that they are kept cool.
- For immediate grinding: place leaf sample in a mortar or some other type of container able to hold liquid nitrogen. Quick-freeze samples in liquid nitrogen. Once frozen do not allow samples to thaw until isolation!
- Grind the tissue samples in the mortar and pestle in liquid nitrogen. Make sure to pre-chill the mortar and pestle. Grind the samples as finely as possible. Trying to grind large amounts of tissue will result in coarse grinding and will greatly reduce the yields.
- If the harvested samples are not processed immediately, they can be stored at -80ºC in air tight bags for a few days.
- For later DNA isolation, the harvested plant organs can be lyophilized and stored in a labeled paper envelop.

2.2. DNA isolation from grinded tissue

Various ready-to-use isolation kits can be purchased and the provided protocols followed to obtain pure DNA. If no such kits are used, several convenient protocols can be followed to obtain pure DNA.

Isolation using CTAB extraction buffer:

- Collect the fresh plant material (20-50 ng) in 1.5 ml eppendorf tubes.
- Add 100 mg quartz sand. Homogenize the fresh plant tissue with the sand then add 300 µl CEP buffer, than vortex the mixture shortly.
- Incubate for 60 min, with continuous gentle rocking at 65ºC. Do not exceed 75 minutes as DNA yield will be compromised.
- Remove tubes from incubation and add 600µl chloroform. Rock gently the samples to mix, for 30 min at room temperature.
- Spin down the samples in a centrifuge for 10 min at 13000 RPM at room temperature
- Collect the supernatant in 520 µl isopropanol (2-propanol), mix gently and keep at -80ºC for 20 minutes.
- Spin down the samples and remove the isopropanol, repeat the procedure and resolve the pellet in 150 µl 0, 1 mg/ml RNase.
- Incubate the samples for 2 hours at 37ºC.
- NH_4Ac/SDS cleaning: add 150 µl TES to the 150 µl DNA, vortex it and incubate 10 minutes at room temperature; add 150 µl NH_4Ac [7,5M] vortex it and incubate for 10 minutes at -20º then centrifuge for 5 minutes at 13000 RMP. Collect the supernatant in 750 µl ethanol, vortex it and incubate at -20ºC for 20 minutes. Centrifuge the samples for 5 minutes with 13000 RMP, remove the supernatant, repeat the centrifugation and remove carefully all the supernatant then dry the DNA pellet.
- Resolve the pellet in 1x TE, double as many mg plant tissue was initially used to isolate DNA.

CEP:
2% CTAB
100 mM Tris
20 mM EDTA
1, 4 M NaCl
0.5% ß-mercaptoethanol

TES:
10 mM Tris (pH 7.5)
1 mM EDTA
1% SDS

1x TE (10:1)
Tris HCl 100 mM (pH 7.6)
EDTA 10 mM

3. **Quantification of the extracted DNA**

3.1. Quantification using the Spectrophotometer

- Calibration of the machine as described in the user manual
- Use water or TE as a reference (or blank). Aliquot 1ml of either water or TE into the cuvette and load the spectrophotometer.
- Use diluted sample for measuring as a 1:100
- The readings should be don at A260 and A280.

DNA concentration (µg/ml) = 50 x A260 x Dilution Factor
RNA concentration (µg/ml) = 40 x A260 x Dilution Factor
260/280 ratio = 1.6 ~1.8 – absorption due to DNA
= 1.6 or less – Protein contamination
= 2.0 or more – Chloroform or phenol contamination.

3.2**. DNA concentration can be approximated** using mass rulers with known concentration of each migration bands. DNA samples are run next to mass rulers, on agarose gel and the concentration can be estimate according to the ladder description

4. Gene specific primer design for PCR amplification

4.1 The intron targeting method

An efficient strategy to generate gene-specific markers for mapping in plants is the **I**ntron **T**argeting (IT) method. IT primer pairs are complementary to the sequences of the exons flanking the targeted intron. Since the targeted intron sequence is generally less conserved than the exons, the amplified product may display polymorphism due to length/nucleotide variation among introns in the alleles of the gene. On the other hand, the higher level of sequence conservation in the exons ensures that all alleles can be effectively amplified. If a single targeted intron is too short, primers may be designed to match exons flanking two introns and an internal exon, thereby fostering the detection of length polymorphism.

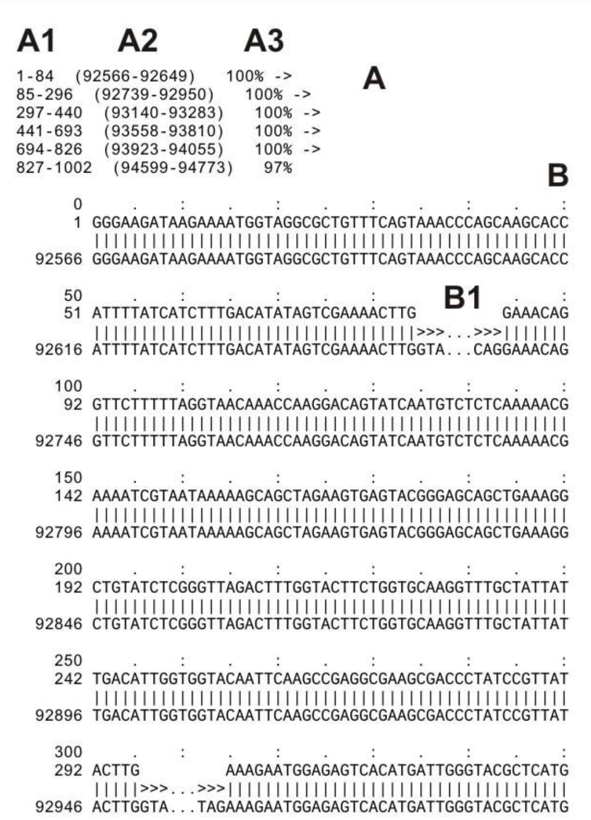
The prerequisite of the method is that the genomic region harboring the gene is sequenced and mRNA, assembled EST consensus or at least EST sequences also exist. EST consensi can be obtained from TIGR: the Gene Index databases contain such Tentative Consensus (TC) sequences (ftp://ftp.tigr.org/pub/data/tgi/). Throughout this chapter, mRNA, TC or EST sequences will be referred to as cDNA sequences. Here we describe a computational method to design primer sequences that can be used to generate IT markers for mapping studies. All software programs are freely available and can easily be built into a processing pipeline. We also present a web application built to carry out the last few steps of the process, i.e. the exon selection and the primer design steps.

4.1.1. Design of intron targeting primers

The method used to find the intron to be targeted and its flanking exons depends on whether cDNA and genomic sequences are both available from the same species (in this case in *M. truncatula*) or not. If the genomic region is not sequenced yet, the sequence of a homologous (preferably orthologous) gene from another species can be of help.

