Running head: siRNA degradation of plant resistance genes

Corresponding author:
Ulf Lagercrantz
Department of Ecology and Genetics, Uppsala University
Norbyv. 18D
752 36 Uppsala
Sweden
Email: ulf.lagercrantz@ebc.uu.se
Phone: +4618 4716418

Journal research area: Genetics, Genomics, and Molecular Evolution
Keywords: siRNA, NBS-LRR, R-genes, miRNA, conifer
A significant fraction of 21 nt sRNA originates from phased degradation of resistance genes in several perennial species

Thomas Källman1, Jun Chen1, Niclas Gyllenstrand2, Ulf Lagercrantz1
1Department of Plant Ecology and Evolution, EBC, Uppsala University, 752 36 Uppsala, Sweden
2Department of Plant Biology and Forest Genetics, Swedish University of Agricultural Sciences, Uppsala, P.O. Box 7080, SE-750 07 Uppsala, Sweden
Financial sources: TK acknowledge financial support by the European Linktree and TipTree (BioDiversa) projects to Martin Lascoux. The project was made possible via support from EC-funded project Noveltree (FP7-211868), and the Erik Philip-Sörensen Foundation.

Corresponding author:
Ulf Lagercrantz
Email: ulf.lagercrantz@ebc.uu.se
ABSTRACT

Small RNA (sRNA) including miRNA and siRNA are important in the regulation of diverse biological processes. Comparative studies of sRNA from plants have mainly focused on miRNA even though they constitute a mere fraction of the total sRNA diversity. In the present study we report results from an in-depth analysis of the sRNA population from the conifer Picea abies and compared the results to those of a range of plant species. The vast majority of sRNA sequences in P. abies can be assigned to 21 nucleotides long siRNA sequences, of which a large fraction originate from degradation of transcribed sequences related to NBS-LRR (Nucleotide Binding Site-Leucine Rich Repeat) type resistance genes. Over 90% of all genes predicted to contain either a TIR or NBS domain showed evidence of siRNA degradation. Data further suggests that this phased degradation of resistance related genes are initiated from miRNA guided cleavage, often by an abundant 22 nt miRNA. Comparative analysis over a range of plant species revealed a huge variation in the abundance of this phenomenon. The process seemed to be virtually absent in several species, including Arabidopsis thaliana, Oryza sativa and non-vascular plants, while particularly high frequencies were observed in P. abies, Vitis vinifera and Populus trichocarpa. This divergent pattern might reflect a mechanism to limit runaway transcription of these genes in species with rapidly expanding NBS-LRR gene families. Alternatively, it might reflect variation in a counter-counter defense mechanism between plant species.
INTRODUCTION

Small RNAs (sRNA) are abundant in plants, and a large portion of them has important roles in the regulation of diverse biological processes, including regulation of patterning and development, response to the environment, defence against pathogens and silencing of endogenous transposable elements (Vaucheret 2006; Chapman and Carrington 2007; Chen 2009; Ruiz-Ferrer and Voinnet 2009; Voinnet 2009). The common origin of all sRNA is double stranded RNA (dsRNA) that can be derived from a diverse set of sources, including transcription of inverted repeat structures, convergent transcription or virus replication, as well as transcription of genes containing a hairpin-structure by RNA polymerase II (miRNA) (Voinnet 2009). dsRNA can also arise from the action of specific RNA-dependent RNA polymerases (RDR). The resulting dsRNA is cleaved into short RNA duplexes typically between 18 and 24 nucleotides (nt) long by a dicer-like protein (DCL). After cleavage, a specific strand is incorporated into an RNA-induced silencing complex (RISC) that upon identification of a complementary sequence might inactive the target by cleavage or translational repression of mRNA. Alternatively, RISCs can target DNA for methylation of DNA or histones (Ramachandran and Chen 2008).

Numerous studies of both model and non-model plant species have revealed that the most abundant class of plant sRNA are short-interfering RNAs (siRNA) (Rajagopalan et al. 2006). These siRNA mediates silencing of both endogenous and exogenous genes. To date, the majority of endogenously produced siRNA have been shown to act on transposable elements and repetitive DNA through DNA and chromatin modification (Ramachandran and Chen 2008; Jin and Zhu 2010). In angiosperms these siRNA are typically 24 nt long and are processed by a specific dicer protein, DCL3 (Xie et al. 2004). These 24 nt siRNA are dependent on RDR2 for their generation and AGO4 (argonaute 4) for their function in DNA or histone methylation (Xie et al. 2004; Kasschau et al. 2007). This pathway also requires the plant specific RNA polymerases Pol IV and Pol V (Haag and Pikaard 2011). Pol IV has a major role to initiate the generating 24 nt siRNA, that are guided to sites transcribed by Pol V.
The second most abundant class of sRNA is miRNA that regulate target transcripts by cleavage or translational repression (Brodersen et al. 2008). miRNA originate from DNA-encoded transcripts that form imperfect fold-back structures. miRNA genes produce pri-miRNA transcripts that are cleaved, mainly by DCL1, to produce miRNA-miRNA* duplexes. Thereafter, the miRNA is selectively incorporated into the RISC. Plant genomes typically contain hundreds of MIR genes i.e. regions encoding pri-miRNA (Cuperus et al. 2011).

Besides these general and shared features of sRNA populations among plant species, there are also major differences, but as many of the available analysis tools to study sRNA rely on already identified sRNA sequences in other species, novelties in individual species has, until recently, been largely overlooked. Further, we have a much better understanding of the biogenesis and function of specific groups of sRNA e.g. miRNA and as a result the vast majority of sRNA studies report putative miRNA sequences and their gene targets and more seldom discuss other groups of sRNA sequences. In two recent studies of short RNAs from Medicago truncatula and Solanum lycopersicum (Zhai et al. 2011; Shivaprasad et al. 2012) several novel aspects were identified and revealed that a substantial fraction of the total population of sRNAs in these species can be classified as siRNA originating from NBS-LRR (Nucleotide Binding Site - Leucine Rich Repeat) genes. Similar findings have not been reported in the model plants Arabidopsis thaliana and Oryza sativa. Further, these siRNAs were often observed in 21 nt spaced (phased) intervals and initiated through cleavage by 22 nt long miRNA, analogous to the generation of tasiRNA (trans-acting siRNA) (Zhai et al. 2011).

As NBS-LRR genes are key genes in plant defence systems this observation has posed an intriguing hypothesis that such miRNAs could be part of a counter-counter defence system. Support for this hypothesis was recently found in a detailed study in tomato where miRNA mediated silencing of NBS-LRR genes was reduced in plants infected by virus and bacteria (Shivaprasad et al. 2012). An alternative explanation could be that the mechanism evolved to constrain the expression of a rapidly expanding and evolving gene family in a similar fashion that has been proposed for the PPR (Pentatricopeptide) gene family in Arabidopsis (Sunkar and Zhu 2004; Cuperus et al. 2011). These differences in siRNA mediated degradation of
NBS-LRR genes between plant species is intriguing and warrants further studies of plant sRNA in additional species.

