Analysis of gene expression in the outer cell layers of Arabidopsis roots during lateral root development

Veth-Tello, L.M.

Chapter 4

Localization of *AIR1A, AIR1B* and *AIR3* genes in the *Arabidopsis* genome and analysis of their 5'-flanking sequences

Luz M. Veth-Tello, Paul J.J. Hooykaas and Bert J. van der Zaal

Abstract

AIR1A, *AIR1B* and *AIR3* are members of a group of auxin-responsive genes from *Arabidopsis* expressed specifically at sites of lateral root emergence. *AIR1A* and *AIR1B* are two almost identical genes encoding putative membrane proteins belonging to the CLCT family. *AIR3* encodes a protein that possesses all the characteristics of a serine protease belonging to the family of subtilisins. The promoters of the *AIR1A* and *AIR1B* genes possess five highly conserved regions that may be involved in the regulation of the expression of these genes. We studied the importance of each of these regions for the expression of the *AIR1B* gene by promoter-deletion::*GUS* analysis. The 5'-flanking sequences of the *AIR3* gene were also analyzed. The shortest *AIR3* 5'-flanking sequence conferring the characteristic *AIR3* expression pattern was identified. Although the *AIR1* and *AIR3* gene expression patterns are almost identical, no similarities were found in their 5'-flanking sequences. The position of *AIR1A, AIR1B* and *AIR3* in the *Arabidopsis* genome was determined. *AIR1A* and *AIR1B* are located in the long arm of chromosome 4 and are part of a cluster of genes encoding putative membrane proteins belonging to the CLCT family. *AIR3* is located in the short arm of chromosome 2, at a position where little gene activity is found.

Introduction

AIR1A, AIR1B and *AIR3* are members of a group of genes isolated in our laboratory by differential screening of a cDNA library from mRNA isolated from auxin-

treated *Arabidopsis* roots. *AIR1A* and *AIR1B* are two almost identical genes with their homology not only restricted to their ORFs but also including their 5' and 3' non-coding sequences (Neuteboom, 2000). The predicted AIR1A and AIR1B proteins are members of a large family of putative membrane proteins identified in many plant species and called CLCT proteins, because the sequence of the four amino acids CLCT is strictly conserved in the C-terminal part of these proteins (Neuteboom et al., 1999a). The N-terminal proline- or glycine-rich stretch present in most members of the CLCT family is possibly involved in cross-linking the plasma membrane with components of the cell wall (Deutch and Winicov, 1995). At difference with most of the other members of the CLCT family, AIR1A and AIR1B do not possess a proline- or glycine-rich region between the signal peptide and the C-terminal part. This characteristic suggests that the proteins encoded by *AIR1A* and *AIR1B* may not be able to link the plasma membrane to the cell wall as proposed for their related counterparts. *AIR3* is a single copy gene encoding a protein that has many characteristics of a serine protease belonging to the family of subtilisins. Like other eukaryotic subtilisin-like serine proteases, AIR3 possesses a signal peptide, a pro-region splicing site and four well-conserved domains that form the active site of the mature protein (Neuteboom et al., 1999b). Since such proteases are usually secreted, we assume that this protease plays a role in degradation of components of the cell wall.

Previous promoter::*GUS* analyses have shown that the *AIR1A, AIR1B* and *AIR3* expression patterns are almost identical and highly specific for cells surrounding lateral roots and for those cells overlaying sites where laterals will protrude (Neuteboom et al., 1999b; Neuteboom, 2000). According to the type of proteins encoded by these genes and their expression pattern it has been assumed that they play a role in weakening plasma membrane-cell wall connections to facilitate lateral root emergence.

Comparison of 2.6 kb and 1.1 kb of *AIR1A* and *AIR1B* 5'-flanking sequence showed three highly homologous regions of 350 bp, 59 bp and 440 bp (Neuteboom, 2000). In *AIR1B* these regions are located immediately adjacent to each other while in *AIR1A* they are separated by sequences probably inserted in the promoter during the course of the evolution. The almost identical expression pattern of *AIR1A* and *AIR1B* genes (Neuteboom, 2000) could be attributed to the presence of these three regions and the highly conserved sequences may indicate that regulatory elements are present there. In this chapter we made a first step towards the identification of regulatory

sequences present in these conserved regions by promoter-deletion::*GUS* expression analysis. We chose for this study the *AIR1B* promoter because it represents a "compact" version of the *AIR1A* promoter as far as these three regions are concerned. In addition, we extended the *AIR1A* and *AIR1B* promoter analysis to unknown 5'-flanking sequences by using information from the *Arabidopsis* genome sequence database (The *Arabidopsis* Genome Initiative, 2000) and we searched for the presence of known auxin and MeJA regulatory elements in the *AIR1A* and *AIR1B* promoter.

We also analyzed the 5'-flanking sequences of the *AIR3* gene. The expression of *AIR3* was studied with plants containing a 7 kb segment of the *AIR3* 5'-flanking sequence fused to a *GUS* reporter gene. By deletion-expression analysis we identified the shortest 5'-flanking sequence required for *AIR3* gene expression and we searched for known regulatory elements in this region.

In the second part of this chapter we analyzed the information from the *Arabidopsis* genome sequence database (The *Arabidopsis* Genome Initiative, 2000) to find the position of the *AIR1A, AIR1B* and *AIR3* genes in the *Arabidopsis* genome. The genes adjacent to *AIR1A*, *AIR1B* and *AIR3* were identified.

Results

The AIR1B promoter

Sixteen constructs containing total or partial deletions of the three highly conserved regions I, II, and III of the *AIR1B* promoter (Figure 1; Neuteboom, 2000) fused to a *GUS* reporter gene were created. Three constructs carrying a heterologous TATA box (the region up to –47 of the CaMV *35S* promoter) were also analyzed. Figure 1A and B show 10 of the most informative constructs studied.

For each construct at least five independent transgenic lines carrying one or more copies of the construct were tested. Seven to ten days old seedlings were induced with 0.1 μM 1NAA and then stained for GUS activity. The results from the observations are summarized in Figure 1C. The 2.6 kb of *AIR1B* 5'-flanking sequence conferred high *GUS* expression to the cortex and epidermis cells immediately

surrounding the site of lateral root emergence giving the appearance of a ring in the outer cell layers. Before lateral root emergence, GUS activity was observed in the cells overlaying the lateral root primordium even at very early stages of lateral root formation (see Figure 2c of Chapter 2). Upon auxin induction, GUS activity was observed in the outer cell layers all along the root except for the root meristem.

As shown in Figure 1C, most of the promoter deletions had a drastic effect on the expression of the *GUS* gene. However, in construct 1 the deletion of 150 bp containing region II clearly increased the *GUS* intensity before as well as after auxin induction. Similarly, deletion of this region had a positive effect on the *GUS* expression of plants carrying construct 4. This construct, containing only part of region III and the complete region I, showed some *GUS* expression after auxin induction. In contrast, plants harboring construct 3, containing region I, II and part of region III, did not show GUS activity after the same treatment. These results indicate that a repressor element is present within the 150 bp sequence containing region II.