cDNA and genomic sequences exist from the same species

The method described here is based on the cDNA sequence of the targeted gene and all genomic sequences of the same species. We use the **blastn** search mode (Altschul et al. 1997) of the **blastall** program from NCBI (ftp://ftp.ncbi.nih.gov/blast/executables/) to find the genomic region coding for the gene. The E-value threshold is set to 10^{-20}. We extract the sequence of the genomic region. To locate the precise boundaries of the exons (i.e. the positions of the introns) within the cDNA sequence , we use the **sim4** program (Florea et al. 1998) (http://globin.cse.psu.edu/html/docs/sim4.html). **Sim4** rapidly aligns a spliced transcript sequence to its parent genomic sequence and attempts to find to correct exon–intron junction. The program generates a list of exons and their positions on the sequences. **Fig. 1** shows an example **sim4** result containing the list and an optional alignment. If the gene contains introns, the intron positions within the cDNA sequence can be determined. Although **sim4** may miss small marginal exons, usually at least one intron suitable for targeting can be identified for a gene.



```
1-84   (92566-92649)    100% ->
85-296   (92739-92950)    100% ->
297-440   (93140-93283)    100% ->
441-693   (93558-93810)    100% ->
694-826   (93923-94055)    100% ->
827-1002   (94599-94773)    97%

    0     .    :    .    :    .    :    .    :    .    :
    1 GGGAAGATAAGAAAATGGTAGGCGCTGTTTCAGTAAACCCAGCAAGCACC
      ||||||||||||||||||||||||||||||||||||||||||||||||||
92566 GGGAAGATAAGAAAATGGTAGGCGCTGTTTCAGTAAACCCAGCAAGCACC

   50     .    :    .    :    .    :              .    :
   51 ATTTTATCATCTTTGACATATAGTCGAAAACTTG         GAAACAG
      ||||||||||||||||||||||||||||||||||>>>...>>>|||||||
92616 ATTTTATCATCTTTGACATATAGTCGAAAACTTGGTA...CAGGAAACAG

  100     .    :    .    :    .    :    .    :    .    :
   92 GTTCTTTTTAGGTAACAAACCAAGGACAGTATCAATGTCTCTCAAAAACG
      ||||||||||||||||||||||||||||||||||||||||||||||||||
92746 GTTCTTTTTAGGTAACAAACCAAGGACAGTATCAATGTCTCTCAAAAACG

  150     .    :    .    :    .    :    .    :    .    :
  142 AAAATCGTAATAAAAAGCAGCTAGAAGTGAGTACGGGAGCAGCTGAAAGG
      ||||||||||||||||||||||||||||||||||||||||||||||||||
92796 AAAATCGTAATAAAAAGCAGCTAGAAGTGAGTACGGGAGCAGCTGAAAGG

  200     .    :    .    :    .    :    .    :    .    :
  192 CTGTATCTCGGGTTAGACTTTGGTACTTCTGGTGCAAGGTTTGCTATTAT
      ||||||||||||||||||||||||||||||||||||||||||||||||||
92846 CTGTATCTCGGGTTAGACTTTGGTACTTCTGGTGCAAGGTTTGCTATTAT

  250     .    :    .    :    .    :    .    :    .    :
  242 TGACATTGGTGGTACAATTCAAGCCGAGGCGAAGCGACCCTATCCGTTAT
      ||||||||||||||||||||||||||||||||||||||||||||||||||
92896 TGACATTGGTGGTACAATTCAAGCCGAGGCGAAGCGACCCTATCCGTTAT

  300     .    :    .    :    .    :    .    :    .    :
  292 ACTTG         AAAGAATGGAGAGTCACATGATTGGGTACGCTCATG
      |||||>>>...>>>||||||||||||||||||||||||||||||||||||
92946 ACTTGGTA...TAGAAAGAATGGAGAGTCACATGATTGGGTACGCTCATG
```

Figure 1: List of exons and optional alignment in the output of **sim4**
A. exon positions, **A1.** exon positions on the cDNA, **A2.** exon positions on the genomic sequence, **A3.** sequence similarity, **B.** spliced alignment (partially shown), **B1.** place of intron.

Having located an intron, the joined sequences of the flanking exons can be passed to a primer designer program. A target position corresponding to the position of the intron between the two exons is specified and the designer program is instructed to find a forward primer for the 5' exon and a reverse primer for the 3' exon. The size of the PCR product is predicted by adding up the distance between the primer positions and the length of the intron. If a PCR product with an effectively amplifiable product size can be predicted for two exons that are not adjacent in the cDNA, the above procedure is carried out for such an exon pair with the target being two introns and another exon between them. We prefer a product size within a range of 300 – 2000 bp and use filtering parameters to select the suitable exons accordingly.

To generate the primer sequence pairs we use the **Primer3** program (Rozen and Skaletsky 2000) (http://frodo.wi.mit.edu/primer3/primer3_code.html) since it can be built into a processing pipeline. The program takes as an input a text file containing the sequence and the parameters (e.g. target position, requested product size).

Plant genomes contain many multigene families. Since IT markers are based on primers matching the relatively conserved protein-coding exons of genes, the amplification of other members of a multigene family cannot be excluded. Therefore, it is important to primarily choose single- or low-copy genes as targets. A computational approach aimed to predict copy number of genes can be based on clustering of homologous cDNA sequences. In case of a species for which extensive EST sequencing and generation of tentative EST consensi (TC) has been performed, the number of non-identical sequences in a cDNA/TC cluster provides an approximation on the copy number of the gene. When comparing two homologous sequences, two thresholds of sequence matching are used: an *identity threshold* to distinguish non-identical sequences from identical ones and a *similarity threshold* to determine whether two sequences belong to the same cluster or not. To account for EST sequencing errors, the identity threshold must be set 1–2 % below true identity. Similarity threshold is usually set to 80%. An all-against-all similarity search using **blastn** can provide the pairwise similarity information that can be post-processed using a simple single-linkage clustering algorithm.

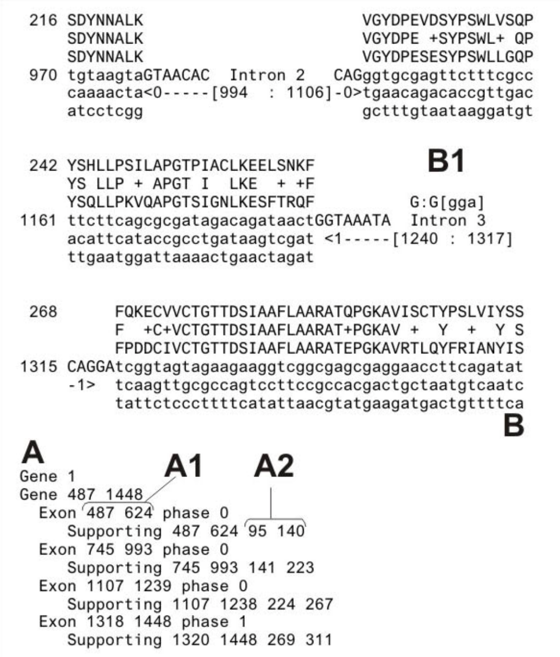
cDNA and genomic sequences are from different species

cDNA from the target species and genomic sequences from a helper species

If the genomic sequence of the targeted gene is not available from the species of interest, the homologous genomic region from a related species may offer the possibility to predict the position of the introns within a cDNA sequence and estimate the intron lengths. To obtain reliable predictions, the following criteria must be met: the proteins encoded by the genes should be sufficiently similar, and the exon/intron structure of the orthologous genes should correspond to each other. To satisfy these criteria, the two species have to be evolutionarily close to each other.

The cDNA sequence of the gene and species of interest is used as query to find the homologous (orthologous) gene in the genomic sequence of the second species. If we want to design primers for *M. truncatula*, we can select another legume or *Arabidopsis thaliana* as the ‘helper’ species. A database of the genomic sequences of the second species can be searched using either **blastn** (E-value threshold: (E-value threshold: 10^{-50}) or **tblastx** (E-value threshold: 10^{-20}). Although the **tblastx** results must be evaluated with caution since artifacts cannot be completely eliminated by filtering for E-value, the higher sensitivity of the search carried out at the protein level may justify its use.

The genomic region coding for the helper gene must be extracted and aligned to the query cDNA. The **sim4** program is designed to compare nearly identical sequences, differing only in the presence or absence of introns, therefore its use for inter-species comparison is very limited. Taking advantage of the higher similarity observed at the protein level, the alignment of genomic DNA to a protein sequence that may even be heterologous (i.e. from a different species) can be carried out using the **genewise** program of the **Wise2** package (Birney et al. 2004) (http://www.ebi.ac.uk/Wise2/). Both **sim4** and **genewise** take into account the canonical exon/intron junction sequence (GT…AG) to predict the correct boundaries of the exons while **genewise** also determines and checks the intron phase at the putative junctions. Before applying **genewise**, the cDNA of the gene has to be translated into the corresponding protein sequence. One can use the **transeq** program of the **EMBOSS** package (Rice et al. 2000) to obtain the protein sequence. If the gene contains introns, **genewise** provides the exon positions with respect to the protein sequence, therefore the numbers need to be converted into positions on the cDNA sequence. An example **genewise** output is shown in **Figure 2**.



```
 216 SDYNNALK                          VGYDPEVDSYPSWLVSQP
     SDYNNALK                          VGYDPE +SYPSWL+ QP
     SDYNNALK                          VGYDPESESYPSWLLGQP
 970 tgtaagtaGTAACAC  Intron 2   CAGggtgcgagttctttcgcc
     caaaacta<0-----[994  : 1106]-0>tgaacagacaccgttgac
     atcctcgg                          gctttgtaataaggatgt


 242 YSHLLPSILAPGTPIACLKEELSNKF
     YS LLP + APGT I  LKE  + +F
     YSQLLPKVQAPGTSIGNLKESFTRQF          G:G[gga]
1161 ttcttcagcgcgatagacagataactGGTAAATA  Intron 3
     acattcataccgcctgataagtcgat <1-----[1240 : 1317]
     ttgaatggattaaaactgaactagat


 268       FQKECVVCTGTTDSIAAFLAARATQPGKAVISCTYPSLVIYSS
           F  +C+VCTGTTDSIAAFLAARAT+PGKAV +  Y  +  Y S
           FPDDCIVCTGTTDSIAAFLAARATEPGKAVRTLQYFRIANYIS
1315 CAGGAtcggtagtagaagaaggtcggcgagcgaggaaccttcagatat
     -1>  tcaagttgcgccagtccttccgccacgactgctaatgtcaatc
          tattctcccttttcatattaacgtatgaagatgactgttttca

Gene 1
Gene 487 1448
  Exon 487 624 phase 0
     Supporting 487 624 95 140
  Exon 745 993 phase 0
     Supporting 745 993 141 223
  Exon 1107 1239 phase 0
     Supporting 1107 1238 224 267
  Exon 1318 1448 phase 1
     Supporting 1320 1448 269 311
```

Figure 2: List of exons and alignment in the output of **genewise**
A. exon positions, **A1.** exon positions on the cDNA, **A2.** exon positions on the genomic sequence, **B.** spliced alignment (partially shown), **B1.** place of intron.