In this study we report a detailed characterization of the sRNA population from the gymnosperm species *Picea abies*. In line with previous reports from Pinaceae species we found a very low abundance of 24 nt sRNA (Morin et al. 2008; Dolgosheina et al. 2008), and the most abundant group of sRNA in spruce is 21 nt long. Among these 21 nt sRNA we could identify and validate several miRNA sequences, but the majority of *P. abies* 21 nt sRNA could be characterised as siRNA. More than 25% of all 21 nt short reads collected originated from phased degradation of NBS-LRR genes. In an attempt to better understand this massive production of short reads from NBS-LRR genes we broadened our study and re-analysed short read data from a diverse set of plant species with emphasis on reads that originates from this type of genes.

**RESULTS**

*Picea abies* sRNA

Small RNA was isolated from newly flushed buds of *Picea abies* and more than 10 million sequence reads were generated from the resulting libraries. After quality filtering, 8.3 million reads were retained for further analysis. Constraining the reads to a length between 18 and 29 nt (7 million reads) and collapsing them into unique sequences yielded 59,616 entities. The copy number of each sequence was highly variable, and the most abundant sequence occurred in 459,908 copies and 1,160,718 sequences occurred less than 10 times. The clearly most abundant size class was 21 nucleotides (nt), constituting 72% of all reads with a size between 18 and 24 nt (Figure 1). The 22 nt class was the second most abundant class (12% of the reads), while the 24 nt class comprised only 1%. The 5’-terminal nucleotide is important for sorting of sRNA to Argonaute complexes in angiosperm plants. In Arabidopsis, the majority of miRNA possess a 5’-uridine, which directs those to AGO1 (Mi et al. 2008). Among all *P. abies* sRNA sequences, no bias for a particular 5’-nucleotide was evident. Calculating the
percentage of each 5’-terminal nucleotide for sRNA of different lengths, indicated an over-representation of uridine for 22 nt sRNA only (36% of all 22 nt sRNA starts with a 5’-Uridine, Figure 2).

24 nt siRNA are known to guide DNA methylation and heterochromatin formation of repetitive and transposable elements in angiosperms (Herr 2005; Morin et al. 2008). As such 24 nt siRNA seems to be rare or absent in *P. abies*, we wanted to investigate to what extent other sizes of Pinaceae sRNA might target repetitive DNA, and mapped *P. abies* sRNA to four *P. abies* and two *P. glauca* BACs (in total 657 kb; De Paoli 2006; Hamberger et al. 2009). Allowing one mismatch and multiple hits resulted in 1% of the sRNA sequences mapping to these references. We also mapped *Pinus contorta* sRNA to genome sequence data from 10 annotated *Pinus taeda* BACs. Despite an estimated high repeat content (Kovach et al. 2010), a low number of sRNA mapped to the genomic sequence of the 10 BACs (in total 1 Mb sequence). Mapping 134,580 sRNA and allowing for 2 mismatches resulted in only 0.5% of the reads mapping, and no clustering of reads on repetitive elements was discernible. These results suggest that the role of 24 nt sRNA in angiosperms, to silence repetitive and transposable elements, is not to a large extent conducted by other types of sRNA in Pinaceae.

**Picea abies** miRNA

The 59,616 unique sRNA sequences and 166,685 *Picea* EST clusters were used to identify potential miRNA and miRNA* and their potential precursor sequences. This identified a large number of potential miRNA/miRNA* that based on sequence similarity could be grouped into 137 families. 50 of these showed high similarity (<=2bp mismatches) to the plant miRNAs in miRBase (Kozomara and Griffiths-Jones 2011), including 26 miRNAs previously reported for *P. abies* by Yakovlev et al. (2010). Among the 87 novel putative families, reads corresponding to both the miRNA and the miRNA* sequence were identified in our short read data for 17 of them (Supplemental Table S1). While the vast majority of miRNA families also identified in angiosperm were associated with a most abundant read of 21 nt in our data (17
out of 18), a larger fraction of those identified previously only in Pinaceae were associated with a most abundant read of 22 nt (18 out of 33).

Target prediction for the 137 potential miRNA/miRNA* sequences identified a large number of putative targets (Supplemental Table S1). Without additional data these predictions should not be taken as strong evidence of true targets, but we noted that in almost one third of the miRNA families with a predicted target, at least one of the targets is annotated as a resistance genes. This targeting of resistance genes is even more pronounced in the families that seem specific to Pinaceae, whereas we found no example of resistance targets among the families that is shared between angiosperms and gymnosperms. Furthermore, among the predicted 87 novel miRNA identified in the present study, 22 were predicted to target resistance genes (Supplemental Table S1).

**Degradation of NBS-LRR genes is a main source of sRNA in *P. abies***

To further characterise the sRNA sequences the collapsed reads were mapped to the *Picea* “EST clusters” from the gene index project (SGI, that in total contains 79,409 clusters). Allowing for multiple hits and one mismatch resulted in mapping of 57% of all sRNA reads (39% of all unique sRNA sequences) to 48,242 clusters. Out of those clusters, 14,231 contained a single mapped sRNA sequence, while 10 or more different sRNA sequences mapped to each of 11,189 clusters (Supplemental Table S2).

As the majority of EST clusters are likely to code for proteins, a surprisingly large number of EST clusters were characterised by a substantial number of reads mapping to each of them. The maximum number of unique sequences mapping to an individual EST cluster was 33,862, which corresponds to 544,261 reads including redundant identical sequences. The mean and median number of reads (including redundant ones) per EST cluster was 468 and 3, respectively. As we allowed for multiple hits, these numbers do not truly reflect the number of reads originating from each EST cluster, but at least suggest that a multitude of sRNA sequences originate from each of a substantial number of different genes corresponding to...
these EST clusters. To obtain a rough minimum estimate of the number of genes spawning these massive numbers of sRNAs, we aligned the sRNA sequences without mismatch and retained only reads with a single hit in the reference. With these parameters, 11,806 EST clusters from the SGI *Picea* gene catalogue obtained at least one hit, and 1,605 of those were hit by 10 or more sequences. The pattern seen for sRNA from newly flushed buds was further compared to data from three other tissues (needles, immature female buds and lateral bud meristem) and the correlation of number of mapped reads per EST cluster was in all cases above 0.9 (data not shown), showing that the observed pattern is not specific for flushing buds.