The 1661 bp deletion at the 5' end of construct 1 (giving construct 2) reduced the *GUS* expression to a minimal level. Transgenic plants carrying this construct showed faint blue staining in the elongation zone of the root after auxin induction. Without auxin induction, no GUS activity was visible. The high *GUS* expression conferred by construct 1 and the low expression conferred by construct 2 indicate that sequences far upstream (between –939 and –2600 bp) contain elements conferring high *AIR1B* expression. Deletions presented by constructs 5 to 7 led to loss of all visually detectable GUS activity.

Replacement of the *AIR1B* TATA box for a heterologous TATA box (–47 bp CaMV, construct 8) did not disturb the expression of the *AIR1B* gene. Based in this observation we designed the constructs 9 and 10. Surprisingly, complete deletion of region I and II (construct 9) had no major effect on the expression of the *GUS* gene compared to plants carrying the complete –2600 bp sequence (construct 8). These observations indicated that all the elements necessary to confer the characteristic *AIR1B* expression are located between position -500 bp and -2600 bp of the promoter. The 350 bp of region I -either with its "own" or with a heterologous TATA box (construct 7 and 10, respectively) - was not enough to confer any GUS activity at all. Besides a TATA box, region I contains a putative CAAT box located 133 bp upstream of the ATG codon (Figure 3).

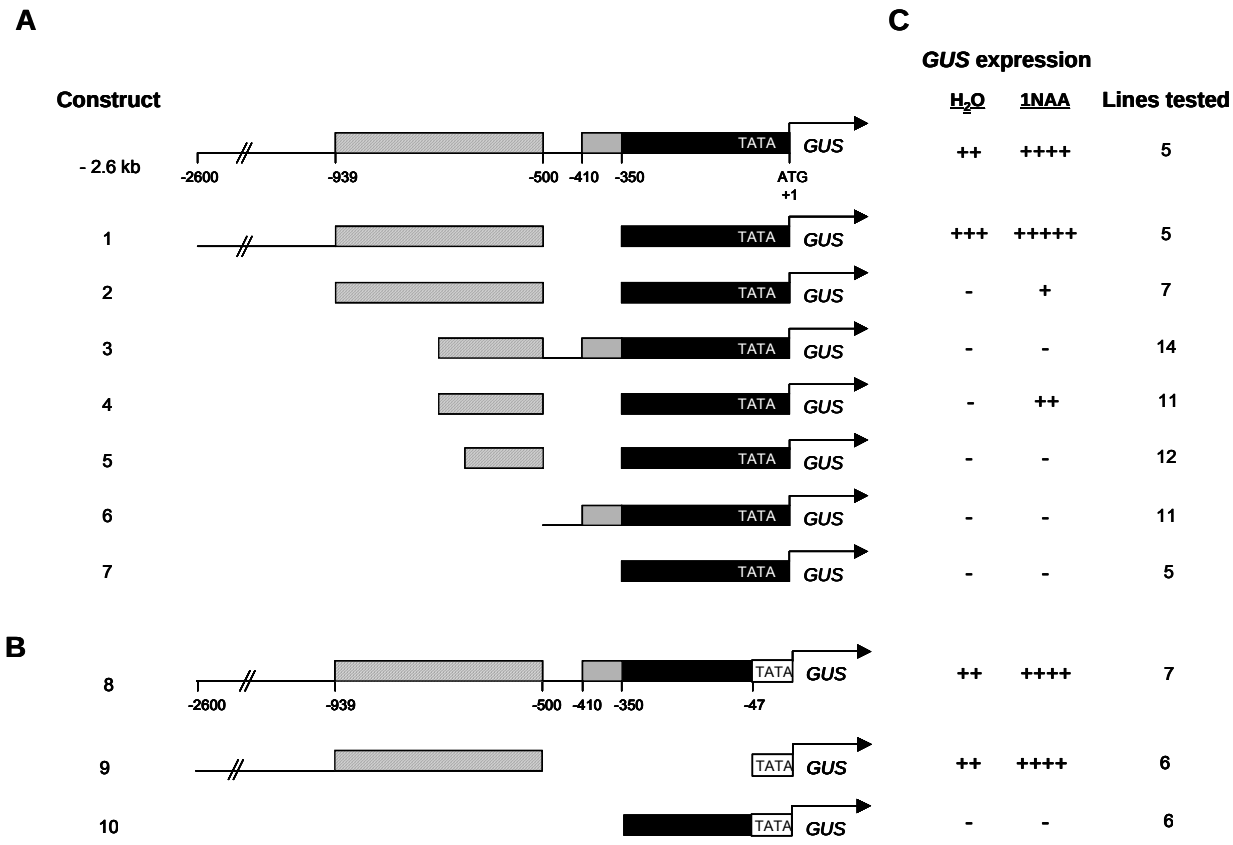

Figure 1. Schematic representation of *AIR1B* promoter::*GUS* constructs and relative *GUS* expression levels in roots of *Arabidopsis* transgenic lines.

Boxes indicate homologous regions in *AIR1A* and *AIR1B* promoters, black= region I, gray= region II, and hatched= region III. Distances are given in bp with respect to the ATG start codon. **A**, constructs with deleted promoter fragments transcriptionally fused to a *GUS* coding region. **B**, constructs containing a heterologous TATA box. Promoter regions were fused to a *GUS* construct containing a CaMV *35S* minimal promoter (only the TATA box) transcriptionally fused to the *GUS* coding region. **C**, *GUS* expression was analysed in several lines per construct after the indicated treatment. The relative expression levels were quantified visually and indicated with + and -. The range of expression goes from highly expressed (+++++) to no *GUS* expression (-). In the H_2O treatment (control) GUS activity was observed only at the sites of lateral root emergence. Therefore, (+) in the H_2O column indicates *GUS* intensity at these specific sites, while (+) in the 1NAA treatment indicates intensity of staining along the root.

Analysis of sequences upstream of region III

Our earlier obtained sequence data covered 2743 bp and 1128 bp of the 5' regions of *AIR1A* and *AIR1B*, respectively (Neuteboom, 2000). Since the complete sequence of the *Arabidopsis* genome (*Arabidopsis* Genome Initiative project, AGI) became known during the course of the experiments, we made use of this source to analyze sequences upstream of region III. When we aligned the complete 4.5 kb of *AIR1A* 5'-flanking sequence (separating the upstream *AIR1B* stop codon from the *AIR1A* start codon) with the 2.6 kb of *AIR1B* 5'-flanking sequence (Figure 2), we found two additional homologous regions: one of 435 bp (named region IV) and one of 343 bp (named region V). Both regions were separated by 15 bp in *AIR1A*. Thus the *AIR1A* and *AIR1B* promoters have in total five highly conserved regions (Figure 2 and 3). Considering the fact that deletion of sequences upstream of region III led to a very weak *GUS* expression (Figure 1), it is expected that regions IV and V possess the elements required for enhanced expression of the *AIR1A* and *AIR1B* genes. Further promoter-deletion and mutational analyses of these regions are necessary to determine whether sequences IV and V are indeed important for expression. Subsequently, they may be used in one-hybrid experiments to identify their binding transcription factors.