Having determined the positions of the exons flanking one or two introns, the design of the primers based on the cDNA sequence is carried out as described above. However, the length of the targeted intron is not known, consequently the size of the PCR product can only be estimated based on the length of the orthologous intron.

Genomic sequence from the target species and cDNA from a helper species

If a cDNA database exists for the helper species, either **blastn** or **tblastx** can be used as above to find the orthologous/homologous pair of cDNA and genomic sequences. The genomic region to be targeted and its cDNA orthologue must then be aligned as above. The primer pair is then designed for an exon pair in the genomic sequence from the species of interest.

Web applications to help primer design

We developed two web applications to facilitate the last two steps of the primer design process described above. One uses the **sim4** program to align cDNA sequence of a gene to its genomic counterpart while the other runs the **genewise** program to align the protein encoded by the cDNA to a homologous genomic sequence from another species. In both cases the corresponding pair of cDNA and genomic sequences must be identified beforehand. The two sequences can either be pasted into the input form or uploaded from sequence files (in fasta format). The exon positions are first determined by the software then the user can select the two exons that will be used by the **Primer3** program to design the forward and reverse primers, respectively. The online applications are available at http://bioinformatics.abc.hu/itprim/.

4.1.2. PCR amplification protocols

Specific amplification using *Pfu* enzyme:

PCR amplification reaction mix consist of 10-10 pmol of forward and reverse primers, 1 U of *Pfu* enzyme, 2mM $MgSO_4$, 10 mM $(NH_4)_2SO_4$, 20 mM Tris-HCl (pH8.8), 10mM KCl, 0.1% Triton X-100, 0.1 mg/ml BSA, 0.75 mM activated calf thymus DNA, 200 mM of each dNTP, and 25 ng total DNA of the individuals

***Pfu* enzyme incorporates nucleotides at 70-80ºC and is more thermostable than *Taq* polymerase

**For *Pfu* enzyme $MgSO_4$ is needed in PCR instead of $MgCl_2$

****Pfu* is eight times more accurate than *Taq* polymerase

Specific amplification using *Taq* enzyme

PCR amplification reaction mix consists of 10-10 pmol of forward and reveres primers, 1 U *Taq* polymerase enzyme, 1.5 mM $MgCl_2$, 200 mM of each dNTP, and 25 ng total DNA of the individuals in 1x *Taq* polymerase buffer in a final volume of 25 µl.

RAPD amplification

The reaction mix consists of 10 pmol 10-mer random primer, 1 U *Taq* polymerase enzyme, 2.4 mM MgCl2, 200 mM of each dNTP, and 25 ng total DNA of the individuals in 1x *Taq* polymerase buffer in a final volume of 25 µl.

* To improve the success of amplification in some cases $MgCl_2$ or $MgSO_4$ gradient is advised between 1 mM and 2.5 mM concentration

4.1.3. PCR programs

- Gradient PCR programs
  In cases when the optimal working condition of primers is unknown, gradient amplification conditions are advised: the reactions can be carried out in 35 cycles of 30 s at 94°C; 1 min at different annealing temperature decreasing by each column with a predetermined number of C grades (e.g. 60-58-56-54-52-50-48-46), 1 or 2 min at 72°C, the reactions can be terminated with a final extension at 72°C for 4 min.
- Touch down amplification programs can be used in case of RAPD amplification or when annealing temperature of a primer is unknown. In 2 cycles: 30 s at 94°C, 1 minute at 60°C and 1 minute at 72°C, repeat this and decrease the annealing temperature by 2°C after each second cycle. Reach the lowest annealing temperature that can be considered useful and apply at least 30 cycles of amplification, than terminated with a final extension at 72°C for 4 min.
- RAPD amplification is in fact a touch down amplification program reaching 37°C as the lowest annealing temperature.

4.2. The candidate gene direct sequencing and SNP discovery method

This method focuses on the mapping of candidate genes in *M.tr.* and other legume species. Putative orthologous sequences are searched for and sequenced in the parents of mapping populations in order to develop SNP markers.

4.2.1. If candidate gene sequences are available in the other legume species

Sequences for the genes of interest should be retrieved from GenBank and EMBL databases and primers designed to amplify 0.3-3.0 kb-sized fragments, depending on the length and type of sequence available -genomic DNA or cDNA. PCR reactions are carried out in a total volume of 25 µl containing 20 ng of template genomic DNA, 0.2 mM of each primer, 0.2 mM dNTP, 1.5 mM $MgCl_2$, 1X Taq buffer, and 1.5 units Taq polymerase. After an initial 3 min denaturation step at 94°C, 35 cycles each of 50 s denaturation at 92°C, and 50 s at the required Tm (locus-dependent) and 3 min elongation at 72°C, are performed. These cycles are followed by a final 5 min elongation step at 72°C. PCR products are purified from 1-2% agarose gels using the NucleoSpin gel-extraction kit (Macherey-Nagel, Düren, Germany) and sequenced directly. Sequences should be aligned and insertions/deletions and/or SNPs looked for among the parents using ClustalW (http://www.infobiogen.fr/services/analyseq/cgi-bin/clustalw_in.pl).

Putative orthologous genes should be searched in *Medicago truncatula.* Several strategies are possible. In some cases, the same primer pairs used in other legume species can amplify the genomic DNA of *M.tr.* parental lines. PCR conditions used for other legume species should be tried initially, and PCR conditions optimized in order to obtain a single band in the electrophoretic profile. When there is no amplification, orthologous sequence should be searched in the *M.tr.* EST databases (http://medicago.toulouse.inra.fr/Mt/EST/ or http://www.tigr.org/tigr-scripts/tgi/T_index.cgi? species=medicago), and specific primers should be designed for *M.tr*. The amplification products are sequenced directly and screened for polymorphism between *Mtr* parental lines. In the remaining cases putative orthologous genes can be searched in *M.tr*. BAC sequences and their linkage group assignment and position when available, on http://www.medicago.org/genome/.

4.2.2. Starting from *M.tr.* mapped gene markers to design putative orthologous gene markers in other legume species

EST-derived microsatellite markers have been designed and mapped in M.tr. (Jemalong x DZA315.16 genetic map (T. Huguet, http://medicago.toulouse.inra.fr/Mt/GeneticMAP/LR4_MAP.html). Physical map and sequences can also be searched on http://www.medicago.org/genome/. Then, EMBL database should be searched for homologous sequences in other legumes. Where good homology is found, primer pairs can be designed in order to amplify, sequence, and detect polymorphism between the other legume species parental genotypes.

5. **Polymorphism detection**:

When no sequence information is available for the parents of the mapping populations (the IT method), different polymorphism detection techniques will be used successively (7.1., then 7.2. and/or 7.3). When sequence information is available (candidate gene direct sequencing and SNP discovery method), the strategy used for polymorphism detection can be adapted to the type of polymorphism revealed by sequence analysis.

5.1. **Length and single dose polymorphism** detection by agarose gel electrophoresis

After PCR amplification loading buffer is added to the samples (5 µl to the 25µl volume) and are run on different concentration of agarose gel. Loading buffer contains 50 mM Tris-HCl pH 8.0, 40% sucrose, 10 mM EDTA pH8.0, 0.05% bromophenol-blue.