Examining the annotation of the EST-clusters to which high numbers of sRNA reads mapped identified a large proportion of genes annotated as NBS-LRR resistance genes. Domain prediction and comparison to known conifer NBS-LRR genes further supported a very large number of them as putative NBS-LRR sequences (Supplemental Table S2). Out of the 100 EST-clusters with the highest number of different sRNA sequences mapping to them, 73 were classified as NBS-LRR sequences and of the 4.7 million reads that in total mapped to SGI *Picea* EST clusters, 2.1 million (44%) mapped to clusters for which the sequence showed high similarity to NBS-LRR genes (Supplemental Table S2). Restricting sRNA reads to those with a length of 21 nt reinforced this observation further: 90 out of the 100 genes with highest read counts were classified as NBS-LRR type sequences, and out of 3.3 million 21 nt reads that mapped to SGI EST clusters 1.6 million (48 %), mapped to clusters classified as NBS-LRR type sequences.

Figure 3 shows the distribution of the number of 21 nt reads per EST cluster, illustrating the preponderance of NBS-LRR type among those with high read counts. As NBS-LRR genes constitute around 1% of all EST clusters, these data reveal a strong enrichment of NBS-LRR genes among those with many sRNA reads mapping to them.
As many of the EST clusters in our reference are most likely not full length, we also mapped our sRNA reads to the newly relased *Picea glauca* gene catalogue (that contains more than 23 thousand mRNA sequences that are annotated as full length; Rigault et al. 2011) and a full-length TIR-NBS-LRR gene from *P. sitchensis* (ACN49932, Liu and Ekramoddoullah 2011). Mapping also to these references confirms the observed pattern and genes with similarity to NBS-LRR type resistance genes, typically had unique reads mapping to them at multiple positions. An example is shown in Figure 4, where the EST cluster from the *P. glauca* gene catalogue with the largest number of mapped unique reads is shown. In total 3,547 different sRNA sequences mapped to this cluster, and on average each sequence mapped 7 times with a range from 1 to 12. This pattern is partly explained by the highly repeated structure of the reference sequence, with 12 tandem repeats of length 72 nt potentially coding for leucine rich repeats (Figure 4).

Figure 5 shows the pattern observed for the *P. sitchensis* full length TIR-NBS-LRR gene (after compensation for the fact that many reads map to multiple positions by dividing the short read counts by the number of places it mapped to). The obtained coverage distribution suggests that the sRNA sequences are derived from a large portion of the transcript, but some areas might spawn more sRNA than others.

The high numbers of sRNA mapping to NBS-LRR type genes suggest that such genes are generally more subjected to siRNA degradation. If this were the case we might expect low levels of intact mRNA from such genes. To compare estimates of mRNA levels we mapped six Illumina DGE libraries to the SGI EST clusters (Supplemental Table S2). Cluster with more than 100 mapped sRNA sequences had a mean of 93 DGE-tags/1000bp, while those with less than 100 sRNA sequences had a mean of 11,931 DGE-tags/1000bp. Thus, in general abundant sRNA mapping is associated with low mRNA levels. For genes with similarity to NBS-LRR type resistance genes, the corresponding number was 91 and 313 DGE-tags/1000bp for those with more and less than 100 mapped sRNA reads, respectively. Thus,
NBS-LRR genes seem to have generally low mRNA levels, but those spawning high numbers of sRNA do not seem to show strikingly lower mRNA levels.

The reports on phased degradation of NBS-LRR genes in angiosperms prompted us to search for phased 21 nt siRNA originating from *P. abies* NBS-LRR loci, applying the phasing algorithm used by Zhai et al. (2011). 208 loci had a phase score larger than 15 and more than 50% of these loci contained either NBS or LRR domains and another significant portion lacked a significant blastx hit against protein databases some of which might constitute tasiRNA loci (Supplemental Table S2).

As repetitive structures, e.g. LRR repeats, in the reference sequence might obscure phasing signals we also focused on genes comprising TIR or NBS but no LRR domains and with more than 100 unique sequences mapping to them. For these the average phase scores were 15 and 12, respectively, compared to the average of 9 observed for all genes (with more than 100 unique sRNA sequences mapping to them) (Supplemental Table S2).

A few highly abundant 22 nt miRNA families that triggered siRNA generation from multiple NBS-LRR genes were identified in *Medicago* and *Solanacea spp.* (Zhai et al. 2011; Li et al. 2012; Shivaprasad et al. 2012). Many of these belong to the related families miR482 and miR2118. Family miR482 was identified also in *P. abies*, where sRNA sequences matching both miRNA482 and miR2118 were identified and were by our pipeline included in the miR482 family. Of the 38 predicted target ESTs for this family, only two were annotated as an NBS-LRR genes (Supplemental table S1).

Focusing on genes with TIR or NBS domains, several examples of genes with a significant phase score and 22 nt sRNA mapping in position consistent with them triggering phased 21 nt sRNA were found. Figure 6 shows data from TC171280, which is predicted to contain an NBS domain but no LRR repeats (Supplemental Table S2). A high phase score was observed with the major phase consistent with the mapping of 22 nt sRNA1865. Additional 22 nt
sRNAs map to downstream in positions that indicate triggering of phased 21 nt sRNA in alternative phases. Figure 7 shows data for transcript TC170604, which is predicted to contain a TIR domain, but no NBS or LRR domains (Supplemental Table S2). Again a 22 nt sRNA, sRNA13 map in a position consistent with the major phase for 21 nt sRNA.

Among the 22 nt sRNA sequences, a clear majority seems to be specific to *Picea*, and of the 52 most abundant 22 nt sRNA sequences (those represented by more than 1,000 reads), 19 were predicted by our stringent pipeline to be miRNA. Furthermore, 20 of them mapped to one or more EST clusters classified as NBS-LRR sequences (in the expected orientation, and with a maximum of three mismatches). Among the 10 most abundant 22 nt sRNA, seven were predicted miRNAs, and nine were predicted to target NBS-LRR like sequences (Table 1). Out of those nine, and six target regions predicted to code for TIR domains, whereas none were predicted to target an NBS domain. Another set (14) of the 52 most abundant 22 nt sRNA sequences mapped to EST clusters that displayed high phase scores but were not classified as NBS-LRR sequences. Only one of those showed significant homology to plant proteins and the rest might thus represent tasiRNA.

As one particular TAS gene (TAS3) has been show to be conserved from mosses to angiosperms, we specifically searched for evidence of a conserved miRNA390-TAS3-ARF pathway in spruce. Among the 14 predicted targets of *P. abies* miRNA390, one EST-cluster displayed a high phase score suggesting siRNA generation (TC161203; Supplemental Table S2). Alignment of TC161203 with TAS3 sequences from Arabidopsis and *Nicotiana tabaccum* revealed conservation of two miRNA390 target sites flanking a conserved area of predicted tasiRNA production (Figure 8). This pattern has previously been reported as strong evidence for conservation of the miR390-TAS3-ARF pathway in multiple plant species, including *Pinus taeda* (Axtell et al. 2006; Krasnikova et al. 2009).
Amounts of sRNA originating from NBS-LRR genes vary largely between species.