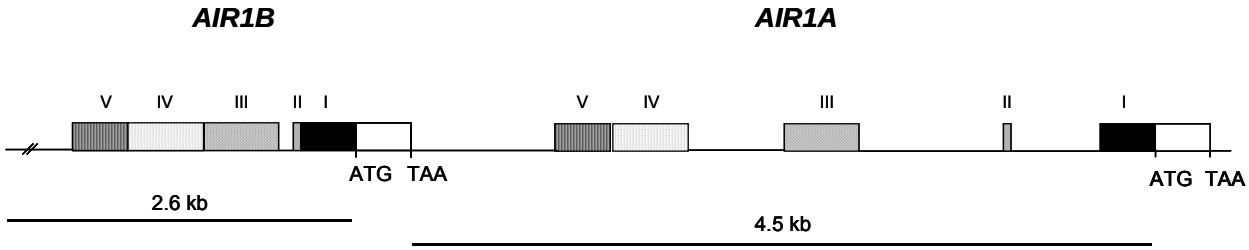

Figure 2. Schematic representation of the position of homologous region I, II, III, IV and V in the promoter of *AIR1A* and *AIR1B* gene. Open boxes represent coding regions.

Region V and IV

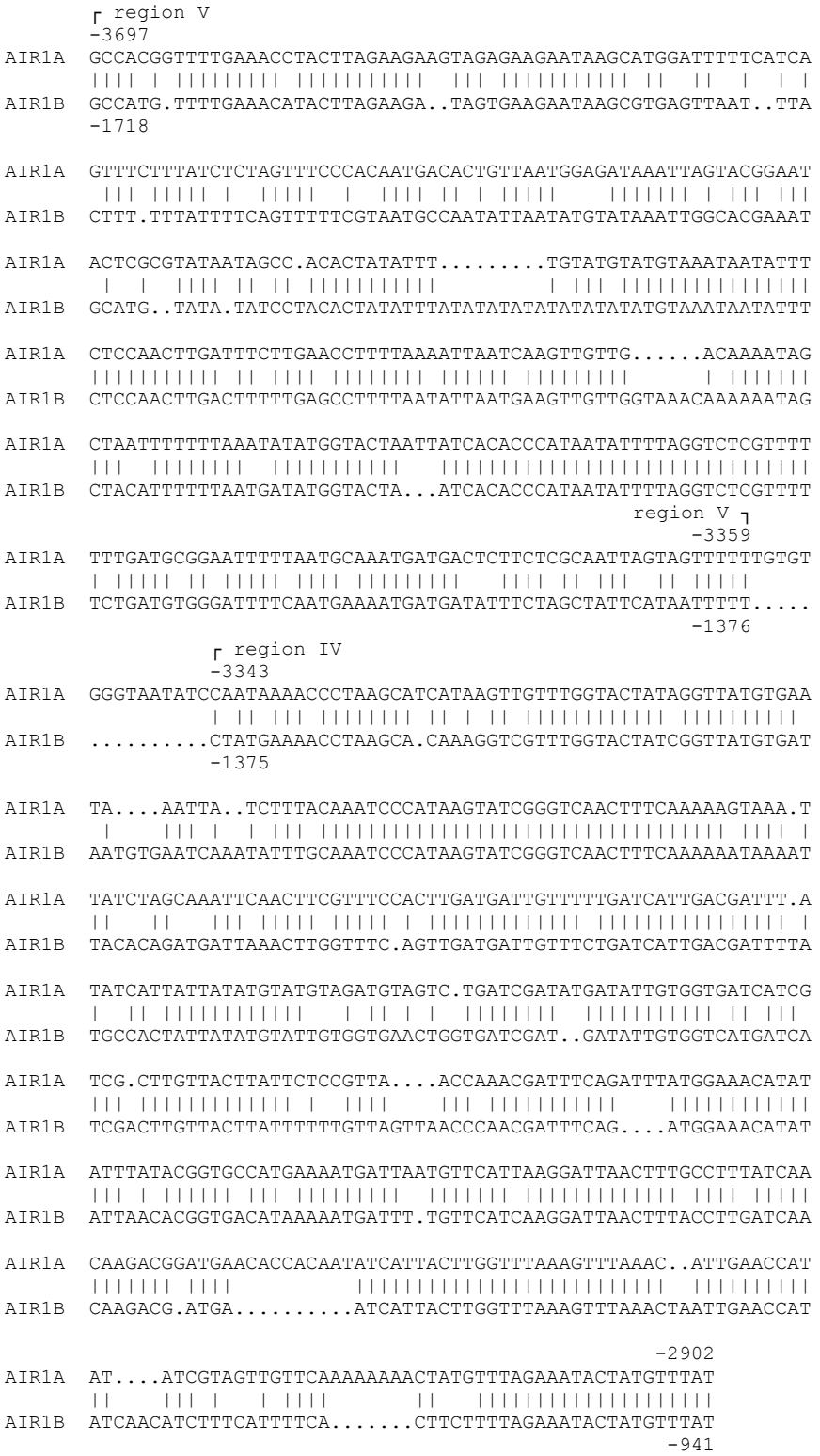

Figure 3. (This page and the next page). Sequence comparison of the *AIR1A* (upper line) and *AIR1B* (lower line) homologous regions I, II, III, IV and V of the promoter. Nucleotide position is indicated with respect to the ATG start codon. The TATA boxes and CAAT boxes are shown in boldface and the ATG transcriptional start codons are shown in italics. *AIR1A* and *AIR1B* 5'-flanking sequences upstream of region III were obtained from the *Arabidopsis* Genome Initiative database (T1P17 BAC clone). The alignments were made using the LALING program [X. Huan and Miller, (1991) Adv. Appl. Math. 12:373-381].

Region III

```
       -2277
AIR1A  AAAGAGAAAC.......AGAGCTTAATAGTAATACAGAAAGTTTTTTAAGTTAGACAAAT
       |||||||||        | |   | |||||||| ||||||||||  |||||||||||| |||
AIR1B  AAAGAGAAATTAATTATAAAATTAAATAGTAAAACAGAAAGTTAGTTAAGTTAGACGAAT
       -939

AIR1A  TCACTCTATCAATTTGACTTTTGATCTCAAATTTGTGTTAAATT.TTTTTGGGACAACTC
       |    |  ||   ||| |||||||||||||||||||||| ||| | ||||||||||  ||
AIR1B  T...TGGAT...TTTCACTTTTGATCTCAAATTTGTGTCAAACTCTTTTTGGGACATTTC

AIR1A  TAACGTTGAAATCATACGGCTGAAATATCACTCCCATT...TAAAAATA.ATAT...TCC
       |||||||| ||| ||||| |||| ||||  |||   ||   |||||||| ||||   |||
AIR1B  TAACGTTGGAATTATACGACTGAGATATACCTCTATTTAATTAAAAATATATATATATCC

AIR1A  AAACTAGTCTATCTGTCCATCTTGTAACCCCTGATATACGGAATTATTTGATAAGTCCAA
       |||||||||||| |||||||||||||||||||||||||| ||||||||| |||||||||
AIR1B  AAACTAGTCTAT.TGTCCATCTTGTAACCCCTGATATACAGAATTATTTAATAAGTCCA.

AIR1A  CCATTTGAGATTACAAAAACAATTATCACATGAAAATGATATCATAACTTTGTCTTCGTG
          |||| |||||||||||||||||||||||||||||| |||||||||||||||| |||
AIR1B  ...TTTGCGATTACAAAAACAATTATCACATGAAAATGAAATCATAACTTTGTCTTTGTG

AIR1A  TGAGACTGT.GTAGATTAACTTATGTTTTAGAAATAATTGTT..ATATAAGTCTGCCCCG
       ||||| |   ||| ||||| |||||||||  ||||||| |||  |||||||||| |||||
AIR1B  TGAGAGTTAAGTATATTAAATTATGTTTTTAAAATAATGGTTTTATATAAGTCTTCCCCG

AIR1A  TGACAAACGGATTAAT.GAAGGATAATAAGATATGTTCACATGGCCAACGTGAAAGCATA
       |||||||||| ||  | |||||   |||||| |||||||||||||||||||||| |||
AIR1B  TGACAAACGGGTTTGTTGAAGGGGCATAAGAAATGTTCACATGGCCAACGTGAAAGAATA

                             -1847
AIR1A  AAAAACAATTATTTATTTTTAGATATATTT
       || ||||||| ||  ||||| |||||  ||
AIR1B  AAGAACAATTGTTCCTTTTTCGATATCCTT
                              -501
```