The PCR fragments can be separated in different concentration of agarose gel depending on their length (**Table 1.**). Ethidium bromide is added to the agarose gel (50mg EB/100ml agarose gel) or in the migration buffer in order to visualize the double stranded linear DNA.

Safety: Ethidium bromide is mutagenic, wear gloves when handling stock and any solution or gel that contains ethidium bromide

Table 1. The range of separation of the linear double stranded DNA on agarose and polyacrylamide gel.

Agarose gel (%)	**Range of separation (bp)**	**Polyacrylamide gel (%)**	**Range of separation (bp)**
0.5	1000-3000	3.5	100-1000
0.7	800-12000	5.0	80-500
1.0	500-10000	8.0	60-400
1.2	400-7000	12.0	40-200
1.5	200-4000	20.0	5-100
2.0	50-2000		

In order to perform electrophoresis different migration buffers can be used. In the case of agarose gel the most commonly used migration buffer is the TEA (Tris/EDTA/acetate) and it contains: 40mM Tris base, 20mM glacial acetic acid, 1 mM EDTA (pH 8.0) and distillated water (to 1liter). The other commonly used migration buffer is TBE (Tris/borate/EDTA) containing: 90 mM Tris base, 90 mM boric acid, 1 mM EDTA (pH 8.0) to 1 liter of distillated water.

The PCR samples should be loaded on the appropriate agarose gel and run at maximum 120 V to separate the amplification fragments. The migration bands can be visualized and photographed under a UV lamp. In case of amplification length polymorphism (ALP), bands of different sizes can be observed (**Fig. 1**). In case of single dose polymorphism the amplification product will be missing in one of the amplified individuals (**Fig. 2**).

Figure 3: Agarose gel electrophoresis of cc. 600/650 bp ALP fragments.

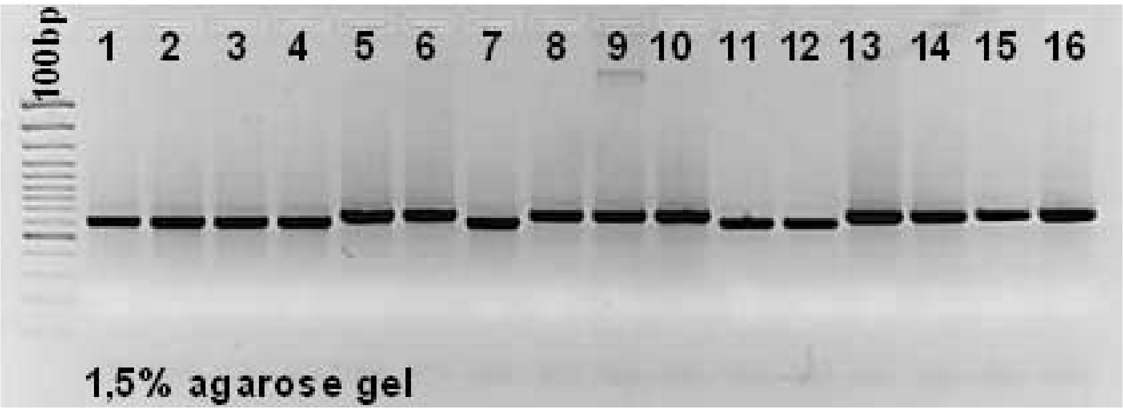
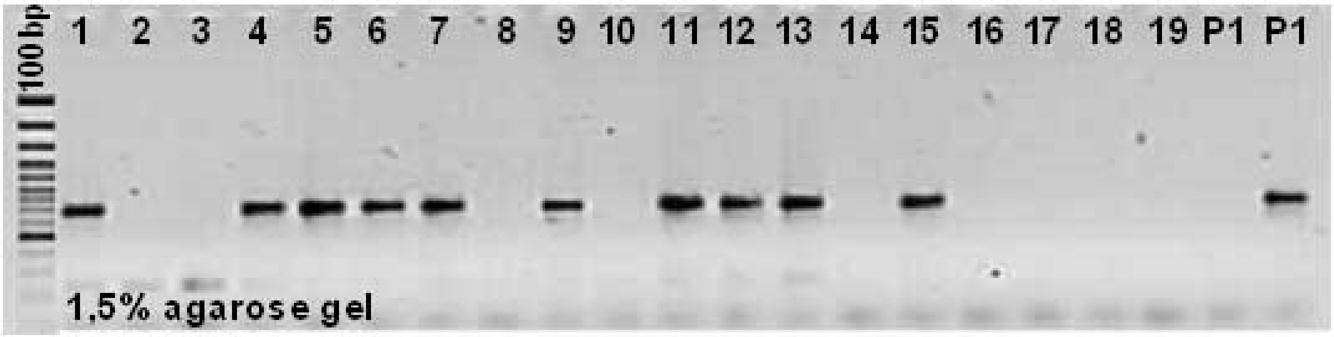


Figure 4: Agarose gel electrophoresis of a cc. 690 bp single dose polymorphic fragment.



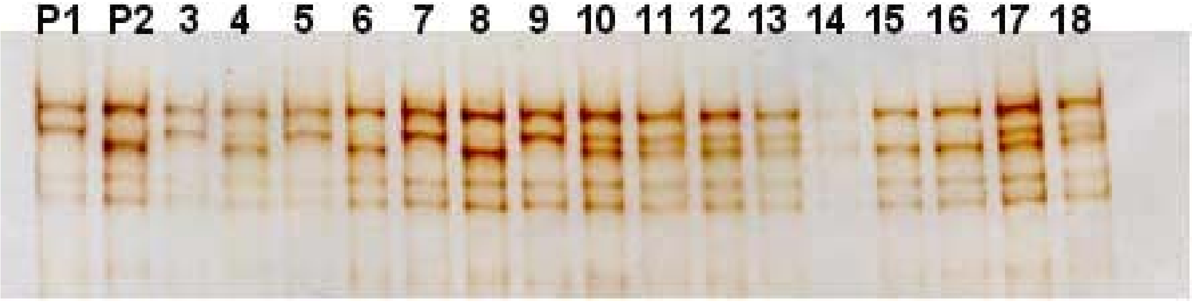
5.2. **Single Stranded Conformation Polymorphism (SSCP)** detection by PAA gel electrophoresis. SSCP analysis detects point mutations and other electrophoretic mobility differences that can result from small changes in nucleotide sequence. Even a single base change can cause conformational change in the DNA molecule. Under non-denaturing conditions and reduced temperature, single stranded DNA molecules have unique conformation which depends on their nucleotide sequence. This different conformation will cause detectable difference in their mobility.

The PCR amplification product containing the potential polymorphism is denatured by boiling in a formamide-containing buffer (95 % formamide, 20 mM EDTA, 0.05 % bromophenol blue, 0.05 % xylene cyanol) followed by a rapid chilling to prevent the re-annealing of the complementary strands. The samples can be separated under non-denaturing conditions like different concentrations of polyacrylamide gel (PAA) (**Table 1**.). After electrophoresis the PAA gel can be stained with silver or other dyes like Sybr Green:

- The PAA gel should be fixed in 1 % HNO_3 for 3-4 min
- than washed 3 times in distillated water for 3×20 sec
- and stained with 0.2 % $AgNO_3$ for 30 min.
- After staining, the gel should be washed 3 times for 20 sec with distillated water
- than developed in a solution containing 0.28 M $NaCO_3$ (5.94 g $NaCO_3$ + 27 µl 37 % HCOH to 200 ml distillated water). The gel should be kept in the developing solution until the bands are clearly visible.
- The developed gel has to be fixed in a 10 % CH_3COOH for 5 min
- washed two times for 20 sec in distillated water
- and impregnate in 10 % glycerin solution for 5-10 min.

The gel can be stored for a long time in a dried form between two layers of cellophane.

Figure 5: SSCP detected on silver stained PAA gel; P1, P2: the two different parental plants, 3–18: individuals from an F2 mapping population.