The high abundance of 21 nt sRNA mapping to NBS-LRR genes in *P. abies*, and a recent report of similar findings in *Medicago* and *Solanum*, prompted us to survey this phenomenon in additional plant species. To facilitate comparisons across species, we first choose to map 21 nt sRNA data from 16 plant species to databases derived from “the gene index project”. Focusing on NBS-LRR type resistance genes (i.e. those EST clusters containing one or more of the following domains: TIR, BED, NBS, LRR), the patterns were highly variable across species. When allowing for one mismatch and multiple hits for each sRNA, around 90% of all *P. abies* genes containing TIR, NBS or BED domains were hit by more than 10 unique sRNA reads. Focusing on the NBS domain containing genes, a high percentage was also obtained in *Amborella, Medicago, Gossypium, Populus* and *Vitis*, while a low percentage of genes hits by 21 nt sRNA was generally found in monocots and non-angiosperm species (except *Picea*) (Figure 9A). Excluding *Picea*, the highest percentages were obtained for the NBS domain, followed by TIR in most species. It should be noted that NBS-LRR genes with TIR domains are rare in monocots (Tarr and Alexander 2009). Mapping 21 nt reads with zero mismatches and retaining only unique hits resulted in a similar pattern across species albeit a generally lower percentage was observed (Figure 9B).

To further compare the sRNA populations of different species, we used sRNA data from four angiosperm species for which genome sequence data is available, two perennial (*Populus trichocarpa* and *Vitis vinifera*), and two annual (*Arabidopsis thaliana* and *Oryza sativa*).

**Arabidopsis**

For the Arabidopsis data, containing close to 300,000 unique sRNA sequences, 36% mapped to 11,859 TAIR cDNA models. A single sRNA mapped to 4040 cDNA models, while 4,356 models spawned more than 10 different sRNA sequences. The maximum number of unique sRNA sequences mapping to an individual cDNA model was 8,157. Among the 500 annotated cDNA models with the highest read counts, 489 were classified as transposable
element gene, while none of those cDNA models were annotated as any type of resistance gene. Mapping only sRNA of 21 nt, still resulted in a predominance of transposable elements among cDNA models with high sRNA read counts (Supplemental Table S3). In addition a few PPR and tetratricopeptide repeat (TPR) proteins as well as transacting siRNA (TAS) genes were identified with high 21 nt sRNA read counts. Calculating the ratio of 21 nt to 24 nt reads mapping to each cDNA model, identified PPR and TPR as those with the highest ratio (Supplemental Figure S1; Supplemental Table S3). High ratios were also typically obtained for TAS genes and miRNA genes. This group also contained a single TIR-NBS-LRR class disease resistance gene.

Oryza

75% of 1,751,364 reads representing 31,474 unique sRNA sequences originating from rice seedlings mapped to O. sativa transcripts (V6.1). The maximum number of unique sRNA sequences mapping to a single transcript was 6,424, and 5,502 transcripts spawned more than 10 unique sRNA sequences. Although many transcripts were annotated as ordinary protein coding genes, a vast majority of transcripts spawning high numbers of sRNA showed similarity to sequences in the TIGR Oryza Repeat Database (Supplemental Table S4). The clearly most prominent type of repeat identified was MITE (171 out of 200 with the highest read counts; Supplemental Table S4). This suggests that many putative protein-coding genes contain short MITE elements that spawn high numbers of sRNA. Additional annotations among transcripts spawning high numbers of sRNA included rDNA like sequences and retrotransposons. For transcripts with high read counts the ratio of 21 nt to 24 nt sRNA was close to 1 (0.89 for transcripts with more than 100 sRNA sequences mapping). The few NBS-LRR type with a significant number of sRNA mapping (e.g. LOC_Os11g12330, Supplemental Table S4) also show similarity to sequences in the TIGR Oryza Repeat Database, and 24nt sRNA generally dominated over 21 nt sRNA at these loci.

Populus
Mapping 2.1 million sRNA sequences from xylem tissue to *P. trichocarpa* transcripts (version v. 2.0) allowing one mismatch and multiple hits, resulted in 10% mapped reads (27% of all 21 nt sRNA and 5% of all 24 nt sRNA). A single sRNA read mapped to 8,723 transcripts, 4,452 transcripts spawned more than 10 sRNA reads, and the maximum number of reads mapped to a single transcript was 17,157. Among the top ranking genes were those with annotations relating to NBS-LRR resistance genes and retroelements (*e.g.* ORF V protein, AP endonuclease/reverse transcriptase, non-LTR retroelement reverse transcriptase, Polyprotein, *etc.*, Supplemental Table S5). For 21 nt sRNA, there was a predominance of NBS-LRR type resistance genes, while for 24 nt sRNA a majority of genes were annotated as transposable elements. In particular, sorting the genes on the ratio of 21 nt to 24 nt sRNA mapping to them, the vast majority of genes with high ratio were NBS-LRR type resistance genes (among the top 100 genes, 91 were annotated as NBS-LRR resistance genes). Table 2 shows the ratio of 21 nt to 24 nt sRNAs mapping to the different categories of genes that were found among those with high read counts. Except NBS-LRR resistance genes a couple of other gene classes (GRAS family transcription factors and Xylem serine proteinase 1) showed high ratios, while those with annotation relating to transposable elements showed ratios below one.

To study the genomic sRNA distribution, the 21 nt and 24 nt sRNA reads were also mapped to scaffold 1 of *P. trichocarpa* genome version 2.0 (length 48.4 Mb). For both types of sRNA 27% mapped to scaffold 1. In general most reads mapped to intergenic regions (93% of 21 nt sRNA, and 97% of 24 nt sRNA), and distributions for 21 nt and 24 nt sRNA were largely similar (Supplemental Figure S2). Focusing on NBS-LRR type resistance genes, the pattern observed for mapping sRNA to transcripts, was evident also when mapping to scaffold 1. Supplemental Figure S3 shows a region with a cluster of resistance genes, where a clear pattern of 21 nt sRNA as opposed to 24 nt sRNA mapping to those genes is observed. This cluster also contains a gene annotated as transposase (POPTR_0001s42600.1) for which the opposite pattern is observed with mainly 24 nt sRNA mapping. For the NBS-LRR resistance
genes, the majority of 21 nt sRNA mapped to exons, indicating that most sRNA are derived from processed transcripts (Supplemental Figure S4).

Vitis
sRNA isolated from leaf and stem comprising 114,688 unique sequences were mapped against V. vinifera mRNA (v. 2). Allowing one mismatch and multiple hits resulted in 10% mapped sRNA sequences (17% of 21 nt sRNA and 4% of 24 nt sRNA). A single sRNA read mapped to 3,887 transcripts, 552 transcripts spawned more than 10 sRNA sequences, and the maximum number of sequences mapped to a single transcript was 919. The majority of mRNA sequences with large number of mapping sRNA showed similarity to NBS-LRR resistance genes. Among the top 200 genes with the highest number of sRNA mapping, at least 121 (60%) were identified as NBS-LRR type resistance genes (Supplemental Table S6). In contrast, among those with from one to ten hits, 74 out of 6,671 were identified as NBS-LRR type resistance genes.

