GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Evolution and epigenetic regulation of RNA-mediated duplicated genes in *Arabidopsis*

Dissertation

For the award of the degree

“Doctor rerum naturalium” (Dr.rer.nat.)

of the Georg August University Göttingen

within the doctoral program: Biology

of the Georg-August University School of Science (GAUSS)

submitted by

Ahmed Mahmoud Abdelsamad Abdrabou

from Cairo, Egypt

Göttingen, 2015

Thesis Committee

Prof. Dr. Christiane Gatz
(Department of Plant Molecular Biology and Physiology / Schwann-Schleiden-Research-Center for Molecular Cell Biology, University of Göttingen)

Prof. Dr. Stefan Jakobs
(Department of Nanobiophotonics/ Structure and Dynamics of Mitochondria Group, Max Planck Institute for Biophysical Chemistry and University of Göttingen)

Dr. Ales Pecinka
(Department of Plant Breeding and Genetics / Genome and Epigenome Evolution Group, Max Planck Institute for Plant Breeding Research)

Members of the Examination Board

Reviewer: Prof. Dr. Christiane Gatz
(Department of Plant Molecular Biology and Physiology / Schwann-Schleiden-Research-Center for Molecular Cell Biology)

Second Reviewer: Prof. Dr. Stefan Jakobs
(Department of Nanobiophotonics/ Structure and Dynamics of Mitochondria Group, Max Planck Institute for Biophysical Chemistry and University of Göttingen)

Further members of the Examination Board:

Prof. Dr. Gregor Bucher
(Department of Developmental Biology / Göttingen Center for Molecular Biosciences)

Prof. Dr. André Fischer
(Department of Psychiatry and Psychotherapy, University Medical Center)

Prof. Dr. Volker Lipka
(Department of Plant Cell Biology / Schwann-Schleiden-Research-Center for Molecular Cell Biology)

Prof. Dr. Burkhard Morgenstern
(Department of Bioinformatics/ Institute for Microbiology and Genetics)

Date of the oral examination: 15. 06. 2015

TABLE of CONTENTS

ABSTRACT

Gene duplications allow for protein functional diversification and accelerate genome evolution. Occasionally, the transposon amplification machinery reverse-transcribes mRNA of a gene, integrates it into the genome and forms an RNA-duplicated gene copy, the retrogene. Although retrogenes have been found in plants, their biology, evolution and epigenetic regulation are poorly understood. We developed a novel bioinformatic retrogene annotation tool (RAT) to screen *Arabidopsis* genomes for retrogenes. We identified 251 (216 novel) and 168 retrogenes in *Arabidopsis thaliana* and *Arabidopsis lyrata*, corresponding to 1% and 0.5% of protein coding genes respectively. Based on our findings, we calculated emergence rate of five to ten retrogenes per million years, which is at least ten times faster than previously estimated. Most of retrogenes were randomly integrated away from their parental gene loci; however, some showed targeted integration replacing their parental genes. Therefore, we developed a bioinformatic targeted retrogene annotation tool (TRAT) to screen *Arabidopsis* genomes for these rare cases. To our knowledge, we report the first natural *in planta* retrogene targeting events.

Arabidopsis retrogenes are derived from ubiquitously transcribed parents and reside in gene rich chromosomal regions, depleted of transposons. Unlike transposon regulation, we found retrogenes and their parents to be targets of gene-specific regulatory 21 nt sRNAs rather than transposon-specific 24 nt sRNAs. Retrogene expression levels are relatively low, but significantly higher than that of transposable elements. Approximately 25% of retrogenes are co-transcribed with their parents, and 3% with head-to-head oriented neighbors. This suggests transcription by novel or modified promoters for at least 72% of *A. thaliana* retrogenes. Many retrogenes reach their transcription maximum in pollen, the tissue analogous to animal spermatocytes where up-regulation of retrogenes has previously been found. This implies an evolutionarily conserved mechanism leading to this transcription pattern of RNA-duplicated genes. During transcriptional repression, retrogenes are depleted of permissive chromatin marks without an obvious enrichment for repressive modifications. However, this pattern is common to many other pollen-transcribed genes independent of their evolutionary origin. Hence, retroposition plays role in plant genome evolution and developmental transcription pattern of retrogenes suggests analogous control of RNA-duplicated genes in plants and animals.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

cDNA	complementary DNA
CDS	Coding DNA Sequence
Chr.	Chromosome
DNA	Deoxyribonucleic Acid
gcRMA	guanine cytosine Robust Multi-array Average
gDNA	genomic DNA
GO	Gene Ontology
GW	Genome Wide
H3K9me2	Histone H3 lysine 9 di-methylation
H3K27me3	Histone H3 lysine 27 tri-methylation
kbp	kilo base pair
LINE	Long Interspersed Nuclear Element
MWW	Mann-Whitney-Wilcoxon
MYA	Million Year Ago
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
RPKM	Reads Per Kilobase per Million mapped reads
TE	Transposable Element
TSS	Transcription Start Site
TTS	Transcription Termination Site

