Evolutionary Conservation of Long Intergenic Non-coding RNA Genes in Arabidopsis

by

Alexander John Hammel

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Botany)

The University of British Columbia
(Vancouver)

August 2013

Abstract

Long intergenic non-coding RNA (lincRNA) genes are a poorly studied class of transcripts, particularly in plants. Because of the low levels of expression, high tissue specificity, and rapid rate of evolution of lincRNA transcripts, the discovery and functional annotation of these molecules is a significant challenge. Here, I report the annotation of 201 new lincRNA transcripts in *Arabidopsis thaliana* discovered using the results of a single RNA-seq experiment of a normalized library. Using these sequences, along with the 6 480 lincRNA genes annotated by Liu et al. (2012), I performed a pairwise sequence alignment experiment with the genomes of 22 plant species in order to discover highly conserved sequences within lincRNA loci. Of the 6 681 lincRNA sequences examined, 3 374 have highly conserved sequences supported by multiple genomic alignments to other species. Six of these show evidence of ongoing reduced sequence rate evolution when single-nucleotide variant data from the recent evolutionary history of *Arabidopsis thaliana*. The rate of retention of these conserved regions within the Brassicaceae suggests a much higher rate of sequence turnover in lincRNA genes compared with protein coding genes. Structural variant data from 80 different *A. thaliana* ecotypes suggests that lincRNA genes suffer deletions of the entire locus from the genome with appreciable frequency: 570 of the lincRNA loci examined are entirely missing from at least one *A. thaliana* strain. These results suggest an intriguing mixture of rapid sequence evolution with short, highly-conserved islands in lincRNA genes.

Preface

This project includes collaborations with David Tack, a Ph.D. candidate in the Adams lab at UBC, and Jon Willinofsky, an undergraduate at UBC. The strategy for filtering Illumina RNA-seq reads associated with annotated genes—described in section 2.1—was developed in collaboration with David Tack and Jon Willinofsky. The software implementation of this strategy was developed by Mr. Willinofsky under the supervision of Mr. Tack and Dr. Adams. I was responsible for all further data analysis once the filtered data were obtained.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank Dr. Keith Adams for his support in this project. I would also like to acknowledge the rest of the members of the Adams lab, particularly Aude Darracq and David Tack, for their support and technical assistance. Drs. Quentin Cronk, Naomi Fast, and Loren Rieseberg made invaluable contributions of their time, ideas, and expertise. I feel it necessary to thank the open source software community as a whole, without whom this project would have been impossible.

Chapter 1

Introduction

It was once possible for biologists to think of RNA mostly as a transitional stage between DNA and protein. With modern sequencing technology, it is becoming clear that many transcripts do not undergo translation, but are functional as RNA. The eurkaryotic genome produces an amazingly broad spectrum of transcript types with a great diversity of functions related to gene expression, including transcription, translation, and chromatin remodelling (Ponting et al., 2009). Although recent advances in nucleic acid sequencing technology have revealed a great number of non-protein-coding RNA transcripts (ncRNAs), relatively few of these have been functionally characterized.

Detailed studies of the transcription of eurkaryotic genomes using modern sequencing technology have consistently found that transcription is surprisingly ubiquitous (Kapranov et al., 2007). In humans, it has been estimated that almost every nucleotide in the genome is transcribed in at least one developmental stage or cell type (ENCODE Project Consortium, 2007), and a recent study suggests that ncRNA transcripts outnumber protein coding transcripts at least two-to-one in both diversity of species and total abundance (Managadze et al., 2013). Although a number of ncRNA transcripts have been functionally characterized, it is difficult to believe that all of these transcripts are functionally important. This suggests the need for a method to determine which non-coding transcripts may be functional, and which are non-functional 'transcriptional noise'.

Long intergenic non-coding RNAs are transcripts longer than 200 nucleotides which do not show experimental or evolutionary evidence for translation (such as codon substitution biases consistent with protein-coding sequences), and which do not overlap with known genic regions. This definition excludes other species of non-coding RNA, such as anti-sense, intronic, and bi-directional transcripts, as well as comparatively well-understood small RNA species such as miRNA, siRNA and snoRNA (Zhang et al., 2007;

Molnar et al., 2011; Scott and Ono, 2011). In practise, a distinction is usually drawn between lincRNA and transcripts associated with transposable elements and other intergenic repeats, as these transcripts are thought to be functionally distinct from other non-coding RNAs (Zhang et al., 2010). Liu et al. (2012) refer to these RNA species as 'repeat-containing transcriptional units' (RCTUs). RNA genes which are involved in transcription, including rRNA and tRNA genes are also normally excluded from the category.

1.1 Evolution of non-coding RNA

Despite recent interest in the function of non-coding RNAs, there have been few studies of the degree to which lincRNA genes are conserved amongst species. In mice, a sequence comparison across 29 mammal species has shown that lincRNA loci are evolutionarily conserved at a rate higher than random intergenic non-coding sequence, but lower than protein-coding genes (Guttman et al., 2010).

Pang et al. (2006) found that long non-coding RNAs are poorly conserved between humans and mice compared to miRNA and snoRNA genes. This holds true for lincRNA genes that are known to be functional as long RNAs, including *Xist* (see below). The authors suggest that this may be because only short stretches of sequence may be essential for function, with the rest of the sequence functioning in secondary structure or as spacer.

Sequence conservation based analyses have been shown to be effective in identifying functional ncRNAs. For example, Willingham et al. (2005) were able to identify a ncRNA with a repressor function by screening a pool of mouse ncRNAs showing sequence similarity to human genomic sequences. NRON, the ncRNA gene which was discovered to repress the expression of the NFAT protein coding gene, is a relatively large transcript at 3.7 kb. It contains two regions of relatively high similarity between rodents and primates: one 289 base pair region with 90% identity between humans and mice, and one 400 base pair region with 89% percent identity. In creating a pool of candidates, the authors used a relatively low cutoff of 50% identity across 70% of the length of the locus (Numata et al., 2003; Willingham et al., 2005).

1.2 Known functions of lincRNA genes

Relatively few lincRNA genes have been functionally characterized in plants. One such gene is *INDUCED BY PHOSPHATE STARVATION 1* (*IPS1*) which is involved in the microRNA-mediated regulation of phosphorus nutrition (Kim and Sung, 2012). Over-

expression of *IPS1* results in a decrease in phosphate accumulation in the shoot, which suggests that it is involved in the mobilization of phosphates in conditions of P_i starvation (Franco-Zorrilla et al., 2007). The mechanism of action of *IPS1* has been determined. Franco-Zorrilla et al. (2007) found that the transcript modifies the activity of the microRNA miRNA399 by competitive inhibition. The IPS1 transcript binds to miRNA399 but does not undergo cleavage. This decreases the proportion of miRNA transcripts which can bind to their protein-coding targets and trigger the transcript degradation pathway.

COLDAIR is another well-characterized plant lincRNA gene. It is required for repression of *FLOWERING LOCUS C* during vernalization. Knockdown mutants for COLDAIR have increased expression of FLC2 following cold treatment and do not display cold-triggered flowering behaviour (Heo and Sung, 2011). COLDAIR is known to act by physically binding the Polycomb Repressive Complex 2, resulting in the formation of repressive chromatin at the *FLC* locus (Heo and Sung, 2011). Remarkably, the mammalian lincRNA gene HOTAIR—after which COLDAIR is named—also binds the PRC2 complex, altering chromatin state and gene expression patterns (Gupta et al., 2010). The two loci are not apparently orthologous (Heo and Sung, 2011), which suggests that PRC complex binding is something of a recurring theme in lincRNA function. This theme is evident in the mammalian lincRNA gene *Xist*, which acts in X chromosome inactivation. *Xist* acts as a molecular bridge, binding both PRC2 and the YY1 protein, which itself binds to DNA motifs along the X chromosome (Jeon and Lee, 2011). The result is that the PRC is brought into proximity with the X chromosome, causing the chromatin modification which inactivates the entire chromosome (Zhao et al., 2008; Jeon and Lee, 2011).

Although lincRNA genes are relatively poorly studied, those that have been functionally characterized tend to be involved in epigenetic gene regulation pathways such as chromatin modification and the miRNA pathway, rather than the "classical" transcription factor pathways or catalytic functions. The mechanism of action of the characterized functions tends to DNA-RNA and DNA-protein binding interactions. Because the known functions of lincRNA genes depend on binding motifs along the length of the transcript, I hypothesize that functional lincRNA genes may be characterized by conserved binding regions. Such conserved regions have been observed in *IPS1* (Franco-Zorrilla et al., 2007; Hou et al., 2005; Liu et al., 1997; Burleigh and Harrison, 1997) in plants, and in many non-coding RNAs in animals (Pang et al., 2006). Guttman et al. (2010) found evidence in mice for short regions of reduced sequence rate evolution in many lincRNA loci, finding that the average conserved region covers approximately 22% of the locus, compared

with a figure of 70% in protein-coding exons. My work is the first systematic attempt to identify conserved regions in plant lincRNA genes on a genome-wide scale.

1.3 Objectives

The goals of this study were to determine the patterns of evolution of lincRNA genes, to identify regions of high evolutionary conservation in these genes, to discover whether there is evidence for ongoing reduced sequence rate evolution in these conserved regions, and to explore the potential of RNA-seq with library normalization for the discovery of lincRNA transcripts. I adopted a sequence-similarity approach to identify putative homologs of *Arabidopsis thaliana* lincRNA genes in a wide variety of plant species. Using the same approach to discover areas of sequence similarity in protein-coding genes allows us to compare the patterns of sequence conservation. Using alignments across many species has allowed me to identify conserved regions of lincRNA genes which have experienced relatively little sequence evolution over the course of macroevolutionary time. By integrating data regarding these conserved regions with data regarding mutations which have arisen in the recent evolutionary history of *Arabidopsis thaliana*, I have been able to test the hypothesis that conserved regions are subject to decreased rates of sequence evolution on both a microevolutionary and macroevolutionary time scale.

Chapter 2

Methods

2.1 Identification of lincRNA transcripts

Novel lincRNA transcripts were identified using RNA-seq. The raw Illumina reads were obtained from a prior study by (Marquez et al., 2012). This data set was chosen because the library was enriched for rare transcripts. Marquez et al. (2012) constructed their cDNA libraries from *Arabidopsis thaliana* (Col-0) flowers and seedlings, and pooled the cDNA from both tissue types. The cDNA pool was normalized using an Evrogen Trimmer-direct Kit, and sequenced with 75-base-pair paired-end reads on five lanes using the Illumina GA system. I mapped the Marquez et al. reads to the TAIR10 version of the *Arabidopsis* genome assembly (Lamesch et al., 2012) using Bowtie (Langmead et al., 2009).

Together with undergraduate Jon Willinofsky, I developed a strategy to filter out reads associated with annotated genes. Our strategy was to remove all of the read pairs which overlap with an genic region annotated in the TAIR10 genome (including 5′ and 3′ UTRs), then remove all of the read pairs which overlap with those read pairs, and so forth until only reads which unambiguously map to intergenic regions remain. We considered reads to be overlapping if they shared at least one base pair of their mapped positions in common, and non-overlapping if there was no shared base pair. Overlap between reads and annotated regions was defined in the same way. We expect this strategy to remove not only reads associated with annotated protein coding genes, but also those associated with unannotated 5′ and 3′ UTR regions and natural antisense transcripts which overlap with annotated genes.

I then identified putative lincRNA loci using the Samtools pileup function (Li et al., 2009). Every RNA molecule identified by the pileup that is at least 200 base pairs long was treated as a putative lincRNA transcript. The genomic positions, consensus

sequences, and average read alignment qualities for the putative lincRNA transcripts were calculated from the pileup.

In order to remove unannotated protein-coding loci, I used GenScan (Burge and Karlin, 1997) to predict open reading frames. I removed all loci which have an open reading frame greater than 100 base pairs. I also removed any loci that overlap with transposable elements annotated by TAIR10 (Lamesch et al., 2012). This is the same strategy that was employed by Liu et al. (2012) to filter their lincRNA annotations for protein-coding loci and transposable elements. Any loci which overlapped with repeat-containing transcriptional units annotated by Liu et al. (2012), and any loci which were covered to an average depth of less than five reads were also removed from further analysis.