DISCUSSION
Studies of sRNA in plants in general, and conifers in particular have focused on the identification of miRNA sequences. In the present study, we aimed at a more general characterisation of sRNA in Picea abies. To set the data into perspective we also analysed some aspects of sRNA patterns in other plant species. The most striking result relates to the very large proportion of sRNA that are derived from NBS-LRR resistance gene sequences in P. abies, a patterns that to a lesser extent is observed only in some other plant species.

In contrast to most angiosperm species, the vast majority of sRNA were 21 nt long in our P. abies sRNA libraries, and the frequency of 24 nt sRNA was very low (estimated at 1% in P. abies). This pattern is consistent with data from several other Pinaceae conifer species: Araucaria araucana, Picea glauca, Pinus contorta, Pinus strobus, Pseudotsuga menziesii, and Thuja plicata (Dolgosheina et al. 2008) and has also been observed previously in P. abies (Yakovlev et al. 2010), suggesting that it might be general for conifers in the Pinaceae family. Concordant with these findings Dolgosheina et al. (2008) found no evidence for a DCL3
enzyme in this group of conifers, which in angiosperm is crucial for the production of 24 nt sRNA. Until a comprehensive survey of siRNA in multiple tissues is available, it cannot be ruled out that that DLC3 dependent 24 nt are present in higher abundance some tissues also in species from the Pinaceae family. The present study included actively growing needles, dormant vegetative buds and female reproductive structures at an early stage of development. It is conceivable that 24 nt siRNA are produced during later stages of reproductive development in *P. abies*, but the frequency observed in vegetative tissues are clearly lower than those in the angiosperms studied so far.

The apparent lack of DCL3 and the low frequency of 24 nt sRNA is not a feature common to all conifers as it was recently shown that *Cunninghamia lanceolata* has a sRNA distribution more similar to angiosperm species, including a large fraction of 24 nt sRNA (Wan et al. 2012). They also reported the presence of sequences predicted to code for proteins necessary for biogenesis of 24 nt sRNA, including DCL3.

An intriguing question stemming from apparent lack of a pathway producing 24 nt siRNA in Pinaceae is whether the role of 24 nt sRNA, in methylation and silencing of transposable elements, is conducted by some other sRNA species, or perhaps an alternative mechanism. As species in the Pinaceae family tend to have exceptionally large genomes compared to both angiosperms and other gymnosperm families (mean 1C in Pinaceae=23.76pg, Cypressaceae=12.43pg and Taxaceae=11.05pg, Angiosperms=5.94pg), it has been speculated that the large size may in part be explained by less efficient mechanisms for silencing of transposable elements (Dolgosheina et al. 2008; Bennett and Leitch 2010; Burleigh et al. 2012). Only 1% of the short reads obtained from *P. abies* mapped to the six available BAC sequences from *Picea*. In an attempt to compare this estimate to a species with a high abundance of 24 nt sRNA mainly associated with repetitive DNA (*Populus*; Klevebring et al. 2009), we mapped *P. trichocarpa* sRNA against scaffold 1 of the *P. trichocarpa* genome assembly (48.4 Mb). Dividing the reference in 74 windows each with a size corresponding to the sum of the six *Picea* BACs, we obtained an average of two percentage of sRNA mapping...
per window. The estimate of one percent obtained with *Picea* BACs lies in the lower 10% of the distribution obtained in Populus, indicating a somewhat, but not substantially, lower frequency of sRNA mapping to genomic DNA. As the majority of the analysed *Picea* BACs contain repetitive DNA (mainly retroelements; De Paoli 2006; Hamberger et al. 2009) the estimated fraction of sRNA mapping to these regions seems to be low. However, it is clear that more data including a reference genome is needed to test if mechanisms involving sRNA are important in silencing of retroelements in conifers. As most retroelement in conifers seems to be old (De Paoli 2006), the majority is likely silenced by mutations, thus reducing the need for mechanism silencing their transcription. However, based on comparisons of repetitive elements between bald cypress and loblolly pine, (Liu et al. 2011; Magbanua et al. 2011) it was suggested that Pinaceae species might contain more recently expanded repetitive element families, which in turn could explain part of the larger genome sizes observed in Pinaceae. Thus, a less efficient siRNA guided silencing of transposable elements in Pinaceae seems plausible.

**miRNA in *P. abies***

The identification of novel miRNA in *P. abies* and their targets was not main focus of the present study, and we can conclude that the vast majority of 21 nt sRNA in *P. abies* can be classified as siRNA, and that only a very small part of the *P. abies* 21 nt sRNA population is comprised of miRNA. The exceptionally large fraction of 21 nt siRNA complicates the identification and analysis of miRNA and their targets. Still, a number of the novel miRNA families have support by both large number of reads and the presence of both miRNA and miRNA* sequence and a potential precursor suggesting that they are true miRNAs.

We found that seven of the predicted miRNA families had predicted target genes with high sequence similarities to genes with characterized functions in other plant species, including MIR156, MIR160, MIR164, MIR166, MIR172, MIR395 and MIR858 (Rhoades et al. 2002; Takuno and Innan 2011; Khraiwesh et al. 2012). Even if our data does not allow us to conclusively identify correct targets of the identified miRNA families, we identified a
substantial number of putative miRNA families that may target resistance genes of the NBS-LRR class (Supplemental Table S1). In particular, several predicted highly abundant conifer specific 22 nt miRNA were identified that are likely to trigger siRNA production from the target loci (see below).

**Generation of siRNA from NBS-LRR genes in *Picea abies***

We found that a surprisingly high portion of sRNA in *P. abies* is derived from NBS-LRR type resistance genes. Out of the close to six million 21 nt sRNA reads included in the analysis, 1.6 million (27%) mapped to sequences related to NBS-LRR genes, and these reads constitute 48% of all reads mapping to transcribed sequences. Considering the small fraction of the transcriptome that these NBS-LRR sequences represent, these estimates are compelling. Recent studies have highlighted a new role for miRNA and siRNA in plant defence (Zhai et al. 2011; Li et al. 2012, Shivaprasad et al. 2012). 22 nt miRNA specifically targeting NBS-LRR genes trigger siRNA at these loci, indicating an important role in regulation of resistance. In the present study, we found evidence that this mechanism is a main source generating 21 nt sRNA from *P. abies* NBS-LRR genes.