ACKNOWLEDGEMENT

I would like to thank Dr. Ales Pecinka for supervision during my work in his group at Max Planck Institute for Plant breeding research. I'm very grateful for Professor Dr. Maarten Koornneef, not only for allowing me to work in his department; but also for being the caring and understanding director he is. I’m also thankful to my supervisors at University of Göttingen, Professor Dr. Christiane Gatz and Professor Dr. Stefan Jakobs for their support and supervision.

Many thanks are to current and former group members of Pecinka and Jiménez-Gómez groups: Dr. Andreas Finke, Björn Pietzenuk, Ganga Jeena, Dr. Inga Schmalenbach, Dr. Marcel von Reth, Mariana Diaz, Mohamed El Hadidi, Navratan Bagwan, Dr. Niels Müller, Dr. Thomas Piofczyk, and Dr. Tripta Jhang for the provided working environment. Special thanks are to Arunkumar Srinivasan for the useful discussions and brain storming, as well as for the bioinformatics consultation. I also thank Barbara Eilts, Regina Gentges and Petra Pecinkova for taking care of plants and technical assistance. Special appreciation to the support and organization provided by the department secretary Birgit Thron. Many thanks are to administration and faculty members of the faculty of Agriculture, Cairo University, with special thanks to members of the department of Genetics, Cairo University for their continuous support.

My success in my studies has been positively influenced by the continuous encouragement and support of Dr. Ahmed Ashoub and his family, to whom I’m deeply grateful. Additionally, I’m thankful to my former teachers and friends who had positive influence on my scientific and personal life, particularly Adham Saleh, Mahmoud Negm, Mohamed Aman, Mohamed Soliman and my dear beloved brother Gaser Abdelsamad.

My love is to my family, specially my parents, whose love and care supported and shaped my life and progress, and to whom I’m totally indebted. My love is also to my wife and mother-in-law who cheered my life. Finally, and before all, all praise is to the God the all-wise, the all-knowing the creator of this wonderful world.

DECLARATION

Herewith I declare that I prepared the PhD thesis entitled “Evolution and epigenetic regulation of RNA-mediated duplicated genes in *Arabidopsis*” on my own and with no other sources and aids than quoted. I also affirm that the facilities of Max Planck Institute for Plant Breeding Research were the only facilities utilized during the practical work.

Göttingen, Germany on 28.05.2015

Signature:

Ahmed M. Abdelsamad

1. INTRODUCTION

Gene duplications are an important factor in genome evolution allowing for functional diversification of genes (Flagel and Wendel, 2009; Innan and Kondrashov, 2010). Duplicated genes are generated by several DNA- and RNA-based mechanisms (Innan and Kondrashov, 2010; Sakai et al., 2011). Whole genome DNA-based duplication (WGD) by polyploidization has occurred in the evolutionary history of all land plants and many animals (De Smet et al., 2013; Dehal and Boore, 2005). Since WGD amplifies the entire genome, it seems to be a solution towards major evolutionary and/or ecological challenges (Comai, 2005; Fawcett et al., 2009). However, WGDs do not alter protein stoichiometry in most cases and therefore, they may be relatively ineffective in situations when an increased amount of a single or few specific proteins is required. In such situation, local DNA and RNA duplication mechanisms may be a better fitting solution. Local DNA duplications amplify individual genes or short chromosomal regions, presumably by an unequal crossing over mechanism (Zhang, 2003). In RNA-based duplication (retroposition), the mature mRNA of a protein-coding gene is reverse transcribed and integrated at ectopic position in the genome using retroviral or retrotransposon machinery (Kaessmann et al., 2009). Therefore, retroposition has a high potential to generate evolutionary innovations, e.g. by expressing genes in a new developmental context, generating chimeric genes with new functional domain combinations or inter-specific horizontal gene transfer (Sakai et al., 2011; Wang et al., 2006; Yoshida et al., 2010).