5.3. **Heteroduplex digestion with CelI enzyme** and polymorphism detection by PAA gel electrophoresis. Plants and fungi contain single-stranded specific nucleases that attack both DNA and RNA, for example: S1 nuclease form *Aspergillus oryzae*, P1 nuclease form *Penicillium citrinum*, and mung bean nuclease from *Vigna radiata*. CelI is similar to these enzymes; it shows single- and double stranded DNase and endonuclease activity and is very active on mismatch substrate. In comparison to the other enzymes, CelI has a neutral pH optimum, prefers double-stranded mismatched DNA substrates, is not inhibited by high GC content, and is stimulated by magnesium ions. CelI characteristically cuts one strand of the double stranded DNA at the 3'end of the mismatches (Yeung et al. 2005). This enzyme can be used to detect SNPs, small insertions/deletions and few nucleotide mismatches, being a new and efficient tool in polymorphism detection (Fig.5).

a. First, PCR amplification has to be performed using the adequate conditions previously established for the respective primer pair (amplification using the genomic DNA of the parental plants and/or the individuals of the mapping population)
b. The presence of the amplification products has to checked by agarose gel-electrophoresis. If amplification was successful
c. heteroduplex has to be generated, in order to detect de missparing between the to amplified DNA strands.
- First, a mixture of (e.g.) 10-10 µl of PCR amplification products, obtained with the genomic DNA of the parental plants, can be used as control probes. If polymorphism can be detected at the control probes lave, the genotyping can further be done for the whole mapping population. In the case of RI populations the obtained PCR products should be mixed with the PCR product of one of the parental plants in order to distinguish one homozygote from the other. Hereby it can be detected if the added parental individual has the same genotype with analyzed individual. In the case of F2 mapping populations heteroduplex cuttings will be obtained in the heterozygote plants and the two homozygous ones can be distinguished after a second digestion. In order to distinguish the two homozygous patterns the amplification products of one parental plant is added to form new heteroduplexes. The new cuttings compared to the previous ones, will mark the homozygous genotype opposite to the added parental plant.
- Heteroduplex can be generated in a thermocycler using the following temperatures:
A. 94ºC for 2 min
B. chilling down with 0.1 ºC/s speed until 20ºC
C. End
d. Digestion of the formed heteroduplex:
A 100-fold dilution of the undiluted CelI enzyme, using the Dilution Buffer, has to be performed for the digestion.

<u>Dilution Buffer</u>
50 mM Tris-HCl pH 8.0
0.5 mM alpha-methyl-mannoside
0.01% Triton X-100
100 µM phenylmethylsulfonyl fluoride (PMSF)
0.5 M KCl

<u>Digestion</u>
template (heteroduplex)
Millipore Water (MQ) to complete to a final volume
10× CelI Buffer
CelI enzyme (100× diluted)

<u>10× CelI Buffer</u>
100 mM $MgSO_4$
100 mM HEPES pH 7.5
100 mM KCl
0.02% Triton X-100
0.002 mg/ml BSA

<u>Incubation</u>: 37ºC for 30-45 min
e. After digestion the samples have to precipitate and redilute to a concentration which can be loaded in the PAA gel

DNA precipitation

Complete the samples with MQ water to 100 µl

+300 µl NH_4Ac-ethanol (Mixture of: 0.5× volume NH_4Ac 7.5M + 2.5× ethanol 96%) to the samples then mix well with a vortex. keep at -20ºC for 1h or overnight

f. Centrifuge the samples for 5 min at 15 krpm and carefully decant the supernatant

g. Centrifuge again for 2 min at 15 krpm and carefully decant the remained supernatant

h. Dry the samples in a Speed Vaq. for 30 min

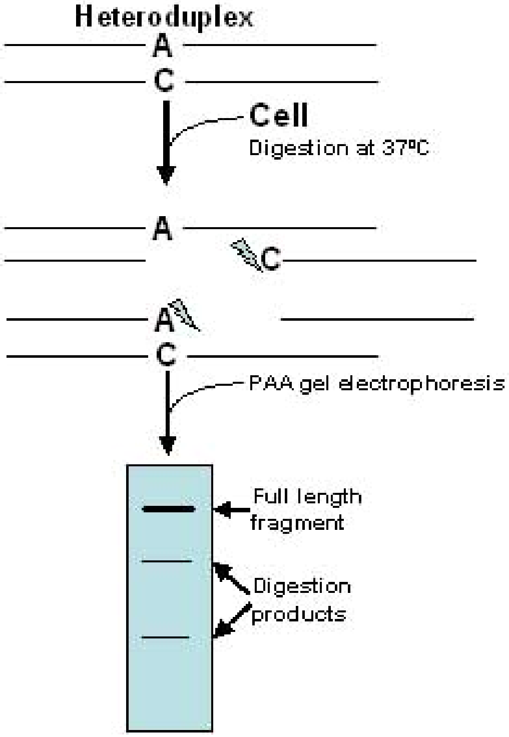
i. Resolve the samples in SSCP loading buffer (see above)

keep the samples 5-10 min at RT than vortex it and spin down

j. PAA gel migration (**Fig.4**)

use 0.5× TEB migration buffer

Figure 6 Digestion of nucleotide mismatches With CelI enzyme



10 % PAA gel + 6 M Urea

14.4g Urea

10 ml 40 % acryl-amide

3.8 ml 2 % bis-acrylamide

1.5 ml 10× TEB buffer

fill up to a final volume of 35 ml

60 µl N,N,N',N'-tetramethylethylene-diamine (TEMED)

150 µl 10 % APS

(this in the case of a gel of 17×17.5 cm large and 1mm thick)

The TEMED and the APS components should be added to form the PAA gel only after the complete dissolving of the urea.

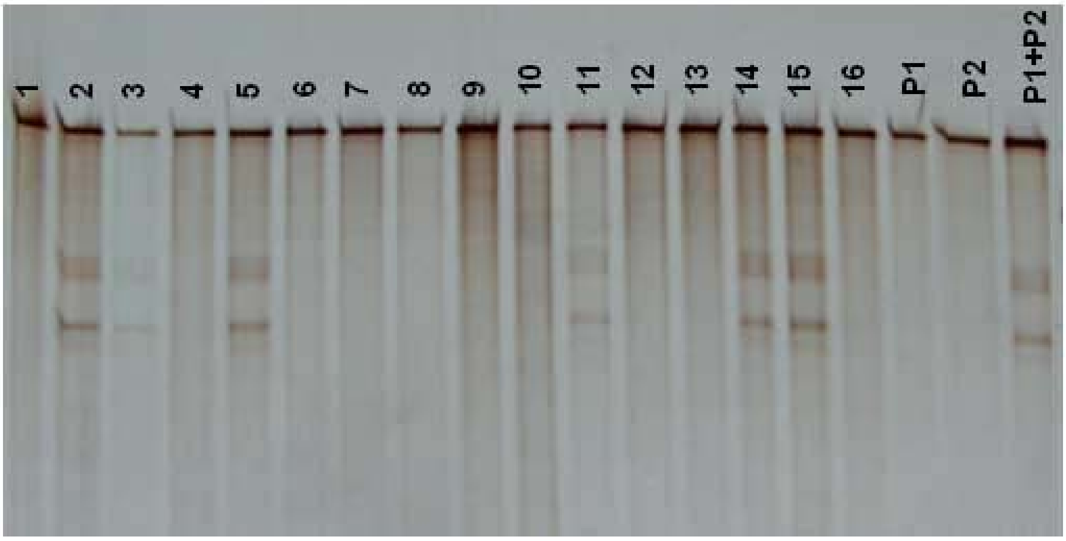
- the PAA gel should be preheated to 45ºC (1000V and 50mA about 20') the migration duration depends of the size of the digested DNA fragments.
- the samples should be first heat-denaturized and chilled down in ice, than quickly loaded on the gel.

- migrate at 50ºC (1000V)

k. PAA gel staining

- First, in order to stain the gel, the urea should be washed away:
- by fixing the gel in 200 ml solution of 50 % methanol + 12% acetic acid + 0,02% formaldehyde for 1 hour
- Washing 3× in 200 ml 50 % ethanol for 3 × 10'
- Treat the gel with 200 ml 1 % nitric acid for 5 min
- Wash the gel 3 times in MQ water for 20 sec
- Silver-stain the gel with 0.2 % silver-nitrate for 30 min
- Wash the gel 3 times in MQ water for 20 sec
- Develop the staining by flushing the gel in 0.28 M $NaCO_3$ solution
- Fix the stained gel in 10% acetic acid
- Steep the gel in 10% glycerin for 15-60 min
- Store the gel as usually

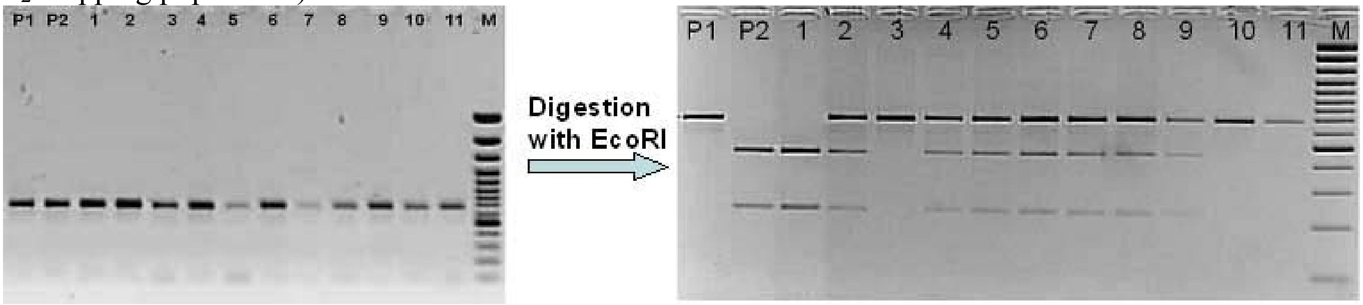
Figure 7: CelI digestion of the amplification products of two parental plants (P1, P2), the mix of P1 and P2 and 16 individuals of a RIL mapping population mixed with P2.