Region II

```
        -958                                                    -908
AIR1A   TATATTTG...ATTAT...ATTTTTAGTAATCGTGAAGACTCGTTCATATAT..TGATA
        ||||||||   ||| |   | ||||| | ||||||||| | ||| |||||||  |||||
AIR1B   TATATTTGTCAATTTTTGAAATTTTAATCATCGTGAAGGCCCGTACATATATATTGATA
        -410                                                    -352
```

Region I

```
       -363
AIR1A  AATTCGTAAATTTTGAGAGTAGTTTTTAACCAACTTACACCACCAATTTATTAGAGTACA
       ||||| ||||||| | ||||| | | |||| |||||| |||||||| || |||||||| |
AIR1B  AATTCATAAATTT.GGGAGTAATGTCTAACAAACTTATACCACCAAATTGTTAGAGTAAA
       -350

AIR1A  TATTTTGATTTTCGATATGCAATCTTGTTACCAACTAAGTTGTTAATTAGGACCATATAC
       ||||||||||||||| |||||     |||||||||||||||||||||||| || |||||
AIR1B  TATTTTGATTTTCGACATGCA.....GTTACCAACTAAGTTGTTAATTAGAACAATATAT

AIR1A  TCTTAGAAATTGCCAGCTAATTTGAACAATGACCATTTAAGTGTTTTCAAATAATCAAAT
       | |||||||||||||| ||||||||||      |||||||||||| ||||||||||||||
AIR1B  TTTTAGAAATTGCCACCTAATTTGAA......CCATTTAAGTGTTCTCAAATAATCAAAT

AIR1A  CATATCCATATAAAAATAATTAGTCAACTAATCCCCTATTAAGAAAAATCAA**CAAT**TAAA
       ||||||  |||||||| |||||||||||||| |||||||||||||||||||||||||||||
AIR1B  CATATC..TATAAAAA.AATTAGTCAACTAAACCCCTATTAAGAAAAATCAA**CAAT**TAAA

AIR1A  ACATTCTCATTTC.CTCCTAATGAATACCC**TATAAA**TACCCACTAAAAACTTACATTTCT
       || |||||||||  |||||||||||||| |||||||||||||||||||| ||||| ||||||
AIR1B  ACCTTCTCATTTTTCTCCTAATGAATATCC**TATAAA**TACCCACTAAACACTTATATTTCT

AIR1A  C.CCACCAATCTCAAAGCAAATTAAATACACTACTACTTCTTGAGCTTTTAAACTACACA
       | |||||||||||||||||||| |||||||||||||||||||| |||||  |||| |||
AIR1B  CTCCACCAATCTCAAAGCAAATTATATACACTACTACTTCTTGATCTTTTGTACTATACA

          -1
AIR1A  AACAA*ATG*
       ||||
AIR1B  AACAT*ATG*
          -1
```

In each *AIR1B* promoter-deletion construct studied, the *GUS* expression pattern and intensity seen after MeJA induction was the same as after 1NAA treatment. These results indicate that regions III, IV and V are involved in the auxin and MeJA response of the gene. We searched for the presence of known auxin-responsive elements (Hagen and Guilfoyle, 2002) within the five homologous regions of the *AIR1A* and *AIR1B* promoters, but we did not find any. We also searched for the presence of known MeJA-response elements like JERE (Menke et al., 1999a), for the sequence motif GATAcAnnAAtnTGATG found in a MeJA-responsive lipoxygenase gene (*tomloxA*) from tomato (Beaudoin et al., 1997), for the *as-1* like elements implicated in MeJA-response (Xiang et al., 1996), and for the presence of a G-box with the CACGTG motif, a common feature in promoters from MeJA inducible genes (Rouster et al., 1997). However, such elements were not found.

BLAST comparisons of the sequences separating the five conserved regions in the *AIR1A* promoter with sequences from the *Arabidopsis* database were performed. A single homology was found within the 544 bp sequence separating region I and II; 450 bp of this sequence is found in all chromosomes and matches (90% identity) a small part of entry AX059539 corresponding to a large sequence (47 kb) present in centromeric regions (Preuss et al., 2000).

AIR3 promoter analysis

Expression of the *AIR3* gene in *Arabidopsis* was studied in plants containing 7 kb of *AIR3* 5'-flanking sequence fused to a *GUS* reporter gene (*AIR3*::*GUS* plants). *AIR3*::*GUS* plants showed blue staining in the cortex and epidermis cells surrounding the site of lateral root emergence and in places where lateral roots were about to emerge. Upon auxin treatment, GUS activity was observed along the root with the exception of the root meristem (Neuteboom et al., 1999b). In general, the *AIR3* gene shows the same expression pattern in the root as *AIR1A* and *AIRB* genes but the expression is less intense. In the upper (above ground) part of the plant *AIR3*::*GUS* expression was observed in the nectaries of flowers. Since this expression was auxin-independent, it was not further studied.

In order to identify promoter regions important for *AIR3* expression, we created 26 constructs containing 5' and/or 3' deletions of the 7 kb *AIR3* 5'-flanking sequence fused to *GUS*. The GUS activity in these transgenic lines before and after auxin

treatment was compared to the *GUS* expression observed in *AIR3*::*GUS* plants. Figure 4 shows a schematic representation of the most informative constructs studied.

In construct A, a large portion of the 7 kb region was deleted at the 5'-end resulting in a 1350 bp sequence fused to *GUS*. Plants carrying this construct displayed the characteristic expression pattern observed in *AIR3*::*GUS* plants although at a slightly lower level. After auxin treatment blue staining along the root was visible with more intensity in the elongation zone of the root. A consensus CAAT box and TATA box are found at 126 and 77 bp upstream of the ATG, respectively (Neuteboom et al., 1999b). The remaining 1350 bp of *AIR3* promoter sequence lacks clear similarity to any of the known auxin responsive elements.