Gene copies generated through retroposition are called retrogenes, and are distinguished from retrotransposons. Their precursor mRNA molecules are transcribed from non-transposable element protein coding genes (parental genes) that are involved in diverse biological processes (Kaessmann et al., 2009). Consequently, retrogenes are also involved in diverse biological processes and human diseases, such as cancer (Cooke et al., 2014; Hirotsune et al., 2003). Relatively few studies have conducted genome-wide search for retrogenes in plants (Zhang et al., 2005; Wang et al., 2006; Zhu et al., 2009; Sakai et al., 2011). They identified retrogenes to be at most 0.38% of protein coding genes, except for a study in rice (*Oryza sativa*) where low stringency selection criteria were applied (Wang et al., 2006). In humans, 19.1% of all genes were identified as retrogene copies;

however, 82% of those copies contain premature stop codons. Therefore, only 3.4% of all human genes are putatively functional retrogene copies producing functional proteins (Marques et al., 2005; Pennisi, 2012). In rice, transcription was observed for two-thirds of retrogenes, indirectly suggesting that there may be higher proportion of functional retrogenes in plants (Sakai et al., 2011).

Since retroposition duplicates only transcribed regions, it is expected to cause the loss of promoter sequences. This may represent a major bottleneck to retrogene evolutionary success. Recent studies in human and rice suggested retroposition including parental gene promoter (Okamura and Nakai, 2008; Sakai et al., 2011). Additionally, there are multiple possible mechanisms of retrogene promoter acquisition that have been demonstrated in individual examples (Kaessmann et al., 2009). Nevertheless, it is often not clear how frequent those mechanisms are at the genome-wide scale.

Retrogenes expression may be suppressed by epigenetic mechanisms that target transposons and repetitive elements (Vaucheret and Fagard, 2001). Retrogenes are generated by retrotransposon reverse transcriptases and represent duplicated copies; therefore they may become targets of epigenetic transcriptional gene silencing (TGS) by repressive chromatin marks. Chromatin is an indispensable component that provides regulatory and protective function to genetic information (reviewed in e.g. (Li et al., 2007). Transcribed protein coding genes are associated with permissive chromatin marks. In contrast, transcriptionally repressed genes and repetitive elements are typically labeled by histone H3 lysine 27 tri-methylation (H3K27me3), histone H3 lysine 9 di-methylation (H3K9me2) and/or high density DNA methylation in all cytosine sequence contexts in plants (Liu et al., 2010; Roudier et al., 2011; Stroud et al., 2013). While H3K27me3 ensures tissue specific developmental transcription (Lafos et al., 2011), the role of H3K9me2 and promoter DNA methylation is to minimize activities of all kinds of repetitive elements, which frequently includes retrotransposons (Ibarra et al., 2012; Mosher et al., 2009; Slotkin et al., 2009). The association of retrogenes with specific chromatin states has been proposed (Boutanaev et al., 2002; Marques et al., 2005), but only few animal and no plant retrogenes have been characterized as to their chromatin states so far (Monk et al., 2011; Pei et al., 2012).

In flies and mammals, many retrogenes show specific transcription in male germ cells (Bai et al., 2008; Marques et al., 2005; Vinckenbosch et al., 2006). This pattern is intriguing and several explanatory models have been proposed (reviewed in Kaessmann et al., 2009; Kaessmann, 2010). First, it could originate from various chromatin modifications affecting chromosomes and leading to hyper-transcription in meiotic and post-meiotic spermatogenic cells. As a consequence of this global chromatin reorganization induced transcription, some of the testis-transcribed retrogenes could also evolve testis-specific gene functions. The second, not mutually exclusive, hypothesis postulates that retrogenes amplify in the germline tissues and insert preferentially into actively transcribed (open) chromatin. This creates a self-reinforcing loop where the retrogenes insert nearby or into germline transcribed genes and consequently would be also germ-line transcribed. The latter hypothesis is partially supported by observations in Drosophila (Bai et al., 2008), but the tissue-specificity in transcription of plant retrogenes has not been studied.