Although high phase scores that indicate secondary siRNA production in a phased pattern, were rare in our data, we identified several loci with high phase score and with a 22 nt sRNA mapping in a position consistent with it acting as a trigger of the phased 21 nt sRNA. A generally low frequency of significant phase score could, even in the presence of phased degradation, be explained by the presence of many partial sequences in the reference database, the repetitive nature of LRRs, and that triggering of siRNA occurs at several positions in a gene. In the partial sequences, the region where siRNA generation is initiated might be lacking and strong phasing is only expected close to the initiation site. Studies of NBS-LRR genes in angiosperms have so far only identified NBR and TIR domains as targets for miRNA mediated initiation of siRNA (Zhai et al. 2011; Shivaprasad et al. 2012). In our data we identified many EST clusters with only LRR domains that spawned large number of siRNA. These generally had low phase scores, probably also as an effect of their structure.
with short highly similar repetitive regions, which may obscure a phasing signal. In the present study we also found indications that more than one 22 nt miRNA might initiate siRNA in different phases in one gene, which would significantly reduce the phase score statistic, despite a phased degradation (Figure 6, 7).

Available data in angiosperms have implicated one diverse miRNA superfamily (including miR482 and miR2118) as initiators of secondary siRNA on NBS-LRR mRNA (Zhai et al. 2011, Shivaprasad et al. 2012). In contrast, our data set contained a substantial fraction of highly abundant and diverse 22 nt sRNA, several of which were predicted miRNA and predicted to initiate secondary siRNA form NBS-LRR transcripts. Most of those miRNA were predicted to target TIR domains, but due to the lack of a comprehensive full-length transcript database for *P. abies*, it is premature to evaluate the target specificity of the spruce miRNA initiating secondary siRNA. Based on available data we find no support for a major role for the miR482/miR2118 superfamily as regulators of NBS-LRR genes in *P. abies*. Rather, a large number of other highly abundant miRNA seems to specifically target such genes in this species.

An alternative mechanism that could potentially generate siRNA is natural antisense transcription (NAT) resulting in double stranded RNA and subsequent siRNA production. Due to the nature of NBS-LRR gene evolution, NAT might be more prevalent among such genes (Yi and Richards 2009). Our DGE-tag data does not support NAT as a general mechanism for sRNA derived from NBS-LRR type resistance genes. Although rarely, EST clusters with high ratio of antisense to sense DGE-tags were identified (Supplemental Table S2), there was no general tendency for such genes to show high ratios.

**siRNA mediated degradation of NBS-LRR genes varies widely between species.**

Our survey of sRNA mapping to NBS-LRR genes revealed striking differences in the prevalence of this phenomenon. On the basis of available data, we can only speculate on reason for this large diversity. Clearly, *P. abies* stands out as an extreme in terms of the
amount of sRNA that is derived from NBS-LRR genes as well as the percentage of such
genes that spawn siRNA. The mechanism seems to reach a significant level only in seed
plants. This increase is correlated with a drastic expansion of the number of NBR-LRR genes
occurring in seed plants (Yue et al. 2012). However, a very low frequency of NBS-LRR genes
from monocots spawn siRNA, and a high frequency was observed only in some dicot species,
with very low frequencies observed in Arabidopsis.

The low frequency observed in monocots refutes the hypothesis that the mechanism is
directly related to the number of NBS-LRR genes. However, estimates of sequence
divergence between members of NBS-LRR genes in Arabidopsis, Oryza, Populus and Vitis
suggest that recent tandem duplication has been more frequent in the two perennial species
(Yang et al. 2008). These data also indicate increased numbers of gene conversion events and
unequal crossing over among NBS-LLR genes in the perennial species, and it was suggested
that the excess of recent duplications and a higher homologous recombination rate in Vitis and
Populus, could compensate for the long generation time coupled to a low nucleotide
substitution rate as compared to the pathogens (Yang et al. 2008). A reliable estimate of the
number of NBS-LRR genes in conifers and their pattern of evolution is premature and has to
await a conifer genome sequence, but recent expansion of the NBS-LRR gene family in some
species might explain part of the divergent patterns of siRNA degradation of NBS-LRR genes
observed between species.

Recently it was suggested that the miRNA-mediated mechanism generating siRNA from
NBS-LRR genes is part of counter-counter defence against pathogens (Shivaprasad et al.
2012). A silencing cascade mediated by miRNA482 that targets NBS-LRR genes was
suppressed in plants infected with virus or bacteria, leading to an increase in mRNA levels of
NBS-LRR target genes, likely contributing to increased defence reactions. To what extent
such a counter-counter defence mechanism are present in different species is currently
unknown, and thus to what extent it could explain the large variation in siRNA production
observed between species. An interesting question is whether the evolution of such counter-
counter defence mechanism is affected by life history, such as perenniality. A common feature of several species with high numbers of siRNA mapping to NBS-LRR genes is that they are long-lived perennial species (*Picea*, *Populus*, *Vitis*, *Gossypium*, but not *Medicago*). A major difference between annuals and perennials, affecting the evolution of resistance, is that perennials will with almost certainty encounter many different pathogens before reproduction, and the long generation time confers problems in matching the evolutionary rates of the pathogens. Besides higher duplication and recombination rates among NBS-LRR genes, additional layers of defense might be beneficial.

Recent data from legumes suggest that repression of NBS-LRR genes by specific miRNA may have evolved to avoid plant defence responses during symbiotic interactions (Bazin et al. 2012). Such a mechanism could potentially explain the high incidence of miRNA guided siRNA mediated degradation of NBS-LRR genes observed in *Medicago*. Such an explanation seems less likely for the high incidence of siRNA degradation of resistance genes seen in perennials. Even though symbiotic mycorhizal interaction are common among those species, a massive degradation of NBS-LRR genes in needles as observed in *Picea* seems costly and detrimental to pathogen resistance.

**MATERIALS and METHODS**

**Sequencing of small RNA from *Picea abies***

Tissue was collected from newly flushed buds of *P. abies* and frozen in liquid nitrogen. Total RNA was isolated using a modified version of that of Azevedo et al. (2003), where precipitation was done in Isopropanol instead of LiCl in order to retain small RNA. Small RNA libraries were constructed using the Illumina Small RNA kit version A, where PAGE gels were used to select small RNA with a size between 18 and 30 nucleotides. Each library was based on a pool of RNA from six trees. Two libraries from separate pools using the Illumina platform.
After removal of adaptor sequence, about 10.6 million reads were obtained. These reads were filtered by removing those matching structural RNA (rRNA, tRNA) using blastn searches against Rfam, and mono-nucleotide repeats. After these filtering steps, 8,310,409 reads, were kept for further analysis.

**DGE-tag data**

To estimate gene expression we used six digital gene expression libraries constructed from individual needle tissue samples of *P. abies* seedlings. The libraries were sequenced on the Illumina platform each to a depth of around 4 million reads. After filtering the reads were mapped against the *Picea* “EST clusters” from the gene index project. Expression levels were estimated by the number of reads mapping to each transcript divided by the length of each transcript.