In order to test the specificity of my lincRNA identification procedure, the remaining putative lincRNA transcripts were compared to the set of lincRNA annotations recently published by Liu et al. (2012). I divided my set of lincRNA loci into two categories: 'overlapping loci' whose genomic positions overlap with a locus in the Liu et al. data set, and 'novel loci' which are non-overlapping with Liu et al. loci. I examined the average fold coverage (the average number of reads covering any base along the length of the locus) and the average *Phred* mapping quality score (Ewing et al., 1998) at each locus in order to determine whether these measures are significantly different in novel and confirmed lincRNA loci.

Both the Liu et al. lincRNA loci and the novel lincRNA loci discovered by my analysis were included in downstream analyses.

2.2 Identification of lincRNA conserved regions

Conserved sequence elements of lincRNA loci were identified by alignment to the genomic sequences of selected plant species (see table 1). Species were chosen on the basis of the quality of the genome available, the depth of coverage of the genome, and to give a broad phylogenetic coverage of the plant kingdom. *Carica papaya*, for instance, was excluded as the genome is sequenced to an average depth of only 3× coverage (Ming et al., 2008).

Alignments were performed using the discontiguous MegaBLAST program (Zhang et al., 2000; Ma et al., 2002) using seed optimized for non-coding sequences with a word size of 11 and a template length of 16. I developed two criteria by which to identify regions of high sequence similarity from MegaBLAST alignments. I call a sequence a 'conserved block' when it can be aligned to the genome of another plant species with an *e*-value of less than 10^{-30}. To this set of conserved blocks, I added any alignments which

were at least 85 % identical across at least 50 base pairs. These are referred to as 'short highly conserved blocks' (SHCBs).

A perennial problem in studies of evolutionary conservation is the decision as to what constitutes a significant alignment. Because the degree of sequence similarity which is sufficient to infer homology is dependent on the organism being studied, the class of molecule, and the researchers' goals, the criteria for calling a 'significant' alignment are necessarily somewhat *ad hoc*. In almost all cases, the false-positive and false-negative rates of a particular alignment criterion are completely unknown. In order to ameliorate this difficulty, I elected to use two different approaches to constructing an alignment cutoff criterion. The SHCB criterion requires a high level of sequence similarity over a relatively small proportion of the lincRNA locus. This criterion was chosen based on what is known about the evolutionary dynamics of known functional lincRNA genes in mammals: many functionally characterized lincRNA genes have been found to have a relatively small conserved region while the rest of the length of the locus is subject to a high rate of sequence evolution (Willingham et al., 2005; Pang et al., 2006; Ponting et al., 2009). The 'conserved block' criterion uses a more general e-value cutoff. This is expected to allow alignments that do not follow the expected patterns of lincRNA evolution at the expense of failing to filter out a larger proportion of spurious alignments. For example: the e-value criterion potentially allows for conserved blocks which are smaller than expected by the low false-positive criterion, or for lincRNAs which have a modest rate of sequence conservation across the entire locus—as is not uncommon in protein-coding genes.

In order to use the presence or absence of a sequence alignment to draw conclusions about evolutionary rates, it is necessary to make comparisons among gene classes using the same alignment criteria. In order to compare the rates of conservation of lincRNA loci to protein coding genes, I took a random sample of 2 000 protein coding sequences annotated by TAIR10 (Lamesch et al., 2012) and aligned them to the plant genomes described above using the same discontiguous MegaBLAST strategy described for lincRNA loci. The resulting alignments were then filtered using the same alignment criteria described above.

2.3 Conserved regions of lincRNA loci

Putative conserved regions in other plant species, for the purposes of this study, are defined as the longest sequence of base pairs of a lincRNA locus which overlaps with all of the alignments in other species at a particular conservation criterion. (Note that this is different from a conserved *block*, which is a region of high sequence similarity between

a lincRNA locus and one particular genomic sequence in another species.) If I found no region which was overlapped by all of either alignments which met the cutoff—or if there was only one alignment to a particular locus—that locus is considered not to have a conserved region. Conserved regions were calculated separately using conserved blocks and SHCBs. These two data sets were combined for downstream analysis. If a single lincRNA locus had a conserved region annotated using both cutoff criteria, the overlap was used in downstream analysis. If the two conserved regions did not overlap, the locus was discarded.

In order to test the hypothesis that the putative conserved regions represent areas which are under stronger purifying selection than the rest of the locus, I used genomic single nucleotide variant (SNV) data from the *Arabidopsis* 80 Genomes project (Cao et al., 2011). Because a SNV present in two different *Arabidopsis* strains may represent a single mutation in their common ancestor, I considered only unique variants. A SNV of the same substitution base present at the same genomic location in more than one strain is only counted as a single unique variant in my experiment. I tested the hypothesis that putative conserved regions contain fewer unique variants than the rest of the length of the lincRNA locus using a one-tailed binomial test and calculated Q-values using false discovery rate correction (Benjamini and Hochberg, 1995).

In order to examine alignments of conserved regions, I used Clustal Omega (Sievers et al., 2011) to realign lincRNA sequences to genome regions aligned by MegaBLAST. The resulting alignments were visualized with MView (Brown et al., 1998).

2.4 Recent gain and loss of lincRNA loci

In order to study the evolutionary dynamics of lincRNA loci within the *Arabidopsis thaliana* species, I used the large deletion annotations from Cao et al. (2011). This data set lists relatively long (more than 10 nucleotide) stretches of genomic DNA which are present in the *Arabidopsis thaliana* Columbia-0 ecotype, but absent in at least one other strain, not counting deletions which are part of a more complex rearrangement. Cao et al. (2011) employed a very similar strategy to study the recent evolution of microRNA genes in *Arabidopsis thaliana*.

In order to determine whether these insertion/deletion events are more likely to represent recent insertions or recent deletions, I determined whether the locus had a significant alignment to any other plant species. Loci which are absent in some *Arabidopsis thaliana* strains but which are similar to genomic sequences in other plants likely represent recent deletion events, while sequences which are present in *Arabidopsis* Col-0 but absent

in other ecotypes and all other plant genomes likely represent recent insertion events. Finally, I determined whether recently deleted lincRNA loci are less likely to contain conserved regions (as defined in section 2.3).

Chapter 3

Results

3.1 Discovery of lincRNA loci through Illumina analysis of normalized libraries

Like other classes of non-coding RNA, lincRNAs are characterized by low levels of expression and, at least in animals, a high level of tissue specificity (Cabili et al., 2011; Young et al., 2012; Liu et al., 2012). Because clone library-based methods are somewhat unreliable for the discover of low copy number transcripts, specialized techniques are required to identify lincRNA genes on a genome-wide scale. Previously, scientists have catalogued lincRNA genes using tiling arrays (Cawley et al., 2004; Matsui et al., 2008), RNA-seq (Guttman et al., 2010), chromatin signature (Guttman et al., 2009), and conserved predicted secondary RNA structure motifs (Hupalo and Kern, 2013). Although the identification of lincRNA genes in plants is in its infancy, both tiling array and RNA-seq based methods have proven effective (Liu et al., 2012). In addition to the set of lincRNA genes annotated by Liu et al. (2012) using a combination of tiling array and RNA-seq techniques, I used a normalized RNA-seq library (Marquez et al., 2012) in order to annotate lincRNA genes *de novo*. Because normalized libraries are enriched for rare transcripts, library normalization for RNA-seq is promising as a low-cost method for the discovery of novel lincRNA genes (although it makes evaluation of expression level impossible). Library normalization combined with high throughput sequencing has previously been shown to be an effective strategy for the discovery of non-coding transcripts and other rare RNA species (Guffanti et al., 2009; Marquez et al., 2012).

The Marquez et al. (2012) data set consisted of 115 883 414 paired-end Illumina RNA-seq reads. Of these, 50 801 105 (43.84 %) mapped concordantly to the *Arabidopsis thaliana* genome using Bowtie2 (Langmead and Salzberg, 2012). After removing reads asso-

ciated with annotated genes, there remained 268 936 intergenic reads (0.5 % of the total mapped reads). I assembled these reads into 1 220 lincRNA loci. The lincRNA loci which mapped to the mitochondrion were discarded, leaving 1 142. Of these, 133 overlapped with one of the 6 728 lincRNA loci annotated by Liu et al. (2012), 82 overlapped with TAIR10 transposable elements, 229 had predicted open reading frames longer than 100 base pairs, and 94 overlapped with RCTUs discovered by Liu et al. These were all removed from further analysis. I also discarded 828 loci which had an average fold coverage of less than five (meaning that each base pair of the locus was covered by fewer than five reads, on average), leaving 201 novel lincRNA loci.

I performed a number of tests to evaluate my confidence in the authenticity of the novel lincRNA loci. These are summarized in figure 1. There is no significant difference between the average *Phred* scores or fold coverage of the lincRNA loci which were discovered *de novo* by my analysis and those which were confirmed by Liu et al. (2012), but the newly discovered lincRNA genes were significantly shorter. The lengths of the lincRNA loci—both those annotated by Liu et al. (2012) and those discovered *de novo* by my analysis–vary greatly, from the minimum length of 200 base pairs up to more than 2100. The distribution in size is roughly exponential: longer lincRNA loci are relatively rare compared to loci around 200 nucleotides.

3.2 LincRNA sequence similarity in other plant genomes

I identified genomic loci in other plant genomes with a high level of sequence similarity to lincRNA genes in *Arabidopsis thaliana* using a pair-wise local alignment method. Alignments with an *e*-value of less than 10^{-30} were considered conserved blocks. In order to identify additional, short regions of high sequence conservation, I examined any further alignments which were at least 85 % identical to a lincRNA gene across at least 50 base pairs. Those are referred to as short highly conserved blocks (SHCBs). These identification criteria were chosen after examination of the distributions of these statistics over the alignments (figures 2 and 3 and table 2) as well as visual examination of the alignments, with the goal of finding regions of high sequence similarity, erring on the side of specificity rather than sensitivity in order to minimize false positives.

Because preliminary analysis suggested that there are very few highly conserved blocks of primary sequence between *Arabidopsis thaliana* and its distant relatives, I chose to focus on finding conserved loci in the four available Brassicaceae species with available

genomes, aside from *Arabidopsis thaliana*: *A. lyrata*, *Capsella rubella*, *Brassica rapa* and *Eutrema parvulum*. Sequence alignments were performed using MegaBLAST. Within the Brassicaceae, I found 34 730 conserved blocks and an additional 8 279 SHCBs. As expected, the SHCBs were identical to their targets at a larger proportion of sites, but covered a slightly lower percentage of the lincRNA gene (table 3). Together with the fact that SHCBs tend to be found in longer lincRNA loci than conserved blocks annotated using the *e*-value criterion, this suggests that the strategy of looking for regions of high similarity at a fixed minimum length is more effective than using the *e*-value of the alignment at annotating short regions of high conservation.

Roughly 90 % of the *Arabidopsis thaliana* lincRNA loci examined have a conserved block in *Arabidopsis lyrata*. This value falls to roughly 50 % in *Capsella rubella*, the next closest relative of *A. thaliana* included in this study, and then to roughly 30 % in both *Brassica rapa* and *Eutrema parvulum* (table 4).

The analysis was expanded to include the genomes of 18 other plant species (table 1) in order to identify regions of deep conservation and compare the rates of sequence evolution between lincRNA genes and protein coding genes. When all species are considered, there were a total of 458 628 pair-wise genomic alignments. Table 2 summarizes the *e*-values of these alignments, and the percent-identity and percent-coverage statistics are summarized in figure 2. The *e*-values of the alignments follow a roughly exponential distribution, with about half of the total alignments having a value greater than 0.005. Most alignments cover a relatively small proportion of the locus (10–30%) of the length of the lincRNA gene), and are identical at fewer that 50 % of the bases of the entire lincRNA locus.