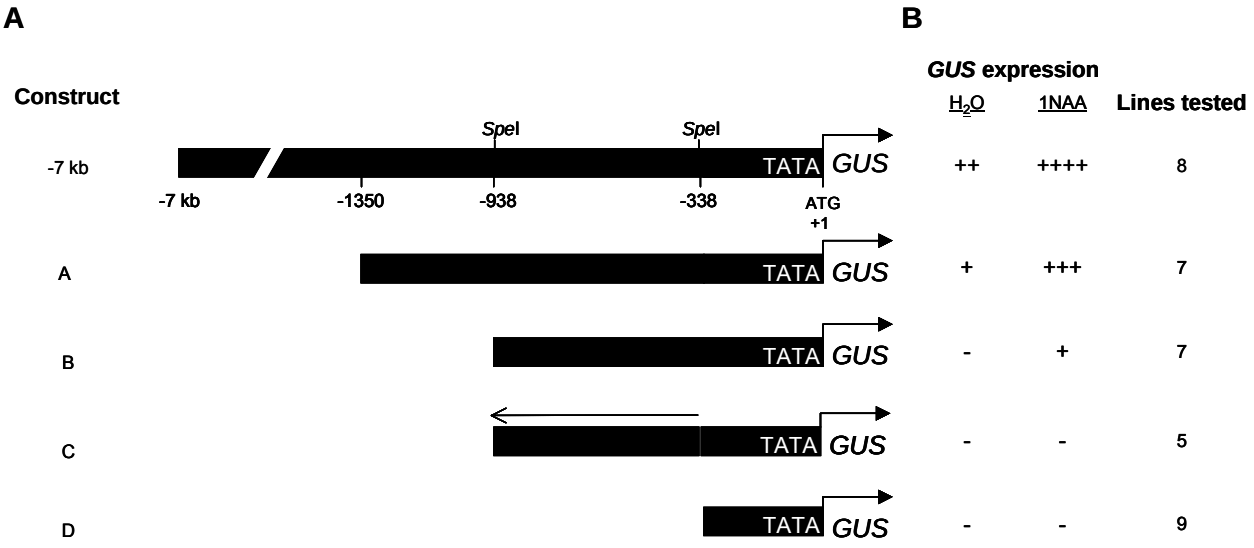

Figure 4. Schematic representation of *AIR3* promoter::*GUS* fusion constructs and *GUS* expression levels in roots of *Arabidopsis* transgenic lines.

A. Distances are given in bp with respect to the ATG start codon. Seven kb of *AIR3* 5'-flanking sequence was transcriptionally fused to a *GUS* reporter gene (= -7 kb construct). A, B and D represent 5'-deletion constructs. In construct C, the 634 bp *Spe*I / *Spe*I region was cloned in reverse orientation as indicated by the arrow. **B.** *GUS* expression in roots of transgenic lines after the indicated treatment. The relative expression levels were quantified visually and indicated with + and -. The range of expression goes from highly expressed (+++), (+) faint, to no *GUS* expression (-). With H_2O treatment (control) only weak blue staining at the sites of lateral root emergence was observed. After 1 NAA treatment *GUS* activity was observed all along the root.

The shortest auxin-responsive *AIR3* 5'-flanking sequence was 938 bp long, represented by construct B. Seedlings carrying this construct showed very weak GUS activity only after auxin induction. After inversion of the 660 bp *Spe*I-*Spe*I region GUS activity disappeared (construct C). Since 338 bp of 5'-flanking sequence (including the TATA box) was not sufficient to confer any GUS activity (construct D), it is expected that the *Spe*I-*Spe*I region contains elements required for a basal *AIR3* auxin responsiveness. Further deletions of the 938 bp upstream sequence (in the direction 5'→3') led to total absence of visible GUS activity (data not shown).

Localization of AIR1A, AIR1B and AIR3 in the Arabidopsis genome

The position of *AIR1A, AIR1B* and *AIR3* in the *Arabidopsis* genome was determined by searching in the *Arabidopsis* sequence database TAIR (The Arabidopsis Information Resource, (www.*Arabidopsis*.org/servlets/). The *AIR1A* genomic sequence (AF09863, EMBL database) was used for a BLAST search resulting in a match corresponding to the locus AT4g12550, located in the long arm of chromosome 4 between position 6,400.000 to 6,500.000 bp (Figure 5). *AIR1A* and *AIR1B* are contained in the BAC clone T1P17 (18a and 18b of Figure 6). The *AIR1A* start codon (ATG) is found 4.5 kb downstream of the *AIR1B* stop codon (TAA).

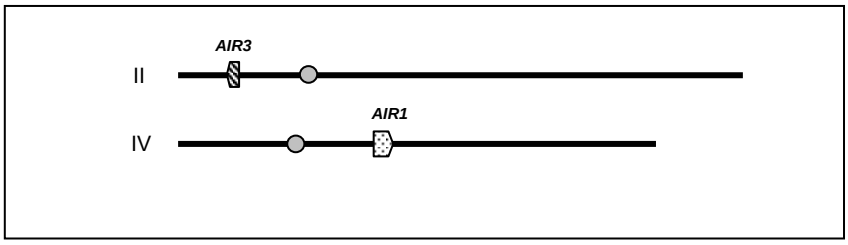

Figure 5. Relative position of *AIR1 (AIR1A and AIR1B)* and *AIR3* in the *Arabidopsis* genome. Roman numerals correspond to chromosome two and four respectively. The relative position of the centromere is indicated by the circle.

We used the same approach to determine the position of the *AIR3* gene (AF098632 EMBL-database). *AIR3* corresponds to the locus At2g04160 located in the short arm of chromosome 2 between position 1,400.000 to 1,500.000 bp (Figure 5).

Genes adjacent to AIR1A and AIR1B

The genes adjacent to *AIR1A* and *AIR1B* are shown schematically in Figure 6, with additional information in Table 1. Upstream of *AIR1B* (18b in Figure 6 and Table 1), the At4g12540 locus is found (17). This gene encodes a hypothetical protein containing a MYB DNA binding domain repeat signature. This gene is so close to *AIR1B* (2.6 kb from the *AIR1B* ATG) that most probably both genes share promoter sequences but in opposite direction.

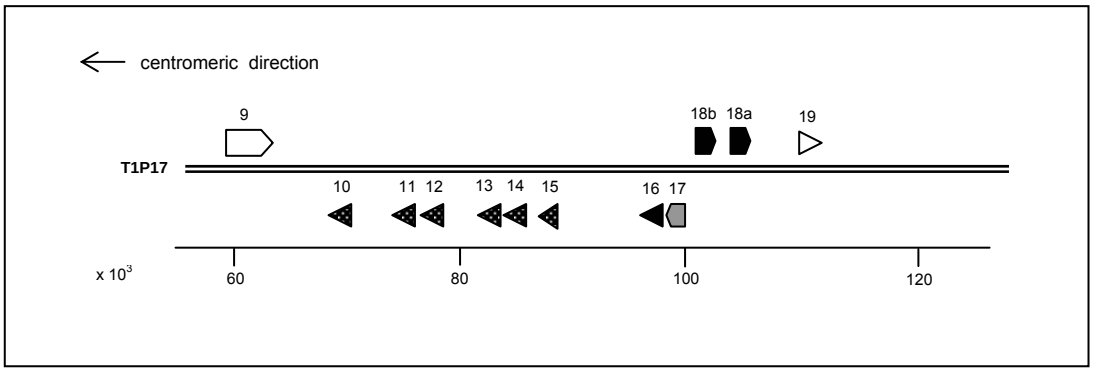

Figure 6. Localization of genes in BAC T1P17 between nucleotide 60,000 to 120,000 of the clone. Numbers indicate the position in the BAC sequence and arrows correspond to the forward or reverse orientation of the genes. A cluster of nine genes encoding proteins belonging to the CLCT family (genes 10-16 and 18a and 18b) is interrupted by gene 17 (in gray). Dotted arrows correspond to *pEARLI1* genes; black arrows correspond to *AIR1-like* (16), *AIR1B* (18b) and *AIR1A* (18a) genes; other genes are represented in white. The name of the genes and description is given in Table 1