5.4. **Cleaved Amplified Polymorphic Sequence (CAPS) detection**

In the CAPS method, gene-specific primers are used to amplify template DNA and polymorphic nucleotides are detected by the loss or gain of restriction enzyme recognition site. The amplification products can be digested with the respective restriction enzyme and the obtained polymorphism can be scored. (**Fig. 8**) The CAPS method is an easy and reliable method to detect SNPs if sequence information is available.

Figure 8: Polymorphism detected with CAPS method: digestion of the amplification products with EcoRI enzyme (P1 and P2 are the parental individuals and 1-11 are individuals form an F_2 mapping population)

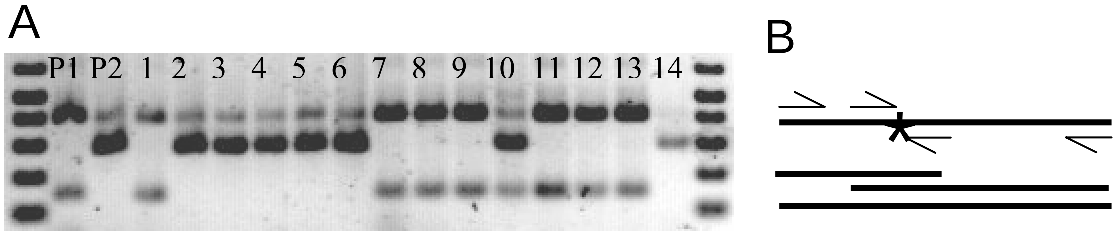

P1 P2 1 2 3 4 5 6 7 8 9 10 11 M
Digestion with EcoRI
P1 P2 1 2 3 4 5 6 7 8 9 10 11 M

CAPS is a powerful and widely used method to detect single nucleotide polymorphism, however it has the limitation of using restriction enzyme sites. Therefore a modification of this technique was developed, where mismatches are in the primer recognition site and they target to detect the polymorphism (dCAPS) (Neff et al. 1998). A restriction enzyme recognition site which includes the SNP, is introduced in the PCR product by a primer, containing one or more mismatches to the template DNA. The PCR product modified in this manner is then subjected to restriction enzyme digestion and the presence and absence of the SNP is determined by the resulting restriction pattern.

5.5. **Allele-specific primers**

In the case of the candidate gene direct sequencing method, SNPs are identified in the parental sequences. Thus, bi-directional allele-specific PCR can be performed as described by Délye et al. (2002): two internal allele-specific primers (ASP) are designed (the 3' end of the primer corresponding to a polymorphic region) and added to the 2 external primers the PCR reaction, producing 3 amplimers (2 specific for each genotype and 1 common to both parents). PCR products are electrophoresed on 1-2.5% agarose gel and visualized after Ethidium Bromide staining (Fig. 9).

Figure 9: Polymorphism detected with ASP method in gene encoding the granule bound starch synthase 2 (*P. sativum,* Aubert *et al.* 2006): (A) amplification carried out with internal allele specific primers and external primers gives 2 two-bands profiles. (B) The upper band is common to all genotypes, the two other bands correspond to the two alleles.



6**. Genotyping** the individuals of the mapping population. After obtaining the polymorphism by different methods, the next step is to establish the genotype of the individuals of the mapping population in concordance with the genotype of the parental plants. The inheritance of the loci can be scored as follows: 1, homozygous for the female allele (A1A1); 3, homozygous for the male allele (A2A2); 4, not-homozygous for the female allele (A1A2 or A2A2); 5, non-homozygous for the male allele (A1A2 or A2A2); 2, heterozygous carrying both alleles (A1A2); and 0 is the missing data.

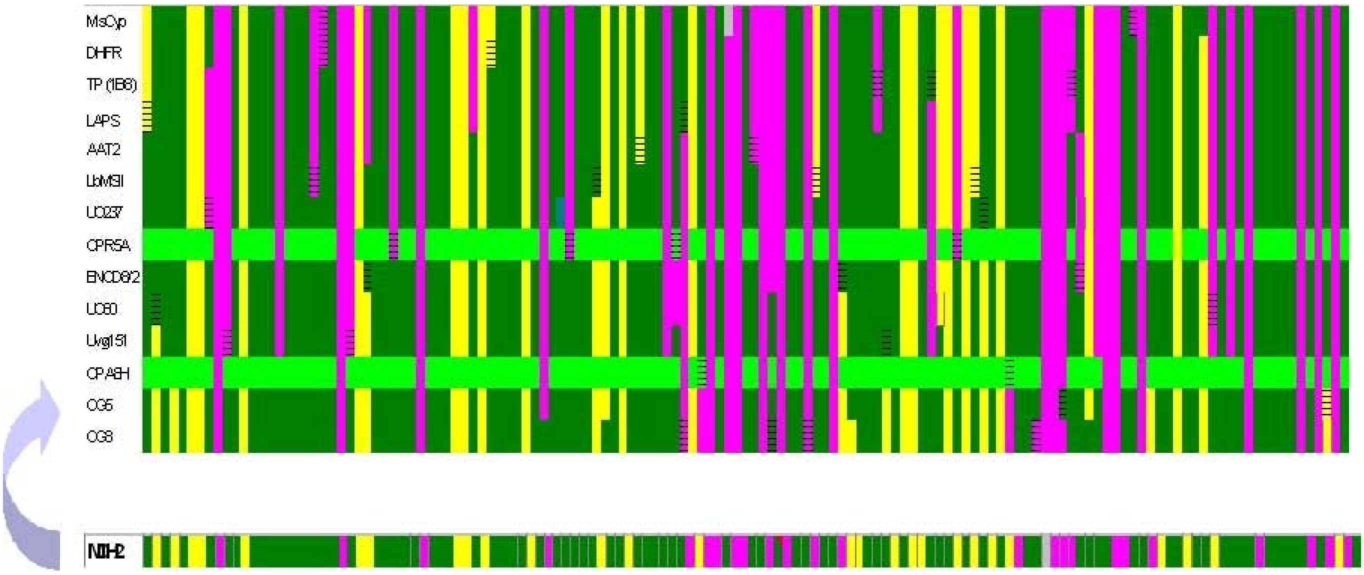
7. **Scoring of genotypes** and positioning of the markers on the genetic map. After determining the genotype of each plant of the mapping population, the position of the newly generated markers has to be found. The alleles of the F_1 plant are assumed to segregate according to Mendelian rules (Mendel 1866) in the F_2 population. Theoretically, the alleles A1 and A2 are segregating according to 1:2:1 ratio. This segregation can be detected only with codominant alleles. The segregation of dominant/recessive alleles will follow the 3:1 ratio. In practice the segregation of the F_2 mapping population is usually distorted to some extent. The degree of distortion depends on the size of the population as well as the number and the quality of the genetically altered loci in the F_1 (Kiss et al. 1993). In the RIL populations usually two genotypes can be observed. The more generations were created after F_1, the smaller is the occurrence of the heterozygous loci.

In plants the genes are linearly arranged along chromosomes. The genes which are on different chromosomes are inherited independently, they are not linked. On the other hand genes which are close to each other on the same chromosome are co-inherited with a given recombination frequency. The loci can be arranged according to their increased recombination frequency. These recombination frequencies can be calculated by different mapping programs and the loci can be arranged manually or by the respective program (e.g. MAPMAKER).