In order to compare rates of evolution in lincRNA genes and protein-coding genes, I carried out an identical pairwise-alignment experiment using the *Arabidopsis thaliana* coding regions annotated by TAIR10 (Lamesch et al., 2012). I aligned the coding sequences to the same genomic data that were used in the alignment of lincRNA genes, and processed the alignments using the same two criteria. The phylogenetic positions of these alignments are summarized in figure 4. Curiously, there were far more SHCBs than conserved blocks discovered using the *e*-value criterion in CDS regions (3 541 208 to 430 728), suggesting that long regions of high similarity are much more rare in lincRNA genes than in protein coding genes. The proportions of conserved blocks confined to the Brassicaceae was much lower in the CDS alignments (43 % of conserved blocks, 8.2 % of SHCBs), which suggests that lincRNA loci are subject to deletions and rapid sequence evolution much more frequently than protein-coding genes. In particular, 32 of 2000 protein coding genes had a significant conserved block in every genome analyzed, while only 4 lincRNA loci out of more than 6 600 had such a block.

The phylogenetic positions of conserved blocks in protein-coding genes and lincRNA genes are markedly different. Within the Brassicaceae, there are far more *Arabidopsis thaliana* lincRNA genes with conserved blocks in *A. lyrata* and no other species than protein coding genes, while protein coding genes are much more likely to be shared by all the Brassicaceae species examined (figure 4). A far larger proportion of lincRNA genes than protein coding genes lack conserved blocks outside of the Brassicaceae than protein coding genes. This difference is particularly striking when comparing alignments of protein coding genes and lincRNA genes between *Arabidopsis thaliana* and the Fabidae[1]. Within this clade, at least 33 % of the protein coding genes sampled have a conserved block, while this is true of only 0.3 % of lincRNA genes at most (table 4)

3.3 Conserved regions of *Arabidopsis* lincRNA loci

Because lincRNA genes with known functions tend to have relatively short, evolutionarily conserved functional regions (Ponting et al., 2009), I identified putatively conserved regions in my lincRNA data set which are present in every genomic region to which the locus aligns. For the purposes of this study, I defined a putative conserved region as the region of a lincRNA locus which is present in all of the significant genomic alignments of that locus at a particular stringency. A locus was considered not to have a conserved region if not all of the significant alignments overlapped, or if there was only one significant alignment. In total, I discovered 3178 conserved regions using the low stringency alignments and 462 using high stringency (table 5). The majority of conserved regions are found in close relatives of *Arabidopsis thaliana*.

In general, the trend in lincRNA genes is toward short regions of conservation in the centre of the gene flanked by relatively long regions which are highly divergent in different organisms. Outside of conserved regions, there is often evidence of dramatic sequence evolution, possibly including large scale insertions/deletions and a high rate of single nucleotide variation. This usually results in a long, unalignable region of the lincRNA locus outside of the conserved region. An example alignment of this is shown in the very deeply conserved alignment is shown in figure 5. Because the rate of interspecific sequence variation is so high, the alignments of non-conserved regions are not of sufficient quality to make a rigorous estimate of the rate of sequence evolution among species.

In order to detect conserved regions with reduced sequence rate evolution, I compared the number of intraspecific variations from the *Arabidopsis* 80 genomes project (Cao et al.,

[1]The Fabidae species included in the analysis are *Glycine max*, *Phaseolus vulgaris*, *Cannabis sativa*, *Malus domestica*, *Populus trichocarpa*, and *Ricinus communis*

2011) in the conserved regions to the number of variants in the surrounding lincRNA locus. The results are summarized in table 6. In total, I found 6 conserved regions which have experienced significantly fewer recent single-nucleotide mutations than the rest of the gene (6, $P \leq 0.05$, false discovery rate= 0.10). The alignments of these six conserved regions are shown in figure 6.

3.4 Recent evolutionary dynamics of lincRNA genes

I was able to use *Arabidopsis thaliana* ecotype resequencing data to examine the degree to which lincRNA loci are subject to large structural variation. Using annotation of structural variants among different *Arabidopsis thaliana* ecotypes from the 80 Genomes project (Cao et al., 2011), I examined the frequency with which entire lincRNA loci are deleted from the *Arabidopsis thaliana* genome relative to the Col-0 genome. Of the 6 681 lincRNA loci included in my analysis, I found 570 which were entirely missing in at least one *Arabidopsis thaliana* ecotype. Of these, 205 lacked any significant alignments to other plant genomes using either the high or low stringency criterion. Figure 7 summarizes the number of ecotypes in which a sequence that is unique to *Arabidopsis thaliana* is deleted. In the majority of cases, the locus is absent in only one or a few ecotypes, suggesting that these are cases of recent deletion of a locus which is present in most *Arabidopsis thaliana* individuals. In a small minority of cases, however, the locus in absent in virtually all *Arabidopsis thaliana* ecotypes except Col-0.

In the majority of cases (352/580) loci with annotated large insertion/deletion events within the *Arabidopsis thaliana* species lack a conserved region (see table 7). None of the loci with annotated insertion/deletion events are among the six which were found to have significantly fewer mutations in their conserved regions.

Chapter 4

Discussion

4.1 Identifying lincRNA genes

In total, my analysis of the Marquez et al. (2012) normalized library RNA-seq data reconfirmed 133 of the 6480 (2.0 %) lincRNA loci identified by Liu et al. (2012), and provided evidence for 201 novel lincRNA genes. Liu et al. (2012) had far more success with RNA-seq analysis: reconfirming more than 2700 loci out of the 6480 that were first identified with tiling array data. However, this is not an entirely fair comparison, since the two data sets differed markedly in the tissues prepared: Marquez et al. (2012) used flowers and whole seedlings, whereas Liu et al. (2012) used flower, leaf, root and silique samples. In addition, the two RNA-seq data sets were created using different platforms: Marquez et al. (2012) used five lanes of 75 nucleotide paired end reads on the Illumina GA system, whereas Liu et al. (2012) had four lanes of 101 nucleotide single end reads on the Illumina HiSequation 2000 platform. Although these differences prevent us from making a rigorous estimate of the degree to which library normalization improves detection of lincRNA transcripts, the relatively large number of novel lincRNA genes discovered by analysis of normalized RNA-seq data suggests that the procedure may provide a valuable increase in sensitivity. The fact that 201 novel lincRNA species were discovered in a single RNA-seq experiment when a novel tissue type is included suggests the possibility that there are many more undiscovered lincRNA transcripts in *Arabidopsis thaliana*.

Although Liu et al. (2012) found evidence for many more lincRNA transcripts through the analysis of tiling array data sets than either they or I were able to confirm through RNA-seq, this does not necessarily indicate that tiling arrays are a more sensitive tool. The technique which Liu et al. describe relies on an enormous volume of data: more than 200 data sets were included in the analysis, including RNA libraries from 14 different *Arabidopsis* mutants, 18 stress conditions and 6 tissue types. If all of these libraries

were submitted to an RNA-seq experiment rather than a tiling array, it is quite possible that many more lincRNA transcripts would have be discovered. Indeed, if the recent findings in mammals are any guide, there may be many thousands of as-yet unannotated *Arabidopsis* lincRNA genes (Managadze et al., 2013).

4.2 Conserved regions in lincRNA loci

Overall, the general pattern in conserved lincRNAs is patches of higher conservation within a poorly conserved overall sequence. The average conserved block discovered using the *e*-value filtering criterion covers slightly more than half of the lincRNA locus (table 5). Expanding the filtering criteria to include any regions of at least 85 % across 50 base pairs or more adds a large number of conserved blocks which cover only 14 % of the locus on average (table 5). Figure 5 shows a good example of an island of high conservation within lincRNA locus: the locus shown is 480 base pairs long, and has a conserved region of approximately 200 base pairs which is present in every plant genome included in this study. Across the rest of its length, however, I was unable to find any conserved blocks. The pattern of conserved regions within loci that are relatively poorly conserved overall is consistent with patterns of sequence evolution that have been found in functional lincRNA genes in mammals (Pang et al., 2006). Curiously, microRNA genes have also been observed to show a high rate of sequence evolution in the nucleotides flanking conserved hairpin structures (Berezikov et al., 2005).

That many lincRNA genes in my analysis lack a conserved region does not necessarily indicate a lack of functional importance, since many of these transcripts could have conserved functional regions too short to detect by primary sequence analysis alone, or that they have conserved secondary structure motifs that function without conserved regions of primary structure. Detecting the evolutionary conservation of such structures will doubtless be extremely challenging. It seems likely that, if any lincRNA genes which are functional only because of extremely short conserved regions or secondary structure, it will not be possible to effectively study the degree to which these structures are preserved by natural selection until they are characterized experimentally in a functional biology setting.

Although lincRNA genes in general have a high rate of sequence evolution, there are many lincRNA transcripts with known function which have short, conserved regions (Pang et al., 2006), and the annotation of conserved regions has been shown to be an effective strategy for finding functional lincRNA transcripts (Willingham et al., 2005). Therefore, the detection of areas of reduced sequence rate evolution within lincRNA loci is

a promising strategy for the discovery of transcripts of functional importance. Several of the conserved regions that were discovered through genomic alignments with other species show signs of reduced sequence rate evolution among different *Arabidopsis thaliana* lines. This is consistent with the hypothesis that these regions are of functional importance, possibly representing miRNA or protein binding sites (although, of course, this can only be conclusively demonstrated with functional studies). Although many conserved regions do not show reduced rates of sequence evolution compared to the rest of the locus, I cannot reject the hypothesis that these regions are of functional importance on this basis alone. In many cases, there is no variant data available at all for a particular lincRNA locus, making it impossible to draw a conclusion one way or the other regarding recent evolutionary conservation of the locus. It is possible that many more of the conserved regions in my data set are under purifying selection that is invisible due to lack of data.

Although reduced sequence rate evolution is not sufficient evidence to conclude that these lincRNA genes are functionally important, these results are very suggestive. The experience of animal researchers suggests that examining evolutionary conserved lincRNA genes is an effective strategy for discovering novel functionally important transcripts (Willingham et al., 2005). Similar studies of lincRNA conservation in animals suggest that there are thousands of non-coding transcripts whose functions have yet to be annotated (Guttman et al., 2009, 2010; Managadze et al., 2013). It would therefore be well worth exploring this set of evolutionarily conserved lincRNA genes in the context of a functional study.

Although evolutionary conservation of primary sequence has been shown to be an effective criterion for the discover of functional lincRNA transcripts, there are other avenues which are worth exploring. Many lincRNA transcripts in mammals have substantial predicted secondary structure (Ponting and Belgard, 2010; Tsai et al., 2011), and enrichment in secondary structures is known to be correlated with both evolutionary conservation and specificity of expression (Marques and Ponting, 2009). However, it has not been demonstrated *in vivo* that disrupting the secondary structure of any lincRNA gene disrupts its function. Nonetheless, it is well worth mining the set of conserved plant lincRNA genes for conserved secondary structural motifs.

4.3 Interpreting conserved sequence between lincRNA transcripts and genomic regions

My strategy for identifying lincRNA homologs is based on pair-wise genomic alignments. It is not possible to conclude with confidence that an alignment between an *Arabidopsis thaliana* lincRNA gene and the genome of another species represents a lincRNA gene in that species. It is possible that the locus is not transcribed in the other species, or even that it is both transcribed and translated into a short peptide. At present the transcriptome data available for non-*Arabidopsis* plant species are too thin to attempt a comprehensive genome-wide analysis of transcription and translation of lincRNA homologs in the rest of the plant kingdom on the scale that can be accomplished with a study of conserved genomic DNA elements. As deeper transcriptome and data sets become available in a variety of plant species, it will be interesting to see to what degree lincRNA genes transition between non-transcribed intergenic space, non-coding transcribed region and protein-coding sequence.

As an alternative to using publicly available transcriptome data, it may be possible to use comparable, matched RNA-seq data sets from two related plant species and compare the rate of conservation of lincRNA genes discovered *de novo*. Although these experiments would necessarily involve a smaller number of species than my approach, this would provide direct evidence that the lincRNA loci in question are expressed in both species. This approach is also likely to identify far fewer lincRNA loci than a tiling-based approach, which, as discussed, is more sensitive but requires many more individual experiments and depth of data.