Number	Gene ID	Description
9	At4g12460	putative protein
10	At4g12470	PEARLI1-like protein
11	At4g12480	PEARLI1
12	At4g12490	PEARLI1-like protein
13	At4g12500	PEARLI1-like protein
14	At4g12510	PEARLI1-like protein
15	At4g12520	PEARLI1-like protein
16	At4g12530	*AIR1A*-like protein
17	At4g12540	hypothetical protein with myb DNA-binding domain
18a/18b	**At4g12550**	**AIR1A and AIR1B.** putative cell wall-plasma membrane disconnecting CLCT proteins
19	At4g12560	Putative protein

Table 1. Genes adjacent to *AIR1A* and *AIR1B* in the *Arabidopsis* genome.
The numbers correspond to the position of the gene in BAC clone T1P17 as indicated in Figure 6. The gene name (ID) and description are given.

Immediately downstream of gene “17” the locus At4g12530 (16) is found. This gene is expected to encode a 117 amino acid CLCT protein with high similarity to

AIR1A and AIR1B, remarkably also lacking the proline-rich region between the signal peptide and the C-terminal part. No other member of the CLCT family of proteins lacking the proline- or glycine-rich region besides AIR1A and AIR1B has been described in *Arabidopsis* so far. This gene contains a single exon with 55% identity to AIR1A at the nucleotide level and 48% identity at the amino acid level (Figure 7). Comparison of 1.3 kb of *AIR1A-like* 5'-flanking sequence (separating the '16' and '17' ORFs) with 1.8 kb of the *AIR1B* promoter (regions I-V) showed no similarity.

Next to gene '16', a tandem of 6 genes (At4g12470 – 12520; genes 10-15 in Figure 6) encoding pEARLI1 and pEARLI1-like proteins is found. These putative proteins show high similarity with AIR1A, AIR1B and AIR1A-like (Figure 7). They belong to the CLCT family of proteins containing a putative signal peptide, a proline–rich region and a highly conserved C-terminal part (Richards and Gardner,1995). No other genes encoding "CLCT proteins" were found upstream or downstream indicating that the CLCT cluster consists of nine members: *AIR1A, AIR1B, AIR1A-like* and six *pEARLI1-like* genes.

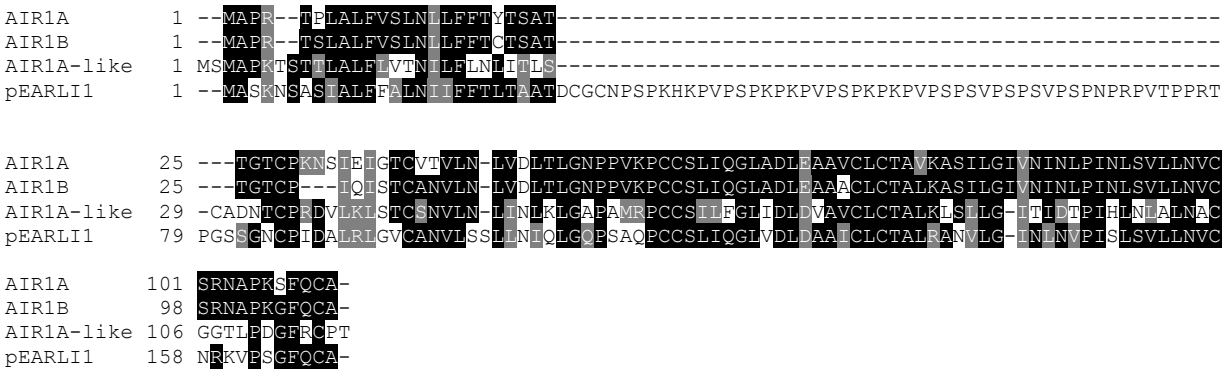

Figure 7. Amino acid alignment of four proteins from the CLCT cluster, AIR1A and AIR1B (At4g12550), AIR1A-like (At4g12530), and pEARLI1 (At4g12480). Identical amino acids are boxed in black, and similar amino acids are boxed in gray. AIR1A shows 92% of identity with AIR1B, 48% of identity with AIR1A-like and 54% of identity with pEARLI1 protein (including the proline-rich region).

The *pEARLI1* (At4g12480) gene is an early aluminium-induced gene from *Arabidopsis* which encodes a highly conserved and highly expressed proline-rich hydrophobic CLCT protein of which the function is unknown (Richards and Gardner, 1995; Richards et al., 1998). FASTA analysis of pEARLI1 (At4g12480) against all *Arabidopsis* proteins showed that seven predicted *pEARLI1*-like genes are present in

Arabidopsis, six located in this tandem on chromosome 4 and one gene located on chromosome 1 (At1g12090). pEARLI1 proteins show significant similarity to other structural cell wall proteins like extensins (Neuteboom et al., 1999a).

Downstream of the At4g12470 locus (10) and downstream of the *AIR1A* locus (18a), genes encoding unknown proteins and repetitive regions are found (Figure 6).

Genes adjacent to AIR3

AIR3 (At2g04160) is located in a heterochromatic region of chromosome 2. This region possesses low gene density and a relatively higher proportion of (pseudo)genes related to retroelements (Lin et al., 1999).

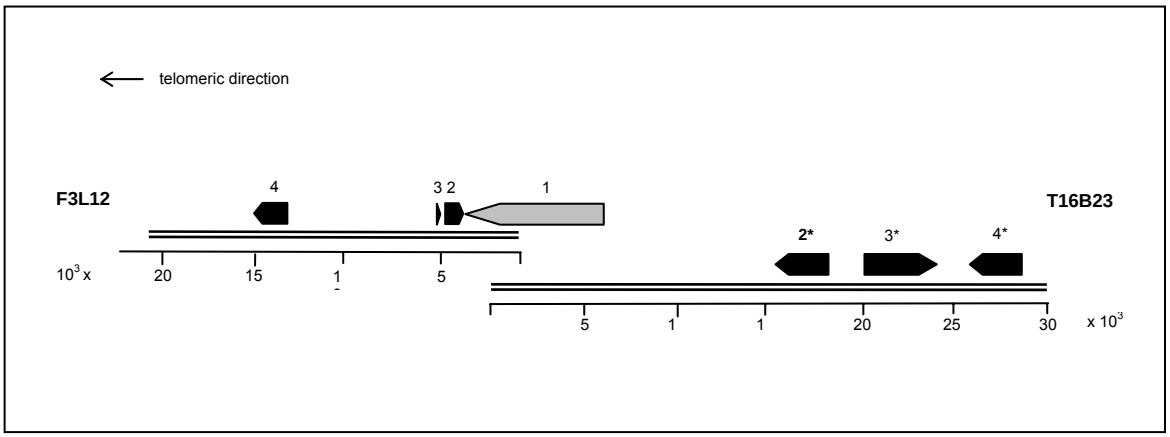

Figure 8. Position of genes in BAC's F3L12 and T16B23.
The *AIR3* gene (1, grey arrow) was found in the overlapping sequences F3L12 and T16B23. Arrows correspond to the forward of reverse orientation of the genes. The names and the descriptions of the genes are given in Table 2. Scale indicates nucleotide position in the BAC sequence.