8. **Genetic mapping** using the color mapping method

Color mapping is a non-mathematical method of genetic mapping. It is based on converting the numerical scores, representing the genotypes, into color codes and these are used to display the genotypes of the markers for each individual of the mapping population. Color genotypes are arranged in a matrix where each row correspond to a marker, ordered according to its position in the respective Linkage Group, and each column represents one plant of the mapping population. The location of a new marker is found by recognizing similarities between the color pattern of the individuals for the new marker and the ordered markers in the color map. The new markers should present minimum recombination regarding to the two neighboring markers. (**Fig 10**) Color mapping can be used to find linkages which can not be determined unambiguously by conventional mapping programs. (Kiss et al. 1998).

Figure 10 Positioning of a new marker, by color mapping method, between the already ordered markers in the genetic map. Each row represents one marker and each column one plant of the mapping population and at the intersection of the row/column the corresponding genotype. The yellow and purple colors represent the maternal and paternal homozygotes, respectively, and the green color indicates the heterozygote genotype. In the case when the markers are dominant, only two genotypes can be scored: one homozygote marked as described above and one homozygote/heterozygote marked with light or dark green.

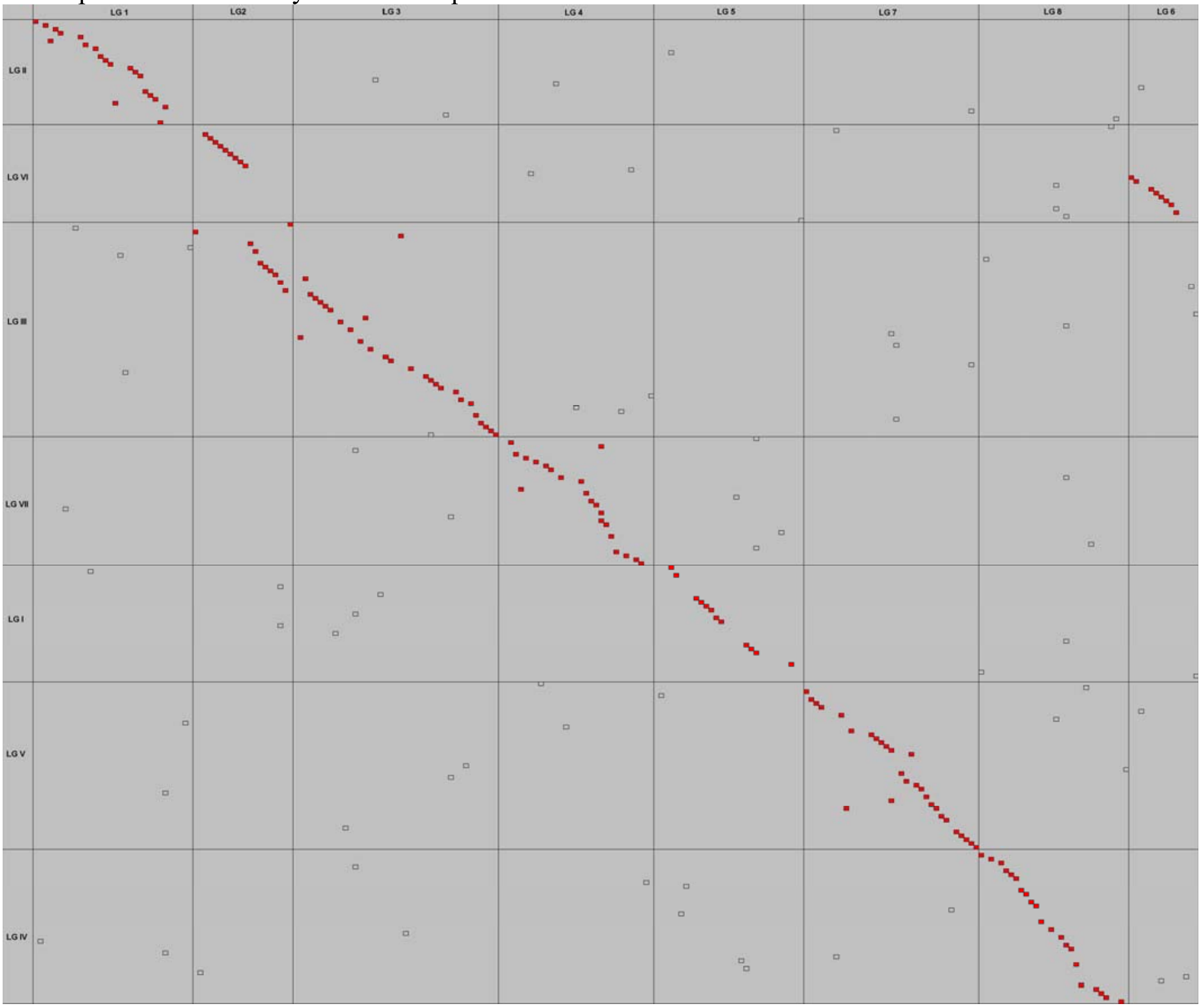


9. **Comparative mapping**: comparing the gene order of homologous genes in *Medicago truncatula* and other plants.

Homologies can be investigated either at the macro- or microsyntenic level depending on the available data. Macrosyntenic studies focus on the genomes as a whole analyzing large regions (e.g., linkage groups) by comparing the order of the genes based on the constructed genetic maps. Microsyntenic comparisons use shorter but continuous stretches of completely sequenced genomic regions in which the order and the orientation of coding sequences as well as the non-coding DNA sections can be investigated.

a) Macrosyntenic comparison. The genetic map of several legumes was compared with the genetic map of the model legume *Medicago truncatula* and an extensive homology in the gene order could be observed. As example we describe the comparison of the composite genetic map of *Medicago truncatula*/*Medicago sativa* and the genetic map of *Pisum sativum* (Kaló et al. 2004). In order to compare the two genetic maps, single copy, gene based genetic markers had to be generated and mapped in both species. Gene-based PCR primers were tested in both species and the obtained markers were mapped on the two genetic maps. Comparison of genetic maps can be performed by analysis of the relative map positions of homologous gene loci plotted on the horizontal and vertical axes, respectively, of a matrix (**Fig. 11**). If the positions of the loci are co-linear in the two genomes, the dots will form an isoclinic or retrograde diagonal axis in the matrix. On the other hand, the positions of those sequences which originated from genetic rearrangements (duplications, translocations, etc.) of smaller genomic regions after the divergence of the two species will be scattered in the matrix. Duplication of larger genetic regions, however, will give rise to additional diagonals.