High sequence similarity between lincRNA genes and genomic regions may be caused by the origins of the lincRNA gene, rather than because the gene really is shared amongst plant species. For example, the lincRNA gene At3NC056191 identified by Liu et al. (2012) has an alignment in every genome included in my study, but a BLAST search of the sequence suggests that these conserved regions are highly similar to ribosomal RNA sequence. This suggests that the locus in question may be descended from an rRNA gene, or possibly a previously unannotated gene copy in the rRNA family. In other cases, the sequence similarity may be due to the inclusion of partial, unannotated repeats, or even the inclusion of conserved DNA elements—such as promoters—in the locus. As with any other sequence alignment, researchers should be cautious about interpreting sequence similarity to be indicative of direct homology without independent confirmation.

4.4 Evolutionary comparisons of lincRNAs to protein coding genes and miRNAs

In contrast to protein coding genes, alignments of lincRNA genes are generally quite short. This is consistent with the hypothesis that lincRNA genes have only short stretches of primary sequence which are required for function, while the rest of the locus is relatively unconstrained in terms of evolution. This is not the case for protein coding genes, in which point mutations along much of the length can cause a disruptive frame shift mutation, dramatically altering the function. However, a similar pattern can be seen to a lesser extent in protein coding genes, in which the rate of sequence evolution is relatively slow in regions where the three dimensional structure of the protein is required for function and relatively rapid in 'intrinsically disordered' regions with no consistent tertiary structure (Brown et al., 2002).

My analysis shows that, compared with protein coding loci, lincRNA loci are generally less broadly conserved. This is consistent with the hypothesis that, in addition to a high rate of primary sequence evolution, lincRNA genes have a very rapid rate of emergence and decline within lineages (Hyashizaki, 2004; Ponting et al., 2009). Compared with protein coding sequences, lincRNA genes are apparently lost very frequently, as is evident in the relatively large number of deletions of lincRNA loci in different *Arabidopsis thaliana* ecotypes (figure 7). This raises the question of how lincRNA genes maintain their diversity in *Arabidopsis thaliana* despite a relatively high rate of loss.

Small RNA transcriptome sequencing studies of microRNA genes have shown that conservation is highly variable: some families are highly conserved throughout the plant kingdom while others are absent from the databases outside of *Arabidopsis thaliana* (Zhang et al., 2006). Although my alignments do not include secondary structure predictions, I find no evidence of a similar core group of highly conserved lincRNA genes. However, lincRNA genes apparently share a tendency with microRNA genes for rapid evolution, and frequent loss within different *Arabidopsis thaliana* ecotypes (Cao et al., 2011). Cao et al. (2011) found that microRNA genes which are deleted in at least one *Arabidopsis thaliana* ecotype are either not conserved in other plant species, or are members of large gene families. In microRNA genes, loss within *A. thaliana* is correlated with the presence of multiple-copy families. If lincRNA genes are also found in large families, that could partly explain their apparent tendency toward frequent deletion. It is also possible that lincRNA genes are frequently deleted due to redundancy in function with unrelated genes (which may or may not be lincRNAs), or that they have nonessential or nonexistent functions.

4.5 The origins of lincRNA genes

My analysis of structural variants in different *Arabidopsis thaliana* ecotypes suggests that there are a small number of lincRNA loci which are absent in the majority of strains aside from Col-0, and which do not appear to be highly conserved in any other plant species (figure 7). This suggests that these loci may have originated very recently as a result of large-scale structural mutations. Although it is extremely challenging to predict lincRNA genes which arose from such sequence rearrangements, there is at least one known case of a lincRNA gene which arose from a chromosomal rearrangement bringing together two previously untranscribed genomic regions (Ponting et al., 2009).

In mammals, lincRNAs do not apparently form large families by comparison to protein coding genes, which has lead to speculation that, while protein coding genes typically arise by duplication and divergence, lincRNAs and other non-coding genes may arise from intergenic space (Ponting et al., 2009). The extent to which lincRNA genes form families in plants is unclear. Research is underway in the Adams lab to determine the extent to which lincRNA genes are conserved after whole-genome and other duplication events. If indeed lincRNA genes are frequently duplicated, this could help to explain the apparently great diversity of lincRNA genes in *Arabidopsis* despite the frequency with which they undergo deletions. On the other hand, if lincRNA genes are not frequently duplicated and retained, they must commonly originate from other classes of genes or intergenic DNA.

There is already evidence for the origins of some lincRNA genes in coding sequences. The *Xist* gene in mammals, for example, has its origins in the pseudogenization of a protein-coding gene (Duret et al., 2006; Elisaphenko et al., 2008). It is also known that protein-coding genes can arise *de novo* from intergenic sequences (Carvunis et al., 2012), a process in which lincRNA genes may play a transitional role. Detailed studies of the origins of specific lincRNA genes are needed to address the issue of how these transcripts maintain their diversity in the face of frequent loss.

Conclusion

Studies in mammal suggest that assembling a comprehensive catalogue of the lincRNAs in a transcriptome requires a tremendous depth of sequencing coverage across many experiments due to the low expression levels and high tissue specificity of lincRNA transcripts. My results suggest that the situation is no different in *Arabidopsis*: a single lane of Illumina analysis has added 201 transcripts to the catalogue of known lincRNAs. There is every reason to expect that deeper coverage of the *Arabidopsis* non-coding transcriptome, aided by library normalization, will uncover many more lincRNA transcripts.

Although lincRNA loci clearly have a much higher rate of sequence evolution and turnover than protein coding genes, many have stretches of highly conserved nucleotides, and a few show signs of ongoing reduced sequence rate evolution. These patterns of evolution are consistent with what has been found in functional lincRNA genes in animals. On the other hand, the relatively high proportion of lincRNA loci which have experienced deletion in the recent evolutionary history of *Arabidopsis thaliana* suggests that many of these transcripts are non-functional, or have redundant functions. As more lincRNA genes are functionally characterized in plants, it will become clear what proportion of lincRNA transcripts have. However, the high rate of deletion of lincRNA loci among *Arabidopsis thaliana* ecotypes suggests that many such loci are not under strong purifying selection.

The high rate of turnover of lincRNA loci in plant genomes suggests that these transcripts may play a role in providing variation in the non-coding transcriptome which provides natural selection with raw material for the evolution of new functions. If lincRNA and other ncRNA transcripts arise frequently from intergenic space, transcripts which, by chance, have secondary structure or binding properties with beneficial functional consequences could be preserved by natural selection, resulting in *de novo* gene birth. New lincRNA loci may be the result of the evolution of new promoter elements in intergenic space by random drift. MicroRNA genes have been found to have originated in this way in *Drosophila* (Nozawa et al., 2010). Detailed studies of the origins of lincRNA transcripts from intergenic space are difficult (Ponting et al., 2009), but will be required in

order to determine the evolutionary roles of lincRNA genes.

Although the importance of lincRNA genes as a class is still unclear, we have tantalizing hints that these transcripts may be of evolutionary and functional importance. Studies of the evolutionary dynamics and degree of sequence conservation of lincRNA genes, such as this one, are the first step in determining what role they play in the function of organisms and the evolution of new genes.

Tables

Arabidopsis lyrata	Hu et al. (2011)
Capsella rubella	Slotte et al. (2013)
Brassica rapa	Wang et al. (2011)
Eutrema parvulum	Dassanayake et al. (2011)
Citrus clementia	International Citrus Genome Consortium (2011)
Gossypium raimondii	Wang et al. (2012)
Eucalyptus grandis	*Eucalyptus grandis* Genome Project (2010)
Glycine max	Schmutz et al. (2010)
Phaseolus vulgaris	DOE-JGI and USDA-NIFA (2013)
Malus domestica	Velasco et al. (2010)
Populus trichocarpa	Tuskan et al. (2006)
Ricinus communis	Chan et al. (2010)
Cannabis sativa	van Bakel et al. (2011)
Vitis vinifera	Jaillon et al. (2007)
Mimulus guttatus	*Mimulus* Genome Project and DOE-JGI (2013)
Solanum lycopersicum	Tomato Genome Consortium (2012)
Aquilegia coerulea	DOE-JGI (2013)
Brachopodium distachyon	International Brachypodium Initiative (2010)
Oryza sativa	Goff et al. (2002)
Zea mays	Schnable et al. (2009)
Selaginella moellendorfii	Banks et al. (2011)
Physcomitrella patens	Rensing et al. (2008)
Chlamydomonas reinhardtii	Merchant et al. (2007)

Table 1: Plant genomes used in the identification of *Arabidopsis thaliana* lincRNA conserved blocks by pairwise alignment.

e value	Frequency	Cumulative Frequecy	Relative Frequency
(0,1e-100]	3125	3125	0.01
(1e-100,1e-50]	18385	21510	0.04
(1e-50,1e-30]	27024	48534	0.06
(1e-30,1e-20]	22006	70540	0.05
(1e-20,1e-15]	18549	89089	0.04
(1e-15,1e-05]	85748	174837	0.19
(1e-05,0.0001]	19794	194631	0.04
(0.0001,0.001]	25307	219938	0.06
(0.001,0.01]	30525	250463	0.07
(0.01,0.1]	57089	307552	0.12
(0.1,1]	67078	374630	0.15
(1,10]	83708	458338	0.18

Table 2: Summary of the e values of the BLAST pairwise alignments. The breaks are exclusive at the lower limit, and inclusive at the upper limit. The total number of alignments at each e-value level includes both lincRNAs identified by Liu et al. (2012) and by my analysis of the Marquez et al. (2012) data (see text).

	Conserved Blocks	SHCBs
n	50955	29653
query length	356.42 ± 310.15	379.31 ± 271.38
% coverage	0.55 ± 0.31	0.14 ± 0.10
% identity	0.82 ± 0.062	0.87 ± 0.024

Table 3: Summary of the results of the two methods of annotating regions of high sequence similarity. 'Conserved blocks' are regions of a lincRNA locus with a MegaBLAST pairwise alignment to another genome with an e-value less than 10^{-30}. 'SHCBs' (short highly conserved blocks) are regions which were not annotated as conserved blocks at the first step, but which have a MegaBLAST alignment of at least 85 % identity across at least 50 base pairs. n is the total number of conserved blocks annotated at each step. 'Query length' is the average length of the lincRNA locus. '% coverage' is the average fraction of the lincRNA locus which is aligned. '% identity' is average fraction of bases which are identical in the lincRNA and genomic sequence. All measures of error are one standard deviation.

	lincRNA		CDS	
Species	Blocks	SHCBs	Blocks	SHCBs
Arabidopsis lyrata	5320	1033	1897	994
Capsella rubella	2547	491	1766	862
Brassica rapa	1655	427	1649	946
Eutrema parvulum	1945	301	1657	792
Citrus clementia	21	41	818	353
Gossypium raimondii	25	45	801	440
Eucalyptus grandis	20	38	689	341
Glycine max	21	45	730	410
Phaseolus vulgaris	18	28	675	327
Malus domestica	25	51	790	387
Populus trichocarpa	15	41	811	398
Ricinus communis	23	43	776	341
Vitis vinifera	21	38	758	378
Mimulus guttatus	18	40	622	296
Solanum lycopersicum	35	47	672	331
Aquilegia coerula	19	44	648	301
Brachopodium distachyon	10	21	383	189
Oryza sativa	14	23	383	224
Zea mays	16	27	380	283
Selaginella moellendorfii	7	16	151	108
Physcomitrella patens	4	22	178	137
Chlamydomonas reinhardtii	1	3	14	53

Table 4: The number of *Arabidopsis thaliana* lincRNA loci and protein coding sequences with conserved blocks in the genomes of other species. 'Blocks' indicates the number of genes with conserved blocks found using the e-value criterion described in the text, while 'SHCBs' (short highly conserved blocks) indicates number of genes which were added to the data set when any alignment with an identity of 85 % over at least 50 base pairs was also included. 'lincRNA' indicates alignments to one of the 6681 lincRNA genes included in this study, while the 'protein coding' alignments where made using the coding sequences of a random sample of the 2000 protein coding genes annotated by TAIR10.