**Identification of miRNA and their targets**

Short reads were searched against the Rfam (Gardner et al. 2009) database and matches to other type of known sRNA were removed. The remaining short reads were compared to the miRBase using NCBI “blast-short” program with an e-value = 1000 and word size = 7. Sequences with no more than 2 mismatches (including overhangs but no gaps allowed) were identified as miRNA conserved between species. Precursors were identified in three steps: First, we searched a spruce transcriptome database that included the *P. abies* PUT (Putative Unique Transcripts) assembly of Chen et al. (2012), the *P. glauca* gene catalogue of Rigault et al. (2011) and the SGI EST clusters (Supplemental Table S7) with the filtered short reads. Sequences with a maximum of two mismatches to a short read (no gaps allowed), and to which the short read mapped less than six times were used predict primary precursors using the miRDeep pipeline (Friedländer et al. 2008). The predicted primary precursors were then filtered to remove protein-coding mRNAs by searching against UniProtKB plant protein database (The UniProt Consortium 2011). In addition, to minimize the number of false positives in the prediction, we used a classification method *PlantMiRNA*Pred (Xuan et al. 2011) for further classification of valid plant pre-miRNA sequences. We used the SGI dataset for prediction of miRNA targets using psRNA Target (Dai and Zhao 2011).
Analysis of siRNA

siRNA data for the various species was mainly downloaded as GEO data sets, series GSE28755 (see Supplemental Table 7 for details). Small RNA was mapped against various reference sequences using bowtie (version 0.12.7; Langmead 2010). In order to facilitate comparisons among the different species we used EST clusters from the Plant Gene Index (Supplemental Table S7) as most species lack a reference genome. For *Arabidopsis thaliana*, *Oryza sativa*, *Populus trichocarpa*, and *Vitis vinifera* we also mapped siRNA against annotated full-length transcripts (Supplemental Table S7). For *P. abies* we also used the Arborea *P. glauca* gene catalogue that contains more than 23 thousand mRNA sequences that are annotated as full length (Rigault et al. 2011). In order to predict protein domains and identify putative resistance genes we translated all EST clusters to protein sequences (all six frames) and used PFAM protein families (Punta et al. 2012) and the program hmmsearch with default settings (http://www.hmmer.org). The following hmm profiles were used; LRR_1: PF00560, LRR_2: PF07723, LRR_3: PF07725, LRR_4: PF12799, LRR_5: PF13306, LRR_6: PF13516, LRR_7: PF13504, LRR_8: PF13855, NBS: PF00931, TIR: PF01582.

In addition *P. abies* siRNA data was mapped against six Picea BACs (four *P. abies* ones supplied by E. De Paoli, and two from *P. glauca*, Hamberger et al. 2009). Finally, *P. contorta* siRNA (ftp://ftp03.bcgsc.ca/public/pine_rice_smRNA/pine_all_454_reads.fna) was also mapped to 10 *Pinus taeda* BACs (Kovach et al. 2010).

Analysis of sRNA phasing

Phasing analysis was conducted based on the calculation described in De Paoli et al. (2009), with some modification, essentially as in Zhai et al. (2011). Reads on the sense and antisense strand, were offset by -1 nt and +1 nt, respectively, to account for a two nt overhang occurring during dsRNA cleavage. The number of phase cycle positions occupied by at least one read in the current 210 nt window must be greater than 3, and the most abundant read at those
positions should not constitute more than 90% of sum of all reads at phase cycle positions. The calculations were performed in R (R Development Core Team 2008).

DATA ACCESS
All sequence data generated in the present study is available from Genbank SRA at accession number: SRR813801, SRR815091, SRR824098-SRR824103, SRR824149, SRR824150.

ACKNOWLEDGEMENTS
We want to thank Kerstin Jeppson for small RNA extractions and Illumina library preparations and Uppsala SciLifeLab for generating the sequence data. We are also in depth to Emanuele De Paoli and Jixian Zhai for answering questions regarding calculation of phase scores from illumina data.
LITERATURE CITED


FIGURE LEGENDS

Figure 1. Observed length distribution of sRNA sequences from *P. abies*. The total number of reads in each size class is plotted.

Figure 2. Percentage of each 5’ terminal nucleotide of *P. abies* sRNA sequences. Estimates were calculated separately for sRNA of different length (20-24).

Figure 3. Distribution of number of sRNA reads mapping to individual EST clusters. The x-axis shows log2 transformed length normalised read counts, where the number of sRNA reads mapping to each EST cluster are normalised by dividing by the length of the cluster and multiplying by average read length (21). Mean and median of non-transformed data for the two classes of genes are 4.4 and 0.07 for non NBS-LRR genes, and 241 and 10.0 for NBS-LRR genes.

Figure 4. Coverage plot of sRNA reads mapping to TC142020. Allowing for one mismatch and multiple hits resulted in 3547 unique sRNA sequences mapping to this EST cluster. Each sRNA sequence is represented by one read. The top panel show a coverage plot of the complete sequence, and the blue shaded area is shown in the bottom panel displaying individual reads. Read orientation is depicted by light or dark blue colour in the bottom panel. The plot was produced with Tablet (Milne et al. 2010).

Figure 5. Coverage plot of sRNA reads mapped to a full-length spruce TIR-NBS-LRR gene (ACN40032). Normalised read count is the number of reads mapping divided by the number of matches for each read in the whole reference.

Figure 6. Phasing pattern observed after mapping *P. abies* 21 nt sRNA to TC171280. A, sRNA abundance plot (reads mapping in forward and reverse orientation are plotted above and below the horizontal line, respectively). B, phasing score distribution is plotted above the horizontal line, and predicted cleavage positions for 22 nt miRNA are shown in red below. C,
a magnification around the predicted miRNA cleavage sites, with the predicted 22 nt miRNA in read, and two phased 21 nt sRNA in black.

Figure 7. Phasing pattern observed after mapping *P. abies* 21 nt sRNA to TC170604. A, sRNA abundance plot (reads mapping in forward and reverse orientation are plotted above and below the horizontal line, respectively). B, phasing score distribution is plotted above the horizontal line, and predicted cleavage positions for 22 nt miRNA are shown in red below. C, a magnification around the predicted miRNA cleavage sites, with the predicted 22 nt miRNA in read, and two phased 21 nt sRNA in black.

Figure 8. Nucleotide sequence alignment of TAS3 loci from Arabidopsis and *Nicotiana tabacum*, and a putative TAS3 locus from spruce (TC161203). Boxes indicate miR390 complementary sites and predicted ta-siARF-encoding sequence regions. Shading shows residues identical for the given position. GeneBank accession numbers of the sequences included in the alignment are as follows: *Arabidopsis thaliana*, BX838290; *Nicotiana tabacum*, FJ804751.

Figure 9. Percentage of NBS-LRR type genes hit by more than 10 unique sRNA reads in different plant species. For each species, genes that contained any of TIR, NBS, LRR or BED domains, were selected, and for each class, the percentage of those genes that were hit by more than 10 different unique 21 nt sRNA sequences was calculated. In the top graph (A), sRNA reads were mapped allowing one mismatch, and reads mapping at multiple positions in the whole reference were retained. In the bottom graph (B), mapping was performed with only perfect matches and discarding reads mapping to multiple positions.
### Table 1. Ten most abundant 22 nt sRNA in P.abies.