Number	gene ID	Description
1	**At2g04160**	**subtilisin-like serine protease AIR3**
2	At2g04150	pseudogene, similar to hypothetical protein AAC24051
3	At2g04140	Putative retroelement pol polyprotein
4	At2g04130	pseudogene , retroelement pol polyprotein
2*	At2g04170	unknown protein
3*	At2g04180	pseudogene, reverse transcriptase
4*	At2g04190	unknown protein

Table 2. Genes adjacent to *AIR3* in the *Arabidopsis* genome.
The number corresponds to the genes in the BAC clone as indicated in Figure 8. Name (ID) and description are given.

The complete *AIR3* sequence including 1350 bp of 5' and 352 bp of 3' non-coding region was found in two overlapping BAC clones, F3L12 and T16B23 (Figure 8). The genes neighboring *AIR3* and displayed schematically in Figure 8 do not show any homology with *AIR3* (Table 2). These results indicate that *AIR3* is the only subtilisin-like serine protease gene present in this area of chromosome 2, although three more members of this family are predicted to be located elsewhere in this chromosome (Table 3).

Discussion

AIR1A, *AIR1B* and *AIR3* are members of a group of auxin-responsive genes from *Arabidopsis* expressed specifically at sites of lateral root emergence (Neuteboom et al., 1999a and b). How the expression of these genes is regulated is unknown. One of the strategies commonly used for the elucidation of signal transduction pathways leading to the expression of genes is the identification of responsive elements in the promoter. These elements may serve to isolate binding proteins and subsequently other components of the signal transduction chain.

AIR1A and *AIR1B* promoters possess five highly homologous regions I, II, III, IV and V. Their conservation could indicate their importance for the regulation of *AIR1A* and *AIR1B* gene expression. However, we found that two of these regions, II and III (from position -1 to –500 relative to the ATG) could be replaced by a heterologous TATA box without major effects on the expression levels of the *AIR1B* gene. We also found that region II had a negative effect on the (*GUS*) expression of different promoter-deletion constructs analyzed. Region III (between –501 and –939 with respect to the ATG) conferred a basal auxin-responsiveness to the *AIR1B* promoter. Regions IV and V (between –941 and –1718 with respect to the ATG) probably enhance the activity to the *AIR1B* promoter.

Most of the auxin responsive elements reported so far have been found in rapidly activated auxin-regulated genes (Hagen and Guilfoyle, 2002). Since *AIR1A* and *AIR1B* are late auxin-responsive genes, whose transcripts accumulate after 4 hours of auxin induction (Neuteboom et al.,1999a), it is not surprising that auxin responsive elements are absent in the promoters of the *AIR1* genes. Consensus sequences of

known MeJA responsive elements (Menke et al., 1999a; Beaudoin et al., 1997; Rouster et al., 1997) were also not found. Identification of jasmonate-responsive elements in the *AIR1A and AIR1B* promoters will be an important step towards the elucidation of the role of MeJA signal transduction pathways leading to the expression of these genes.

A segment of 1.3 kb of the *AIR3* 5'-flanking sequence is sufficient to confer the characteristic *AIR3* expression pattern. Although the *AIR1* and *AIR3* expression patterns are almost identical, no similarities were found between the 5'-flanking sequences of the *AIR1* and the *AIR3* genes. No known consensus regulatory sequences were found in the 1.3 kb of the *AIR3* promoter region analyzed besides a CAAT and a TATA box located at 126 and 77 bp from the ATG (Neuteboom et al., 1999b). The 1.3 kb promoter region possesses AT-rich repeated sequences and these kinds of sequences have been reported to act as silencer-like regions in the promoter of the ascorbate oxidase gene from pumpkin (Kisu et al., 1997]). If *cis*-acting sequences in the 1.3 kb region of the *AIR3* promoter would bind repressors, it could be expected that partial or total deletion of these sequences would lead to increased GUS activity. However, various 5' and 3' deletions of the 1.3 kb 5'-flanking sequence led to total absence of GUS activity, indicating that the AT-rich sequences do not represent repressor elements.

Promoter-deletion::*GUS* analyses showed that important regulatory sequences of the *AIR1* and *AIR3* genes are located relatively far (>500 bp) upstream of the transcription start site. Studies of many eukaryotic genes have shown that important *cis*-regulatory elements can be found scattered over large distances upstream of the gene coding region. For instance, the hormone-dependent expression of the uteroglobin gene from rabbit is mediated by two clusters of hormone-responsive elements located between 2.4 and 2.7 kb upstream of the transcriptional start site (Scholz et al., 1999). Fusion of the core (-35) promoter to this –2.4 to –2.7 enhancer region results in enhanced tissue-specific expression of this gene (Scholz et al., 1999). A more recent study (Mortlock et al., 2003), has show that the expression of the mouse *Gdf6* gene, which is involved in the patterning of embryonic skeletal and soft tissues, is regulated by a 2.9 kb fragment containing highly conserved regions located at a -60 kb distance.

Accumulative evidence indicates that the position of *cis*-regulatory elements is important to achieve their regulatory role. Locus activation does not take place on

naked DNA but in a chromatin environment where repressive histone-DNA interactions in the nucleosome have to be counteracted. The positioning of *cis*-acting elements within the nucleosome seems to be a key factor for the modification of the structure of surrounding chromatin by facilitating interaction with specific DNA-binding proteins and the recruitment of chromatin remodeling enzymes such as histone acetylases (HATs) (reviewed by Bonifer, 1999). Acetylation of lysine residues of the N-terminal tails of the histones H3 and H4 reduces their positive charge and weakens their affinity for DNA which, in turn, facilitates the assembly of the transcription machinery. Interestingly, HAT activity has also been observed in components of the basal transcription machinery (Struhl, 1998). In the same way that transcriptional activators can recruit HATs, transcriptional repressors can recruit different histone deacetylases. Histone deacetylation, as well as acetylation, can be promoter-specific or site-specific (Bonifer, 1999). Thus, we hypothesize that the relatively far-upstream localization of important regulatory sequences within the *AIR1* and *AIR3* promoter regions studied, may play a role in the regulation of the expression of these genes at the level of the chromatin structure.

AIR1A, AIR1B and AIR3 in the Arabidopsis genome

We made use of the *Arabidopsis* database (The *Arabidopsis* Genome Initiative) to determine the position of the *AIR1A, AIR1B* and *AIR3* genes in the genome and to identify their neighboring genes. *AIR1A* and *AIR1B* are located in a region of chromosome 4 with a high density of genes which are transcribed at high levels (Mayer et al., 1999). *AIR1A* and *AIR1B* are part of a cluster of genes encoding putative membrane proteins belonging to the CLCT family. A gene encoding a putative protein with a MYB-DNA binding domain interrupts the cluster between the *AIR1B* and an *AIR1A-like* gene. In the CLCT cluster, two subfamilies of genes can be distinguished, one of six members with a proline-rich stretch between the signal peptide and the C-terminal part and one of three members lacking the proline-rich region. The proline-rich subfamily of genes encode pEARLI1 and pEARLI1-like proteins, while *AIR1A*, *AIR1B* and the newly identified *AIR1A-like* gene (At4g12530) form the proline-rich region lacking subfamily. The C-terminal parts of all CLCT members are highly similar, which makes suggest that all originated from a common ancestor. FASTA searches show

that in *Arabidopsis* 19 CLCT gene members are present, 8 p*EARLI1-like*, 3 *AIR1A-like* and 8 extensin-like types.