	Conserved Blocks	SHCBs
Length	185.73 ± 91.70	121.98 ± 84.83
% Locus Length	58.8 ± 2.1	37.6 ± 2.6
% Identity	81.5 ± 6.1	88.3 ± 2.5

Table 5: Summary statistics of the conserved regions of lincRNA loci. '% Locus Length' is the length of the conserved region divided by the length of the locus, and '% identity' is the fraction of the bases of the conserved region which are identical in all alignments. All values are in the format 'mean ± standard deviation'. 'Conserved Blocks' indicates the conserved regions which were found using conserved blocks defined solely by the *e*-value criteria described in the text, while 'SHCBs' indicates the additional conserved regions which were annotated using short, highly conserved blocks (see text).

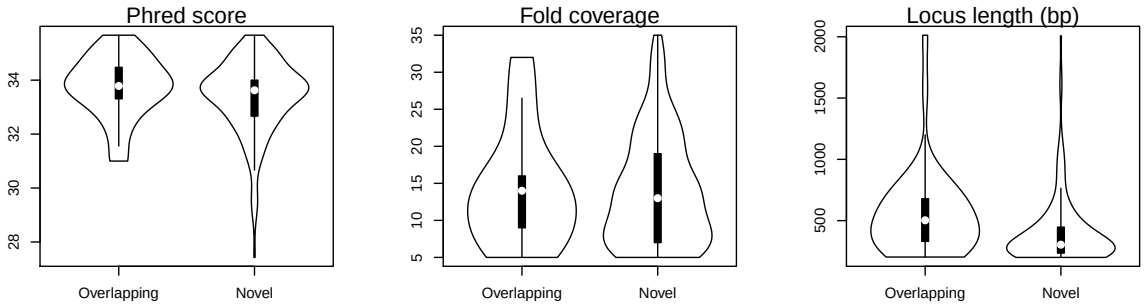
Locus ID	SNVs (inside/outside)	% length	P	Q
At5NC004520	11/21	73.2	$5.00 \cdot 10^{-6}$	$6.78 \cdot 10^{-3}$
At1NC064140	1/15	61.1	$7.23 \cdot 10^{-6}$	$6.78 \cdot 10^{-3}$
At5NC061480	9/75	31.2	$9.07 \cdot 10^{-6}$	$6.78 \cdot 10^{-2}$
At1NC027691	11/8	91.3	$1.05 \cdot 10^{-4}$	$6.25 \cdot 10^{-2}$
At1NC030450	0/25	29.4	$1.68 \cdot 10^{-4}$	$8.12 \cdot 10^{-2}$
At3NC014370	0/24	30.0	$1.90 \cdot 10^{-4}$	$8.12 \cdot 10^{-2}$

Table 6: Evidence for reduced rate of sequence evolution in conserved regions of lincRNA genes. SNVs are the number of distinct single nucleotide variants annotated by Cao et al. (2011) inside the conserved region and along the rest of the locus respectively. '% length' is the proportion of the lincRNA locus covered by the conserved region. P is the result of a binomial test with the null hypothesis that SNVs are equally likely or more likely to occur within the conserved region than in the non-conserved portions of the locus. Q values were obtained using Benjamini and Hochberg false discovery rate multiple test correction. Only loci with significantly fewer SNVs within their conserved regions at $\alpha = 0.05$ and $FDR = 0.1$ are shown.

		Conserved region	
		Present	Absent
Deletion	Present	218	352
	Absent	3156	3045

Table 7: Co-occurrence of conserved regions and whole-locus deletions in lincRNA loci. 'Deletion' refers to a deletion spanning the entire locus in at least one *Arabidopsis thaliana* ecotype as annotated by Cao et al. (2011). Conserved regions are defined in the text. Loci with conserved regions are significantly less likely to have whole-locus deletions ($P < 10^{-5}$, Fisher's exact test).

Figures



	Overlapping	Novel	
N	31	209	
Phred Score	33.73	33.39	
Fold Coverage	14.33	13.85	
Length	550.20	405.07	***

Figure 1: Violin plots of the differences in confidence statistics between putative lincRNA genes which are new in my analysis and those which were discovered independently by Liu et al. (2012). The white circle indicates the median, while the black rectangle spans the first through third interquartile range. The thin curves represent the density estimator. 'N' is the number of alignments in each category. '*Phred* score' is the average *Phred* score of the reads supporting the alignment (Ewing et al., 1998). 'Fold coverage' is the average number of reads which cover the locus at any base. 'Length' is the length of the locus in base pairs. The table gives the average values in each case. *** indicates a significant difference at $P < 0.0001$ (Wilcoxon rank sum test).

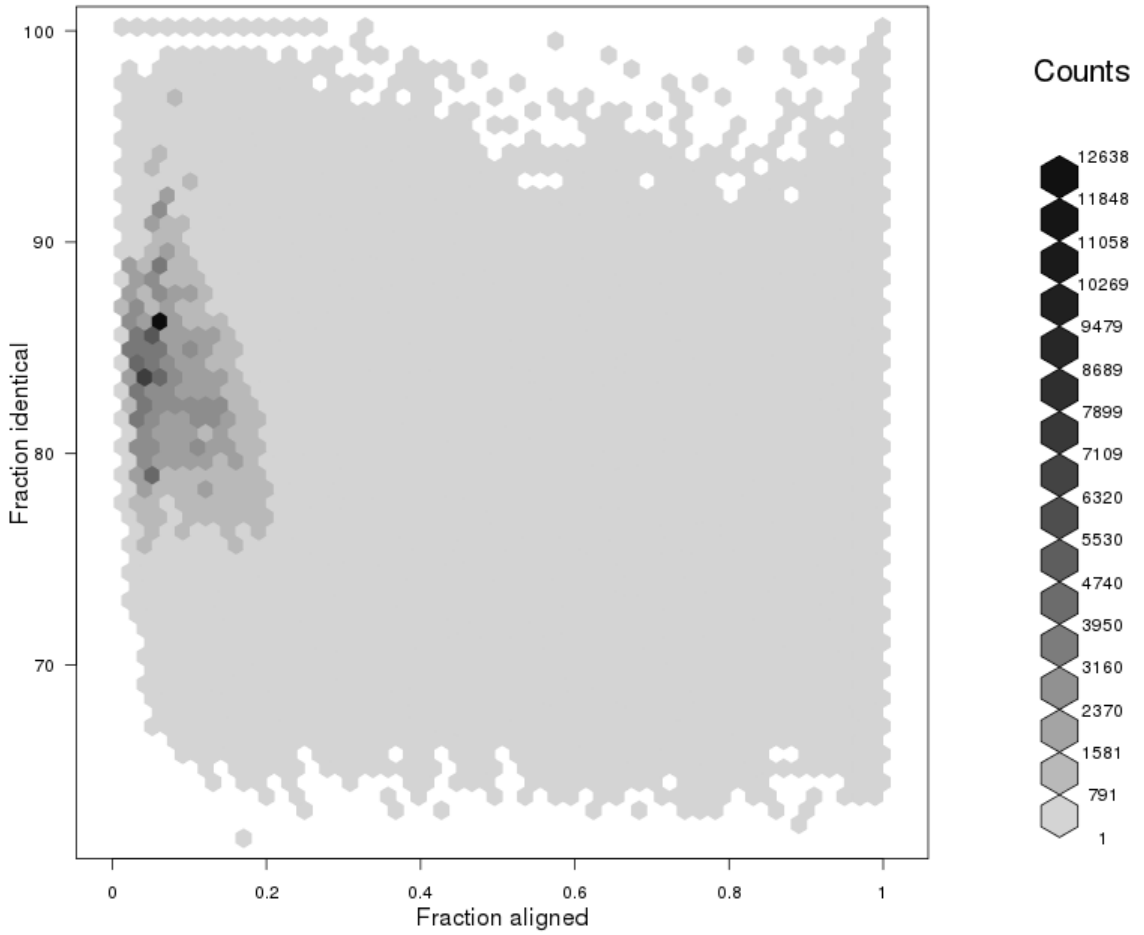


Figure 2: Scatter plot of the characteristics of the alignments to lincRNA genes in all species. 'Fraction aligned' is the proportion of the length of the lincRNA gene which can be aligned to a plant genome by MegaBLAST. 'Fraction identical' is the proportion of the alignment which is a perfect match to the target genome. The points have been binned into cells for ease of reading. Darker cells indicate a larger number of points.

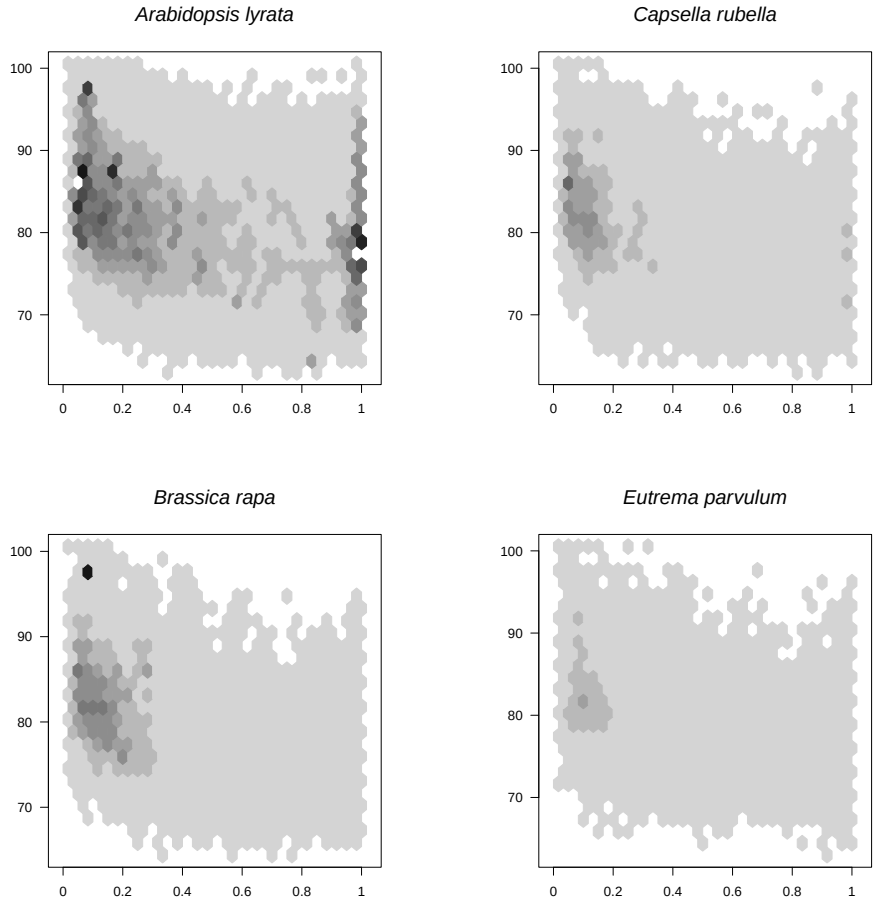
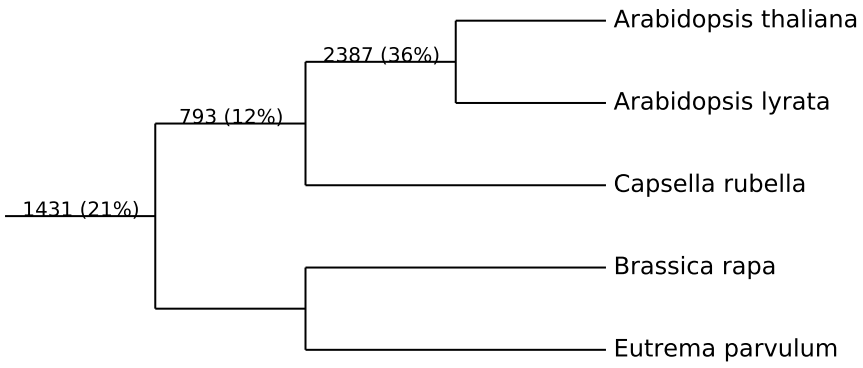
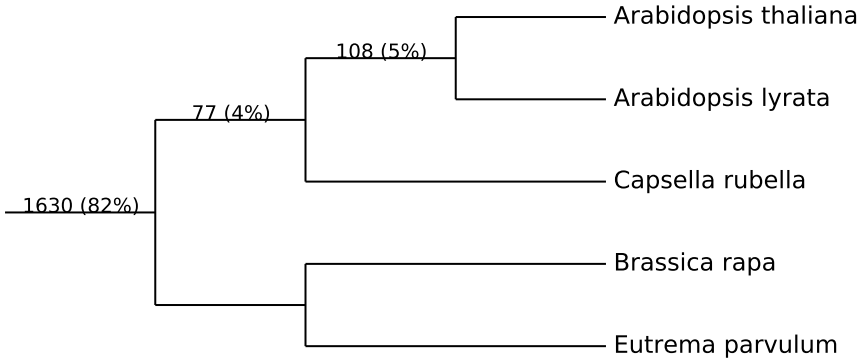


Figure 3: Scatter plots of the characteristics of the alignments to lincRNA genes in Brassicaceae species. The x-axis of each plot is the is the proportion of the length of the lincRNA gene which can be aligned to the indicated plant genome by MegaBLAST. The y-axis is the proportion of the alignment which is a perfect match to the target genome. The points have been binned into cells for ease of reading. A black cell indicates more than 1 000 hits, while the lightest grey shading indicates a single hit.