<table>
<thead>
<tr>
<th>sRNA</th>
<th>Reads</th>
<th>Target 1</th>
<th>Phmax†</th>
<th>Target 2</th>
<th>Phmax†</th>
<th>NBS-LRR‡‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>2*</td>
<td>148890</td>
<td>TC161536</td>
<td>7.3</td>
<td>TC130897</td>
<td>6.1</td>
<td>3</td>
</tr>
<tr>
<td>3*</td>
<td>133356</td>
<td>TC139913</td>
<td>25.1</td>
<td>CO219226</td>
<td>19.4</td>
<td>14</td>
</tr>
<tr>
<td>5*</td>
<td>62691</td>
<td>TC143748</td>
<td>23.8</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>55848</td>
<td>TC143748</td>
<td>23.8</td>
<td>CO234506</td>
<td>15.3</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>53201</td>
<td>TC143748</td>
<td>23.8</td>
<td>CO234506</td>
<td>15.3</td>
<td>7</td>
</tr>
<tr>
<td>13*</td>
<td>41914</td>
<td>DR540978</td>
<td>45.7</td>
<td>TC143748</td>
<td>23.8</td>
<td>4</td>
</tr>
<tr>
<td>14*</td>
<td>40208</td>
<td>TC128174</td>
<td>35.4</td>
<td>TC167323</td>
<td>21.2</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>25571</td>
<td>TC137450</td>
<td>6.6</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>35*</td>
<td>20410</td>
<td>TC160176</td>
<td>13.7</td>
<td>TC158558</td>
<td>4.9</td>
<td>3</td>
</tr>
<tr>
<td>41*</td>
<td>17814</td>
<td>TC143748</td>
<td>23.8</td>
<td>TC161536</td>
<td>7.3</td>
<td>9</td>
</tr>
</tbody>
</table>

† Max phase score observed for Target 1 and 2, respectively.

‡‡ Number of EST-clusters with similarity to NBS-LRR genes identified as targets.

* predicted as miRNA
Table 2. Ratio of 21 nt sRNA to 24 nt sRNA for selected gene classes in Populus. The ratio was calculated for those annotation classes containing five or more genes among those 500 with the highest sRNA counts.

<table>
<thead>
<tr>
<th>Annotation</th>
<th>No. of genes</th>
<th>21nt/24nt</th>
<th>Total reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBS-LRR resistance protein</td>
<td>512</td>
<td>15.8</td>
<td>183,882</td>
</tr>
<tr>
<td>Unknown function</td>
<td>472</td>
<td>1.9</td>
<td>74,508</td>
</tr>
<tr>
<td>Reverse transcriptase</td>
<td>64</td>
<td>0.2</td>
<td>15,173</td>
</tr>
<tr>
<td>Polyprotein</td>
<td>21</td>
<td>0.5</td>
<td>10,647</td>
</tr>
<tr>
<td>ORF V protein</td>
<td>5</td>
<td>0.4</td>
<td>7,429</td>
</tr>
<tr>
<td>Retroelement</td>
<td>14</td>
<td>0.4</td>
<td>6,796</td>
</tr>
<tr>
<td>GRAS family transcription factor</td>
<td>22</td>
<td>16.6</td>
<td>2,987</td>
</tr>
<tr>
<td>Retrotransposon</td>
<td>11</td>
<td>0.3</td>
<td>1,590</td>
</tr>
<tr>
<td>Xylem serine proteinase 1</td>
<td>5</td>
<td>12.9</td>
<td>1,006</td>
</tr>
</tbody>
</table>
Figure 1. Observed length distribution of sRNA sequences from *P. abies*. The total number of reads in each size class is plotted.
Figure 2. Percentage of each 5’ terminal nucleotide of P. abies sRNA sequences. Estimates were calculated separately for sRNA of different length (20-24).
Figure 3. Distribution of number of sRNA reads mapping to individual EST clusters. The x-axis shows log2 transformed length normalised read counts, where the number of sRNA reads mapping to each EST cluster are normalised by dividing by the length of the cluster and multiplying by average read length (21). Mean and median of non-transformed data for the two classes of genes are 4.4 and 0.07 for non-NBS-LRR genes (A), and 241 and 10.0 for NBS-LRR genes (B).
Figure 4. Coverage plot of sRNA reads mapping to TC142020. Allowing for one mismatch and multiple hits resulted in 3547 unique sRNA sequences mapping to this EST cluster. Each sRNA sequence is represented by one read. The top panel show a coverage plot of the complete sequence, and the blue shaded area is shown in the bottom panel displaying individual reads. Read orientation is depicted by light or dark blue colour in the bottom panel. The plot was produced with Tablet (Milne et al. 2010).
Figure 5. Coverage plot of sRNA reads mapped to a full-length spruce TIR-NBS-LRR gene (ACN40032). Normalised read count is the number of reads mapping divided by the number of matches for each read in the whole reference.
Figure 6. Phasing pattern observed after mapping P. abies 21 nt sRNA to TC171280. A, sRNA abundance plot (reads mapping in forward and reverse orientation are plotted above and below the horizontal line, respectively). B, phasing score distribution is plotted above the horizontal line, and predicted cleavage positions for 22 nt miRNA are shown in red below. C, a magnification around the predicted miRNA cleavage sites, with the predicted 22 nt miRNA in red, and two phased 21 nt sRNA in black.
Figure 7. Phasing pattern observed after mapping P. abies 21 nt sRNA to TC170604. A, sRNA abundance plot (reads mapping in forward and reverse orientation are plotted above and below the horizontal line, respectively). B, phasing score distribution is plotted above the horizontal line, and predicted cleavage positions for 22 nt miRNA are shown in red below. C, a magnification around the predicted miRNA cleavage sites, with the predicted 22 nt miRNA in read, and two phased 21 nt SRNA in black.
Figure 8. Nucleotide sequence alignment of TAS3 loci from Arabidopsis and *Nicotiana tabacum*, and a putative TAS3 locus from spruce (TC161203). Boxes indicate miR390 complementary sites and predicted ta-siARF-encoding sequence regions. Shading shows residues identical for the given position. GeneBank accession numbers of the sequences included in the alignment are as follows: Arabidopsis thaliana, BX838290; Nicotiana tabacum, FJ804751.
Figure 9. Percentage of NBS-LRR type genes hit by more than 10 unique sRNA reads in different plant species. For each species, genes that contained any of TIR, NBS, LRR or BED domains, were selected, and for each class, the percentage of those genes that were hit by more than 10 different unique 21 nt sRNA sequences was calculated. In the top graph (A), sRNA reads were mapped allowing one mismatch, and reads mapping at multiple positions in the whole reference were retained. In the bottom graph (B), mapping was performed with only perfect matches and discarding reads mapping to multiple positions.