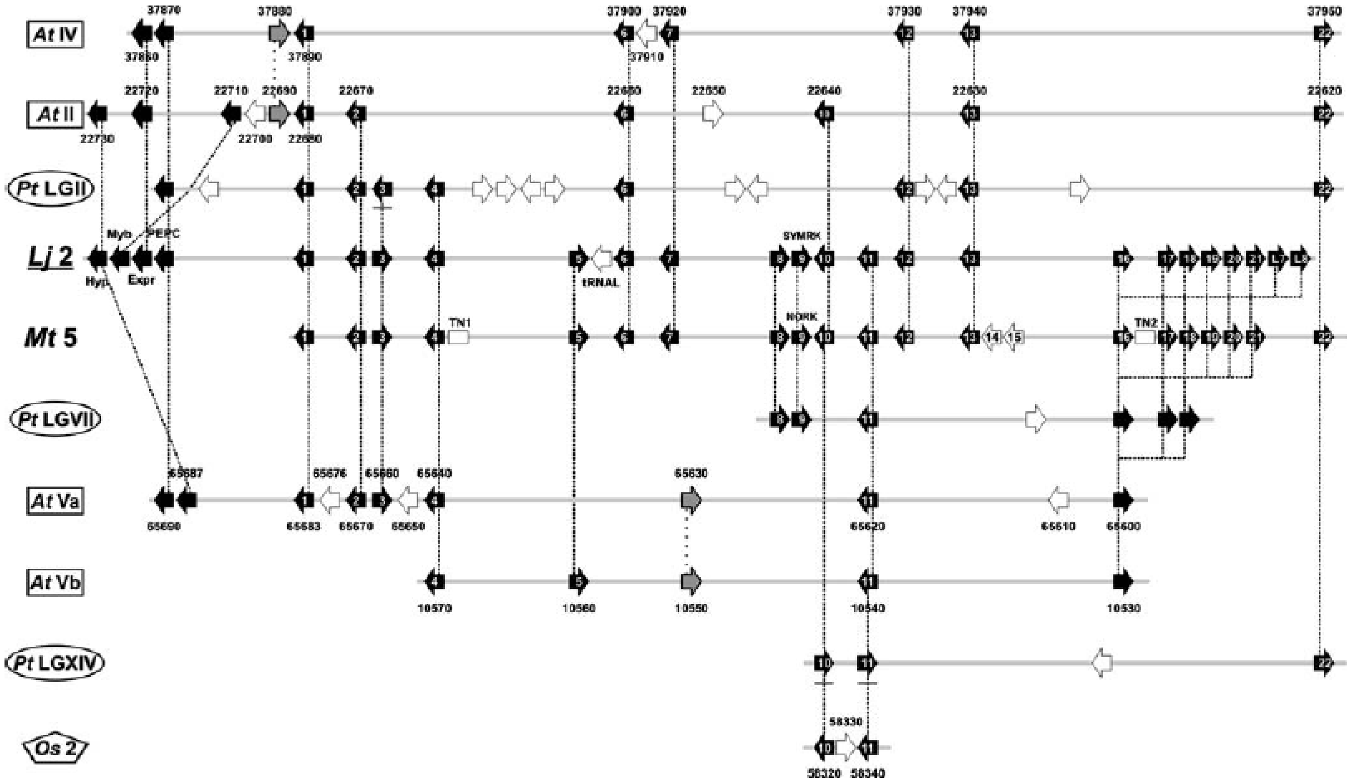
Figure 11. Matrix plot of the common gene-specific loci mapped in alfalfa and pea. The 233 *Medicago* and 252 pea loci are listed horizontally and vertically, respectively, according to their order on the linkage groups. The dots are positioned at each intersection where homologous loci meet (e.g. in the case of both single- and multigene families). The clear diagonals are composed of markers representing syntenic homologous genes. Nonsyntenic homologous positions for which no syntenic homologous counterparts were identified in the other species are shown by uncolored squares in the matrix.



b) Several microsyntenic comparisons were performed within legumes and non legumes, which revealed that the gene content of shorter sequenced regions is highly syntenic or shows limited homology. One example of microsyntenic comparison is the analysis of the NORK region (NOD/SYMRK) (Endre et al. 2002) which has been sequenced (~300kb) and compared to homologous regions which have been found syntenic in other plants like *Lotus japonicus*, *Arabidopsis thaliana* and *Populus trichocarpa* (Kevei et al. 2005). In order to perform microsynteny analyses of this region, the annotation of the NOD-contig was first carried out by homology searches (BlastN, BlastX) in different databases, which resulted in the identification of 22 genes and two retrotransposons (Tn1, Tn2). During the isolation of the SYMRK gene of *Lotus japonicus*, which is an orthologue of the NORK gene, a 345 kb long genomic sequence was determined in the corresponding chromosomal region of this legume species (Stracke et al. 2002).

The two investigated regions exhibit almost complete colinearity in gene content and order. Although this region in *Lotus* is 36 kb shorter, besides the 20 syntenic genes (the two isoleucine-tRNA genes are not present) it contains two additional lectin genes indicating more duplication events in Lotus. Additionally, a small housekeeping gene (aminoacyl-tRNA-ligase) was found in the Lotus region (between gene 5 and gene 6), which is absent in the *Medicago* NOD-contig. In order to reveal microsynteny between the genomes of legumes and other dicotyledonous plants, the NORK/SYMRK region was used to seek homologous genomic regions—defined by homologous genes located close to each other with the same orientation—in the genomes of *Arabidopsis thaliana* and *Populus trichocarpa*. As the first step in the analysis, homologous sequences were identified using the coding sequences in the NORK/SYMRK region as query sequences. Searches were done in the TAIR and NCBI databases by BLAST programs and the resulting best hits were ordered according to the degree of similarity detected. As the second step, the chromosomal positions of the identified *Arabidopsis* and *Populus* genes were analyzed for their possible linkage similar to that found in the NOD-contig In fact, all protein-coding genes in the NORK/SYMRK region of *Medicago* (20 genes) and *Lotus* (24 genes) could be correlated to four *Arabidopsis* and three *Populus* genomic blocks located in chromosomes II, IV, Va and Vb in *Arabidopsis* and linkage groups (LG) II, VII and XIV in *Populus* (**Fig. 12**).

Figure 12 (following page). Microsyntenic relations of the *Medicago* NOD-contig to *Arabidopsis*, *Populus* and *Lotus*. The four syntenic regions of *Arabidopsis* (At IVa, At II, At Va and At Vb, marked with boxes), the NOD-contig of *Medicago* (Mt 5), the SYMRK region of *Lotus* (Lj 2), the three syntenic regions of *Populus* (Pt LGII, Pt LGVII and Pt LGXIV, marked with circle) and the syntenic region of rice (Os 2, marked with pentagon) are present on the figure. The arrows show the genes with their orientation, black arrows mark the syntenic genes (the genes showed an opposite orientation to that of *Medicago* NOD-contig are underlined), and grey arrows indicate syntenic genes exhibiting no relation with NORK/SYMRK region. Thin lines show the syntenic relations. The additional genes of *Lotus* and the LTR-retrotranposons of *Medicago* (TN1 and TN2) are indicated. The *Arabidopsis* and rice genes are marked by their gene-number of TAIR (http://www.arabidopsis.org) and TIGR database (http:// www.tigr.org)



Considering the significant similarity in gene sizes, orders and orientations between *Medicago* and *Lotus*, the other legumes probably have undergone similar genome development resulting in high micro colinearity of the genes.

References

Aubert G, Morin J, Jacquin F., Loridon K, Quillet MC, Petit A, Rameau C, Lejeune-Hénaut I, Huguet T, Burstin J. (2006) Functional mapping in pea, as an aid to the candidate gene approach and for investigating the synteny with the model species *Medicago truncatula*. *Theor. Appl. Genet*, **112**: 1024-1041.

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.

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* **14**: 988-95.

Délye C, Calmes E, Matejicek A (2002) SNP markers for black-grass (*Alopecurus myosuroides* Huds.) genotypes resistant to acetyl CoA-carboxylase inhibiting herbicides. *Theor Appl Genet*. **104**: 1114-1120.

Florea L, Hartzell G, Zhang Z, Rubin GM, Miller W (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res* **8**: 967-74.

Endre G, Kereszt A, Kevei Z, Mihacea S, Kaló P, Kiss GB (2002) A receptor kinase gene regulating symbiotic nodule development. *Nature* **417**: 962–966.

P. Kaló, A. Seres, S. A. Taylor, J. Jakab, Z. Kevei, A. Kereszt, G. Endre, T. H. N. Ellis, G. B. Kiss (2004) Comparative mapping between *Medicago sativa* and *Pisum sativum Mol Gen Genomics* **272**: 235–246.

Zoltán Kevei, Andrea Seres, Attila Kereszt, Péter Kaló, Péter Kiss, Gábor Tóth. Gabriella Endre, György B. Kiss (2005) Significant microsynteny with new evolutionary highlights is detected between Arabidopsis and legume model plants despite the lack of macrosynteny. *Mol Genet Genomics* **274**: 644-57.

György .B. Kiss, Gyula Csanádi, Katalin Kálmán, Péter Kaló, László Ökrész (1993) Construction of a basic genetic map for alfalfa using RFLP, RAPD, izozyme and morphological markers. *Mol. Gen. Genetics* **238**: 129-137.

G.B. Kiss, A. Kertész, P. Kiss, Gabriella Endre (1998) Colormapping, a non-mathematicl procedure for genetic mapping. *Acta Biologica Hungarica* **49**: 125-142.

Michael M. Neff, Joseph D. Neff, Joane Chory, Alan E. Pepper (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphism: experimental applications in Arabidopsis thaliana genetics. *Plant J.* **14**: 387-392.

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**:276-7.

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Misener S, Krawetz SA (eds) Bioinformatics Methods and Protocols. Humana Press, Totowa, NJ, p 365-86.

Stracke S, Kistner C, Yoshida S, Mulder L, Sato S, Kaneko T, Tabata S, Sandal N, Stougaard J, Szczyglowski K, Parniske M (2002) A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* **417**: 959–962.

Anthony T. Yeung, Deepali Hattangadi, Lauryn Blakesley, and Emmanuelle Nicolas (2005) Enzymatic mutation detection technologies. *BioTechniques* **38**: 749-758.