(a) lincRNA genes



(b) protein coding genes

Figure 4: Phylogenetic positions of conserved blocks of lincRNA genes and protein coding sequences within the Brassicaceae. Internal node labels indicate the number of loci with a conserved block in all of the members of that clade, but none of the other species in the phylogeny. The lincRNA genes are the 6681 *Arabidopsis thaliana* lincRNA loci annotated by Liu et al. (2012) and myself (see text). The protein coding genes are 2000 randomly selected coding sequences from the TAIR10 *Arabidopsis* annotations (Lamesch et al., 2012).

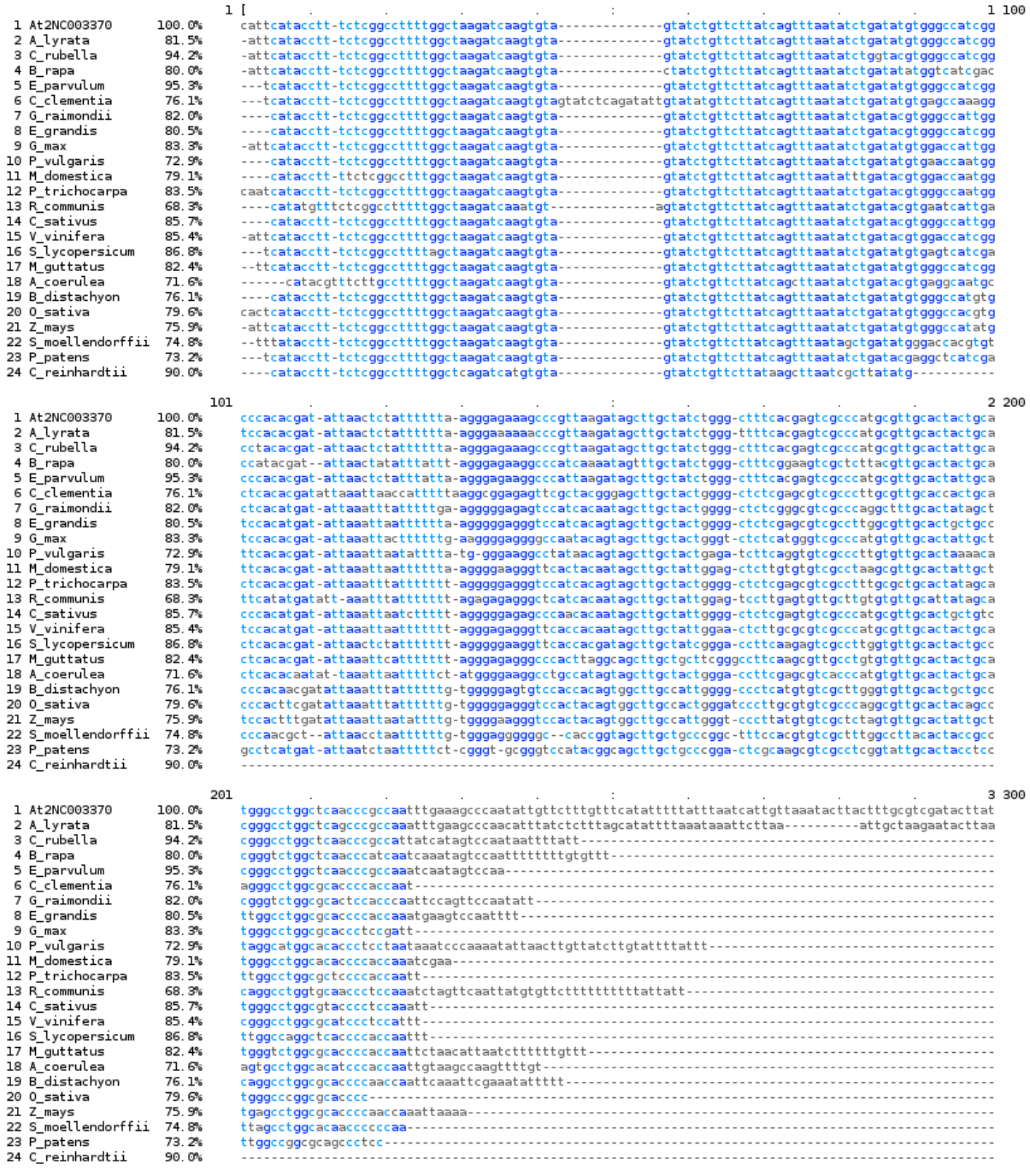


Figure 5: Partial alignment of a very broadly conserved lincRNA locus. The entire alignment is shown for the *Arabidopsis thaliana* locus, while the only the aligned portion of the locus is shown in other species. Alignments were performed using Clustal Omega (Sievers et al., 2011) and visualized using MView (Brown et al., 1998). Highlighting indicates identity to the reference *Arabidopsis thaliana* lincRNA sequence. When there were multiple alignments in a single species, the alignment with the fewest gaps is shown.