FASTA analysis indicates that *AIR1A*, *AIR1B* and *AIR1A-like* are the only genes encoding CLCT proteins lacking a proline- or glycine-rich region in the *Arabidopsis* genome. In *Lithospermum erythrorhizon,* mRNA corresponding to a root specific proline-less CLCT gene with high homology to AIR1 has been identified (Yazaki et al., 2001). The corresponding gene, *LeDI-2,* is dark-induced and encodes a putative membrane-associated protein involved in the pathway of shikonin production, a secondary metabolite in the synthesis of red naphthoquinone pigments. LeDI-2 was found to be localized in the membrane of intra-cellular vesicles derived from the endoplasmic reticulum (ER). It is suggested that LeDI-2, provides stability or helps the vesicles containing shikonin (a lipid) in the transport from the place of synthesis to the outside of the cell where the content is secreted. Interestingly, AIR1 possesses a signal peptide that directs the protein to the ER as well. The localization of AIR1A in the cortical ER was corroborated by expression analysis of an *AIR1A* promoter::*AIR1:GFP* fusion construct in *Arabidopsis* (Neuteboom, unpublished results). The high homology of LeDI-2 with AIR1 and their similar subcellular localization might suggest a similar function. Thus one may speculate that AIR1 plays a role in membrane stabilization during the process of lateral root formation.

AIR1A and *AIR1B* are part of a cluster of genes encoding CLCT proteins. However, it is not known whether the functions of the members of this cluster are related. In *Arabidopsis*, high levels of *pEARLI1* mRNA accumulate after treatment with toxic levels of alluminium (Richards and Gardner, 1995). A common characteristic of all aluminium (Al) induced genes isolated so far is that they are also induced by other stresses and wounding in addition to Al, indicating their involvement in reactions to oxidative stress in the plant (Ezaki et al., 1995 and 1996; Richards et al., 1998 and references therein). *AIR1A* and *AIR1B* are neither wound- nor stress-induced (Neuteboom et al., 1999a), showing that they are not involved in the same responses as *pEARLI 1* genes. On the other hand, transcripts of *pEARLI1* and *pEARLI1-like* were not found in a microarray experiment to search for *Arabidopsis* auxin-induced genes in the root (Van der Zaal, unpublished), while transcripts of *AIR1* and *AIR3* were abundantly induced in the same experiment. These results indicate that the *pEARLI1* and *pEARLI1-like* genes are not co-induced by auxin together with the *AIR1* genes.

AIR3 is located in a position of chromosome 2 where little gene activity is found. Most of the genes located in this region are repetitive sequences and pseudogenes related to proteins found in retrotransposons (Lin et al., 1999). *AIR3* is a relatively weakly expressed gene (Neuteboom et al., 1999a). One could assume that its (low) expression level is due to its position in a relatively silent part of the genome. However, different *AIR3* promoter::*GUS* lines showed low *GUS* expression levels as well. This observation indicates that the low activity of the *AIR3* gene is not a consequence of its chromosomal position. Estimations of gene transcription levels of the *Arabidopsis* subtilisin-like serine protease family revealed that genes corresponding to this family are transcribed at a relatively low level (Beers et al., 2004). These estimations support the assumption of a cell-type-specific role of this type of proteases in the plant.

In *Arabidopsis*, 54 members of subtilisin-serine proteases (subtilases) have been annotated (Table 3). These genes are distributed over all chromosomes but most of them are found in chromosomes 1, 4 and 5. In contrast to *AIR3*, most of the other subtilase genes are found duplicated or in clusters of 3 to 5 members.

Table 3. Number of predicted subtilisin-like serine protease genes in *Arabidopsis*.

Chromosome	Number of genes
1	12
2	4
3	4
4	16
5	18
Total	54

The data are derived from a compilation made by T. Altmann (http://www.unifrankfurt.de/fb15/botanik/mcb/AFGN/altmann.htm).

Materials and Methods

AIR1B promoter-deletion constructs

The construction of 2.6 kb of *AIR1B* 5'-flanking sequence fused to a *GUS* reporter gene is described by Neuteboom (2000). This 2.6 kb 5'-flanking sequence

was used to generate the different promoter-deletion::*GUS* constructs made by cloning procedures except for construct 2. To make construct 2 (Figure 1), a PCR fragment containing the region III was generated with the primers 1030F: 5'-CGGGATCCGATGAATCATTACTTGG and 499R: 5'-GCGAATTCAGATCTAAGGATA TCGAAAAAGG aligning at the 5' and 3' border of region III, respectively. The 1030F and 499R primers introduced a *Bam*HI and *Eco*RI site respectively to allow the cloning of region III in front of a construct containing region I fused to *GUS* (construct 7 in Figure 1).

For constructs containing a heterologous TATA box (Figure 1B), a *Sal* I site was introduced (by PCR) in region I, 53 bp upstream from the *AIR1B* transcription start and 18 bp upstream from the *AIR1B* TATA box, to allow the cloning of 5' sequences in front of a –47 *GUS*XX vector containing the *35S* CaMV TATA box transcriptionally fused to the *GUS* coding region (Pasquali et al., 1994). All the *AIR1B* promoter-deletion constructs were cloned into the *Xba*I/*Xho*I sites of the binary vector pMOGλCAT (Pasquali et al., 1994).

AIR3 promoter constructs

The construction of a 7.1 kb segment of *AIR3* 5'-flanking sequence fused to a *GUS* reporter gene is described by Neuteboom et al., (1999b). Constructs containing deletions of this region were made by standard cloning procedures and introduced into the *Xba*I/*Xho*I sites of the binary vector (Pasquali et al., 1994).

Transformation of Arabidopsis thaliana

The *AIR1B* and *AIR3* promoter-deletion constructs in pMOGλCAT were mobilized to *Agrobacterium tumefaciens* strain MOG 101 (Hood et al, 1993) using a triparental mating procedure (Ditta et al, 1980). *Arabidopsis* plants were transformed using the 'floral-dip' method (Clough and Bent, 1998) with minor modifications. Seedlings from the F2 generation were used for *GUS* expression analysis. In all the experiments, al least 5 independent transgenic lines of each construct were analyzed to exclude variations in the *GUS* expression due to position effects. The histochemical staining for *GUS* was performed as described in Chapter 2.

Analysis of AIR1A, AIR1B and AIR3 sequences in the Arabidopsis genome

For these analyses the data available in the The Arabidopsis Information Resource (TAIR, Rhee et al., 2003) was used. Other relevant internet sites like TIGR (http://tigr.org) and MIPS (http://mips.gsf.de) were also used.

Acknowledgement

We thank Dr. W. Ferro for critical reading and useful comments on the manuscript.