1 At5NC004520 100.0%
2 A_lyrata 89.6%
3 C_rubella 75.4%
4 B_rapa 63.1%
5 E_parvulum 75.2%


1 At1NC064140 100.0%
2 A_lyrata 91.7%
3 C_rubella 68.2%
4 B_rapa 80.7%
5 E_parvulum 91.3%


1 At5NC061480 100.0%
2 A_lyrata 79.9%
3 E_parvulum 66.8%


1 At1NC027691 100.0%
2 A_lyrata 64.1%
3 E_parvulum 55.3%

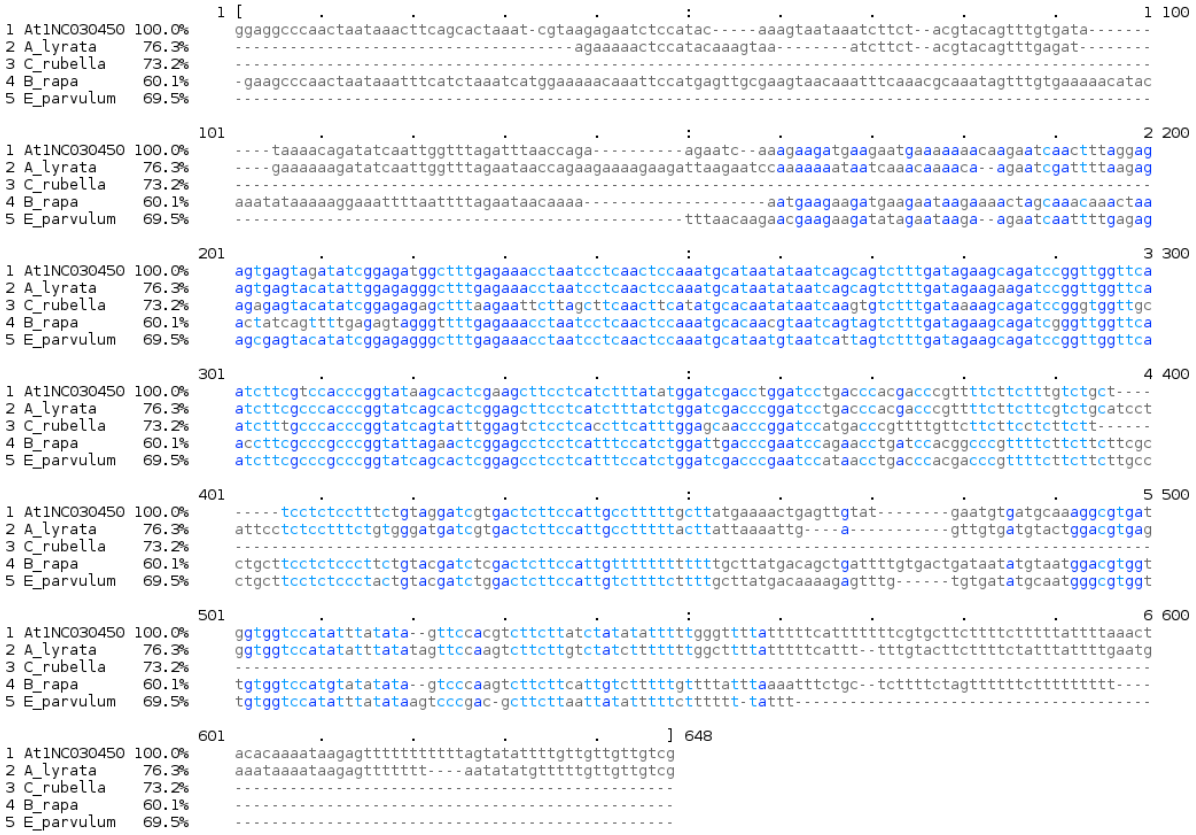

1 [. . . . : 1 100
1 At1NC030450 100.0% ggaggcccaactaataaacttcagcactaaat-cgtaagagaatctccatac-----aaagtaataaatcttct--acgtacagtttgtgata-------
2 A_lyrata 76.3% ----------agaaaaactccatacaaagtaa--------atcttct--acgtacagtttgagat--------
3 C_rubella 73.2% ----------
4 B_rapa 60.1% -gaagcccaactaataaatttcatctaaatcatggaaaaacaaattccatgagttgcgaagtaacaaatttcaaacgcaaatagtttgtgaaaaacatac
5 E_parvulum 69.5% ----------
101 . . . . : 2 200
1 At1NC030450 100.0% ----taaaacagatatcaattggtttagatttaaccaga----------agaatc--aaagaagatgaagaatgaaaaaacaagaatcaactttaggag
2 A_lyrata 76.3% ----gaaaaaagatatcaattggtttagaataaccagaagaaaagaagattaagaatccaaaaaaataatcaaacaaaaca--agaatcgattttaagag
3 C_rubella 73.2% ----------
4 B_rapa 60.1% aaatataaaaaggaaatttttaattttagaataacaaaa----------aatgaagaagatgaagaataagaaaactagcaaacaaactaa
5 E_parvulum 69.5% ----------tttaacaagaacgaagaagatatagaataaga--agaatcaatttttgagag
201 . . . . : 3 300
1 At1NC030450 100.0% agtgagtagatatcggagatggctttgagaaacctaatcctcaactccaaatgcataatataatcagcagtctttgatagaagcagatccggttggttca
2 A_lyrata 76.3% agtgagtacatattggagagggctttgagaaacctaatcctcaactccaaatgcataatataatcagcagtctttgatagaagaagatccggttggttca
3 C_rubella 73.2% agagagtacatatcggagagagctttaagaattcttagcttcaacttcatatgcacaatataatcaagtgtctttgataaaagcagatccgggtggttgc
4 B_rapa 60.1% actatcagttttgagagtagggttttgagaaacctaatcctcaactccaaatgcacaacgtaatcagtagtctttgatagaagcagatcgggttggttca
5 E_parvulum 69.5% agcgagtacatatcggagagggctttgagaaacctaatcctcaactccaaatgcataatgtaatcattagtctttgatagaagcagatccggttggttca
301 . . . . : 4 400
1 At1NC030450 100.0% atcttcgtccacccggtataagcactcgaagcttcctcatctttatatggatcgacctggatcctgacccacgacccgttttcttctttgtctgct----
2 A_lyrata 76.3% atcttcgcccacccggtatcagcactcggagcttcctcatctttatctggatcgacccggatcctgacccacgacccgttttcttcttcgtctgcatcct
3 C_rubella 73.2% atctttgcccacccggtatcagtatttggagtctcctcaccttcatttggagcaacccggatccatgacccgttttgttcttcttcctcttctt------
4 B_rapa 60.1% accttcgcccgcccggtattagaactcggagcctcctcatttccatctggattgacccgaatccagaacctgatccacggcccgttttcttcttcttcgc
5 E_parvulum 69.5% atcttcgcccgcccggtatcagcactcggagcctcctcatttccatctggatcgacccgaatccataacctgacccacgacccgttttcttcttcttgcc
401 . . . . : 5 500
1 At1NC030450 100.0% -----tcctctcctttctgtaggatcgtgactcttccattgccttttgcttatgaaaactgagttgtat--------gaatgtgatgcaaaggcgtgat
2 A_lyrata 76.3% attcctctcctttctgtgggatgatcgtgactcttccattgccttttttacttattaaaattg----a----------gttgtgatgtactggacgtgag
3 C_rubella 73.2% ----------
4 B_rapa 60.1% ctgcttcctctccctttctgtacgatctcgactcttccattgtttttttttttttgcttatgacagctgattttgtgactgataatatgtaatggacgtggt
5 E_parvulum 69.5% ctgcttcctctccctactgtacgatctggactcttccattgtcttttctttttgcttatgacaaaagagtttg------tgtgatatgcaatgggcgtggt
501 . . . . : 6 600
1 At1NC030450 100.0% ggtggtccatatttatata--gttccacgtcttcttatctatatatttttgggtttatttttcatttttttcgtgcttcttttctttttattttaaact
2 A_lyrata 76.3% ggtggtccatatatttatatagttccaagtcttcttgtctatctttttttggcttttatttttcattt--tttgtacttctttttctatttatttttgaatg
3 C_rubella 73.2% ----------
4 B_rapa 60.1% tgtggtccatgtatatata--gtcccaagtcttcttcattgtctttttgttttatttaaaatttctgc--tcttttctagttttttctttttttttt----
5 E_parvulum 69.5% tgtggtccatatttatataagtcccgac-gcttcttaattatatttttctttttt-tattt----------
601 . . .] 648
1 At1NC030450 100.0% acacaaaataagagttttttttttttagtatattttgttgttgttgtcg
2 A_lyrata 76.3% aaataaaataagagtttttttt----aatatatgttttttgttgttgtcg
3 C_rubella 73.2% ----------
4 B_rapa 60.1% ----------
5 E_parvulum 69.5% ----------

Figure 6: Alignments of the *Arabidopsis thaliana* lincRNA genes which have evidence of reduced sequence rate evolution within their conserved regions. The entire alignment is shown for the *Arabidopsis thaliana* locus, while the only the aligned portion of the locus is shown in other species. A gap in the alignment outside of the conserved region therefore does not necessarily indicate that the lincRNA gene is shorter in that species, only that the sequence is so different as to be unalignable. Alignments were performed using Clustal Omega (Sievers et al., 2011) and visualized using MView (Brown et al., 1998). Highlighting indicates identity to the reference *Arabidopsis thaliana* lincRNA sequence. When there were multiple alignments in a single species, the alignment with the fewest gaps is shown.

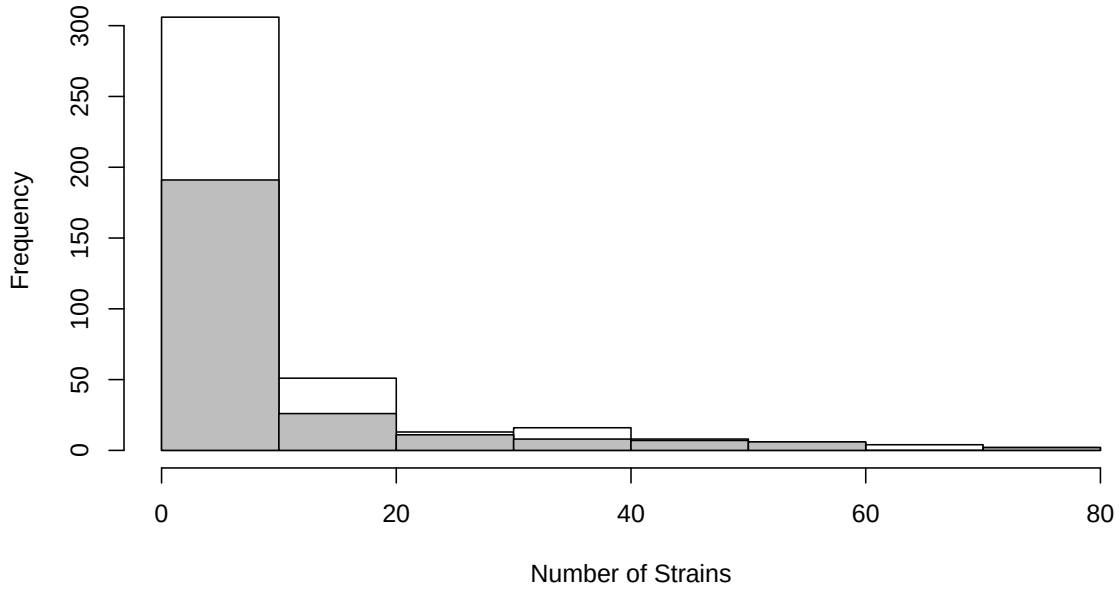


Figure 7: Frequency of deletions of *Arabidopsis thaliana* lincRNA loci which do not have a significant alignment to any other plant genome. Grey bars represent loci which do not have low stringency alignment while white bars represent loci which are unaligned using the high stringency criterion. A deletion of the locus is defined as a deletion event predicted by Cao et al. (2011) which includes the entire lincRNA locus in at least one of the 80 strains examined.

Bibliography

Banks, J. A., Nishiyama, T., Hasebe, M., Bowman, J. L., Gribskov, M., DePamphilis, C., Albert, V. A., Aono, N., Aoyama, T., Ambrose, B. A., Ashton, N. W., Axtell, M. J., Barker, E., Barker, M. S., Bennetzen, J. L., Bonawitz, N. D., Chapple, C., Cheng, C., Correa, L. G. G., Dacre, M., DeBarry, J., Dreyer, I., Elias, M., Engstrom, E. M., Estelle, M., Feng, L., Finet, C., Floyd, S. K., Frommer, W. B., Fujita, T., Gramzow, L., Gutensohn, M., Harholt, J., Hattori, M., Heyl, A., Hirai, T., Hiwatashi, Y., Ishikawa, M., Iwata, M., Karol, K. G., Koehler, B., Kolukisaoglu, U., Kubo, M., Kurata, T., Lalonde, S., Li, K., Li, Y., Litt, A., Lyons, E., Manning, G., Maruyama, T., Michael, T. P., Mikami, K., Miyazaki, S., Morinaga, S.-i., Murata, T., Mueller-Roeber, B., Nelson, D. R., Obara, M., Oguri, Y., Olmstead, R. G., Onodera, N., Petersen, B. L., Pils, B., Prigge, M., Rensing, S. A., Riaño Pachón, D. M., Roberts, A. W., Sato, Y., Scheller, H. V., Schulz, B., Schulz, C., Shakirov, E. V., Shibagaki, N., Shinohara, N., Shippen, D. E., S.ø rensen, I., Sotooka, R., Sugimoto, N., Sugita, M., Sumikawa, N., Tanurdzic, M., Theissen, G., Ulvskov, P., Wakazuki, S., Weng, J.-K., Willats, W. W. G. T., Wipf, D., Wolf, P. G., Yang, L., Zimmer, A. D., Zhu, Q., Mitros, T., Hellsten, U., Loqué, D., Otillar, R., Salamov, A., Schmutz, J., Shapiro, H., Lindquist, E., Lucas, S., Rokhsar, D., and Grigoriev, I. V. (2011). The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science*, 332:960–9633.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. A., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120:21–24.

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., and Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution*, 55:104–110.

Brown, N. P., Leroy, C., and Sander, C. (1998). MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, 14:380–381.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*, 268:78–94.

Burleigh, S. H. and Harrison, M. J. (1997). A novel gene whose expression in *Medicago truncatula* roots is suppressed in response to colonization by vesicular-arbuscular mycorrhizal (VAM) fungi and to phosphate nutrition. *Plant Molecular Biology*, 34:199–208.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & Development*, 25:1915–1927.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Müller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K. J., and Weigel, D. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nature Genetics*, 43:956–963.

Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., and Vidal, M. (2012). Proto-genes and de novo gene birth. *Nature*, 487:370–374.

Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J., Wheeler, R., Wong, B., Drenkow, J., Yamanaka, M., Patel, S., Brubaker, S., Tammana, H., Helt, G., Struhl, K., and Gingeras, T. R. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116:499–509.

Chan, A. P., Crabtree, J., Zhao, Q., Lorenzi, H., Orvis, J., Puiu, D., Melake-Berhan, A., Jones, K. M., Redman, J., Chen, G., Cahoon, E. B., Gedil, M., Stanke, M., Haas, B. J., Wortman, J. R., Fraser-Liggett, C. M., Ravel, J., and Rabinowicz, P. D. (2010). Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology*, 28:951–6.

Dassanayake, M., Oh, D.-H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R. A., Zhu, J.-K., Bohnert, H. J., and Cheeseman, J. M. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. *Nature Genetics*, 43:913–918.

DOE-JGI (2013). *Aquilegia coerulea* v1.0. Retrieved August 2, 2013, from http://www.phytozome.net/aquilegia.php.

DOE-JGI and USDA-NIFA (2013). *Phaseolus vulgaris* v0.9. Retrieved August 2, 2013, from http://www.phytozome.net/commonbean.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science*, 312:1653–1655.

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogozin, I. B., Nesterova, T. B., Brockdorff, N., and Zakian, S. M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS ONE*, 3:11.

ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816.

Eucalyptus grandis Genome Project (2010). Retrieved August 2, 2013, from http://www.phytozome.net/eucalyptus.php.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using *Phred*. I. accuracy assessment. *Genome Research*, 8:175–185.

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garca, J. A., and Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nature Genetics*, 39:1033–1037.

Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B. M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.-l., Chen, L., Cooper, B., Park, S., Wood, T. C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R. M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., and Briggs, S. (2002). A draft sequence of the rice genome (*Oryza sativa* L. *ssp. japonica*). *Science*, 296:92–100.

Guffanti, A., Iacono, M., Pelucchi, P., Kim, N., Soldà, G., Croft, L. J., Taft, R. J., Rizzi, E., Askarian-Amiri, M., Bonnal, R. J., Callari, M., Mignone, F., Pesole, G., Bertalot, G., Bernardi, L. R., Albertini, A., Lee, C., Mattick, J. S., Zucchi, I., and De Bellis, G. (2009). A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. *BMC Genomics*, 10:163.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464:1071–1076.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology*, 28:503–510.

Heo, J. B. and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*, 331:76–79.

Hou, X. L., Wu, P., Jiao, F. C., Jia, Q. J., Chen, H. M., Yu, J., Song, X. W., and Yi, K. K. (2005). Regulation of the expression of OsIPS1 and OsIPS2 in rice via systemic and local Pi signalling and hormones. *Plant, Cell & Environment*, 28:356–364.

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.-F., Clark, R. M., Fahlgren, N., Fawcett, J. A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J. D., Ossowski, S., Ottilar, R. P., Salamov, A. A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M. E., Bergelson, J., Carrington, J. C., Gaut, B. S., Schmutz, J., Mayer, K. F. X., Van de Peer, Y., Grigoriev, I. V., Nordborg, M., Weigel, D., and Guo, Y.-L. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*, 43:476–481.

Hupalo, D. and Kern, A. D. (2013). Conservation and functional element discovery in 20 angiosperm plant genomes. *Molecular Biology and Evolution*, 30:1729–1744.

Hyashizaki, Y. (2004). Mouse transcriptome: Neutral evolution of non-coding complementary DNAs (reply). *Nature*, 431:757.

International Brachypodium Initiative (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463:763–768.

International Citrus Genome Consortium (2011). Haploid *Clementia* genome. http://int-citrusgenomics.org/. Retrieved August 2, 2013, from http://www.phytozome.net/clementine.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M. E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.-F., Weissenbach, J., Quétier, F., and Wincker, P. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449:463–7.

Jeon, Y. and Lee, J. T. (2011). YY1 tethers Xist RNA to the inactive X nucleation center. *Cell*, 146:119–133.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni,

A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488.

Kim, E.-D. and Sung, S. (2012). Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends in Plant Science*, 17:16–21.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., and Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, 40:D1202–D1210.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25:2078–2079.

Liu, C., Muchhal, U. S., and Raghothama, K. (1997). Differential expression of TPS11, a phosphate starvation-induced gene in tomato. *Plant Molecular Biology*, 33:867–874.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N.-H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in arabidopsis. *The Plant Cell*, 24:4333–4345.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445.

Managadze, D., Lobkovsky, A. E., Wolf, Y. I., Shabalina, S. A., Rogozin, I. B., and Koonin, E. V. (2013). The vast, conserved mammalian linc-RNome. *PLoS Computational Biology*, 9:e1002917.

Marques, A. C. and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biology*, 10:R124.

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, 22:1184–1195.

Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T. A., Okamoto, M., Nambara, E., Nakajima, M., Kawashima, M., Satou, M., Kim, J.-M., Kobayashi, N., Toyoda, T., Shinozaki, K., and Seki, M. (2008). Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array. *Plant & Cell Physiology*, 49:1135–1149.

Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K., Maréchal-Drouard, L., Marshall, W. F., Qu, L.-H., Nelson, D. R., Sanderfoot, A. A., Spalding, M. H., Kapitonov, V. V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S. M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C.-L., Cognat, V., Croft, M. T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P. A., Lemaire, S. D., Lobanov, A. V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J. V., Moseley, J., Napoli, C., Nedelcu, A. M., Niyogi, K., Novoselov, S. V., Paulsen, I. T., Pazour, G., Purton, S., Ral, J.-P., Riaño Pachón, D. M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S. L., Allmer, J., Balk, J., Bisova, K., Chen, C.-J., Elias, M., Gendler, K., Hauser, C., Lamb, M. R., Ledford, H., Long, J. C., Minagawa, J., Page, M. D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A. M., Yang, P., Ball, S., Bowler, C., Dieckmann, C. L., Gladyshev, V. N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R. T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y. W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W. C. A., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I. V., Rokhsar, D. S., and Grossman, A. R. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science*, 318:245–250.

Mimulus Genome Project and DOE-JGI (2013). Retrieved August 2, 2013, from http://www.phytozome.net/mimulus.php.

Ming, R., Hou, S., Feng, Y., Yu, Q., Laporte, A., Saw, J. H., Senin, P., Wang, W., Ly, B. V., Lewis, K. L. T., Salzberg, S. L., Feng, L., Jones, M. R., Skelton, R. L., Murray, J. E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R. E., Michael, T. P., Wall, K., Rice, D. W., Albert, H., Li, Zhu, Y. J., Schatz, M., Nagarajan, N., Acob, R. A., Guan, P., Blas, A., Wai, C. M., Ackerman, C. M., Ren, Y., Liu, C., Wang, J., Wang, J., Kuk, Shakirov, E. V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J. E., Gschwend, A. R., Delcher, A. L., Singh, R., Suzuki, J. Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Perez, R., Torres, M. J., Feltus, F. A., Porter, B., Li, Y., Burroughs, A. M., Cheng, Liu, L., Christopher, D. A., Mount, S. M., Moore, P. H., Sugimura, T., Jiang, J., Schuler, M. A., Friedman, V., Olds, T., Shippen, D. E., dePamphilis, C. W., Palmer, J. D., Freeling, M., Paterson, A. H., Gonsalves, D., Wang, L., and Alam, M. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature*, 452:991–996.

Molnar, A., Melnyk, C., and Baulcombe, D. C. (2011). Silencing signals in plants: a long journey for small RNAs. *Genome Biology*, 12:215.

Nozawa, M., Miura, S., and Nei, M. (2010). Origins and evolution of microRNA genes in *Drosophila* species. *Genome Biology and Evolution*, 2:180–189.

Numata, K., Kanai, A., Saito, R., Kondo, S., Adachi, J., Wilming, L. G., Hume, D. A., Hayashizaki, Y., and Tomita, M. (2003). Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Research*, 13:1301–1306.

Pang, K. C., Frith, M. C., and Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics*, 22:1–5.

Ponting, C. P. and Belgard, T. G. (2010). Transcribed dark matter: meaning or myth? *Human Molecular Genetics*, 19:R162–R168.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, 136:629–641.

Rensing, S. A., Lang, D., Zimmer, A. D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.-F., Lindquist, E. A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin-I, T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S.-I., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W. B., Barker, E., Bennetzen, J. L., Blankenship, R., Cho, S. H., Dutcher, S. K., Estelle, M., Fawcett, J. A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K. A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D. R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P. J., Sanderfoot, A., Schween, G., Shiu, S.-H., Stueber, K., Theodoulou, F. L., Tu, H., Van de Peer, Y., Verrier, P. J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A. C., Hasebe, M., Lucas, S., Mishler, B. D., Reski, R., Grigoriev, I. V., Quatrano, R. S., and Boore, J. L. (2008). The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science*, 319:64–69.

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.-C., Shinozaki, K., Nguyen, H. T., Wing, R. A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R. C., and Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, 463:178–183.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A.,

Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326:1112–1115.

Scott, M. S. and Ono, M. (2011). From snoRNA to miRNA: Dual function regulatory non-coding RNAs. *Biochimie*, 93:1987–1992.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7:539.

Slotte, T., Hazzouri, K. M., Agren, J. A., Koenig, D., Maumus, F., Guo, Y.-L., Steige, K., Platts, A. E., Escobar, J. S., Newman, L. K., Wang, W., Mandáková, T., Vello, E., Smith, L. M., Henz, S. R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., Andolfatto, P., Hu, T. T., Blanchette, M., Clark, R. M., Quesneville, H., Nordborg, M., Gaut, B. S., Lysak, M. A., Jenkins, J., Grimwood, J., Chapman, J., Prochnik, S., Shu, S., Rokhsar, D., Schmutz, J., Weigel, D., and Wright, S. I. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nature Genetics*, 45:831–835.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485:635–41.

Tsai, M.-C., Spitale, R. C., and Chang, H. Y. (2011). Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Research*, 71:3–7.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Déjardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood,

J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjärvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leplé, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouzé, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa (Torr. & Gray)*. *Science*, 313:1596–1604.

van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., and Page, J. E. (2011). The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology*, 12:R102.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., Pruss, D., Salvi, S., Pindo, M., Baldi, P., Castelletti, S., Cavaiuolo, M., Coppola, G., Costa, F., Cova, V., Dal Ri, A., Goremykin, V., Komjanc, M., Longhi, S., Magnago, P., Malacarne, G., Malnoy, M., Micheletti, D., Moretto, M., Perazzolli, M., Si-Ammour, A., Vezzulli, S., Zini, E., Eldredge, G., Fitzgerald, L. M., Gutin, N., Lanchbury, J., Macalma, T., Mitchell, J. T., Reid, J., Wardell, B., Kodira, C., Chen, Z., Desany, B., Niazi, F., Palmer, M., Koepke, T., Jiwan, D., Schaeffer, S., Krishnan, V., Wu, C., Chu, V. T., King, S. T., Vick, J., Tao, Q., Mraz, A., Stormo, A., Stormo, K., Bogden, R., Ederle, D., Stella, A., Vecchietti, A., Kater, M. M., Masiero, S., Lasserre, P., Lespinasse, Y., Allan, A. C., Bus, V., Chagné, D., Crowhurst, R. N., Gleave, A. P., Lavezzo, E., Fawcett, J. A., Proost, S., Rouzé, P., Sterck, L., Toppo, S., Lazzari, B., Hellens, R. P., Durel, C.-E., Gutin, A., Bumgarner, R. E., Gardiner, S. E., Skolnick, M., Egholm, M., Van de Peer, Y., Salamini, F., and Viola, R. (2010). The genome of the domesticated apple (*Malus x domestica* Borkh.). *Nature Genetics*, 42:833–839.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., Zhu, S., Zou, C., Li, Q., Yuan, Y., Lu, C., Wei, H., Gou, C., Zheng, Z., Yin, Y., Zhang, X., Liu, K., Wang, B., Song, C., Shi, N., Kohel, R. J., Percy, R. G., Yu, J. Z., Zhu, Y.-X., Wang, J., and Yu, S. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. *Nature Genetics*, 44:1098–1103.

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., Bai, Y., Mun, J.-H., Bancroft, I., Cheng, F., Huang, S., Li, X., Hua, W., Wang, J., Wang, X., Freeling, M., Pires, J. C., Paterson, A. H., Chalhoub, B., Wang, B., Hayward, A., Sharpe, A. G., Park, B.-S., Weisshaar, B., Liu, B., Li, B., Liu, B., Tong, C., Song, C., Duran, C., Peng, C., Geng, C., Koh, C., Lin, C., Edwards, D., Mu, D., Shen, D., Soumpourou, E., Li, F., Fraser, F., Conant, G., Lassalle, G., King, G. J., Bonnema, G., Tang, H., Wang, H., Belcram, H., Zhou, H., Hirakawa, H., Abe, H., Guo, H., Wang, H., Jin, H., Parkin, I.

A. P., Batley, J., Kim, J.-S., Just, J., Li, J., Xu, J., Deng, J., Kim, J. A., Li, J., Yu, J., Meng, J., Wang, J., Min, J., Poulain, J., Hatakeyama, K., Wu, K., Wang, L., Fang, L., Trick, M., Links, M. G., Zhao, M., Jin, M., Ramchiary, N., Drou, N., Berkman, P. J., Cai, Q., Huang, Q., Li, R., Tabata, S., Cheng, S., Zhang, S., Zhang, S., Huang, S., Sato, S., Sun, S., Kwon, S.-J., Choi, S.-R., Lee, T.-H., Fan, W., Zhao, X., Tan, X., Xu, X., Wang, Y., Qiu, Y., Yin, Y., Li, Y., Du, Y., Liao, Y., Lim, Y., Narusaka, Y., Wang, Y., Wang, Z., Li, Z., Wang, Z., Xiong, Z., and Zhang, Z. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. *Nature Genetics*, 43:1035–1039.

Willingham, A. T., Orth, A. P., Batalov, S., Peters, E. C., Wen, B. G., Aza-Blanc, P., Hogenesch, J. B., and Schultz, P. G. (2005). A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, 309:1570–1573.

Young, R. S., Marques, A. C., Tibbit, C., Haerty, W., Bassett, A. R., Liu, J.-L., and Ponting, C. P. (2012). Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biology and Evolution*, 4:427–442.

Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006). Conservation and divergence of plant microRNA genes. *The Plant Journal*, 46:243–259.

Zhang, B., Wang, Q., and Pan, X. (2007). MicroRNAs and their regulatory roles in animals and plants. *Journal of Cellular Physiology*, 210:279–289.

Zhang, Y., Liu, J., Jia, C., Li, T., Wu, R., Wang, J., Chen, Y., Zou, X., Chen, R., Wang, X.-J., and Zhu, D. (2010). Systematic identification and evolutionary features of rhesus monkey small nucleolar RNAs. *BMC Genomics*, 11:61.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, 7:203–214.

Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J., and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science*, 322:750–756.