This study aims to investigate plant retrogenes and their parental genes concerning their abundance, distribution in the genome, expression pattern, relation to transposable elements, epigenetic regulation, emergence rate and evolution. We generated deep sequencing transcriptome data, and used the comprehensive genome and transcriptome resources for the closely related *Arabidopsis thaliana* and *Arabidopsis lyrata* to investigate these open questions. We had manually identified retrogenes in *A. thaliana* genome that were not reported in previous screens (Zhang et al., 2005; Zhu et al., 2009). Therefore, we developed a novel bioinformatic retrogene annotation tool (RAT) to screen both genomes, and initially identified 251 *A. thaliana* retrogenes, 216 of which are novel. We used this set together with the retrogenes found previously (Appendix A) to analyze retrogene and parent-specific features. We show that parents are usually ubiquitously transcribed while retrogenes are mainly low and stage-specific transcribed. Most *A. thaliana* retrogenes acquired novel *cis*-regulatory elements at their integration sites. Importantly, throughout plant development, retrogenes show peak of transcription in pollen. This pattern can also be observed for many lowly transcribed genes genome-wide and resembles retrogene transcription in testis of animals. We found that pollen-specific activation of *A. thaliana* retrogenes is associated with global transcriptional reprograming (Abdelsamad and Pecinka, 2014).

In the second part of this study, we used our enhanced version (v2) of *A. lyrata* genome annotation and identified 168 *A. lyrata* retrogenes representing the first of identified retrogenes in *A. lyrata*. We show that *Arabidopsis* retrogenes emerge in the genome at least ten-times faster than previously calculated (Zhang et al., 2005). Most of the identified retrogenes are transcribed (putatively functional); and targeted by 21nt sRNA molecules, unlike retrotransposons that share the same duplication machinery. Retrogenes tend to acquire introns, which significantly extend retrogene mRNA half-life. We show that not all nascent retrogenes integrate randomly in the genome. Some retrogenes specifically replace their parental genes in a process called retrogene targeting. We developed targeted retrogene annotation tool (TRAT), as an additional tool, to screen the genomes for these cases. Based on current literature, we believe that we report the first natural *in planta* retrogene targeting events.

2. RESULTS

We aimed to study evolution, expression, epigenetic regulation and abundance of retrogenes and their parental genes in plant genomes. We had manually identified retrogenes that were not reported in the previous annotations of retrogenes in *A. thaliana* (Zhang et al., 2005; Zhu et al., 2009). Therefore, we developed a novel bioinformatic retrogene annotation tool to conduct a genome-wide search for retrogene-specific features in *A. thaliana* genome (TAIR10). The identified retrogenes were then studied extensively.

2.1 Annotation of *A. thaliana* retrogenes by a novel retrogene annotation tool (RAT)

We developed a novel bioinformatic Retrogene Annotation Tool (RAT) to conduct a genome-wide screen for retrogenes (Figure 1A). In total 251 retroposition events satisfying stringent quality criteria were annotated in *A. thaliana* genome (Appendix A). Among retrogenes identified in our list, 36 were shared with two previous genome-wide retrogene screens (Zhang et al., 2005; Zhu et al., 2009) and 216 were novel (Figure 1B). The total number of retrogenes identified in all three studies is 309 (291 were considered for downstream analyses; see Appendix A) and corresponds to approximately 1% of *A. thaliana* protein coding genes and pseudogenes (n = 27,416 and 924, respectively).

Generally, retrogenes are intron-less copies of intron-containing paralogous genes. They integrate randomly in the genome; and potentially have downstream poly(A)-tails. The RAT screened for theses retrogene-specific characters. The principal steps in retrogene identification are given in (Figure 1A). First, the paralogy groups between sets of intron-less and intron-containing protein coding genes according to TAIR10 were established using protein homologies in InParanoid Version 4.1 with default parameters (Remm et al., 2001). When the paralogy group had multiple intron-containing 'inparalogs' with ≥ 2 different introns, they were also considered for downstream analysis. Similarly, paralogy groups between pseudogenes and intron-containing protein coding genes were identified as the best reciprocal BLAST hits using cDNA sequences (Altschul et al., 1990; Swarbreck et al., 2008). Accepted retrogene-parent candidate pairs had a minimum homology score 10^{-10} and a minimum difference in intron number of two introns. A single intron difference was only accepted if a poly(A)-tail was detected within 150 or 250 bp downstream of the stop codon of the retrogene candidate with or without annotated 3′ UTR, respectively. Poly(A)-tail was defined as a stretch of consecutive adenines with minimum length of 15 adenine nucleotides, allowing a single mismatch. We determined Poly(A)-tail minimum length as the shortest non-random stretch of (A) nucleotides present in *A. thaliana* genome (materials and methods; Figure 18).

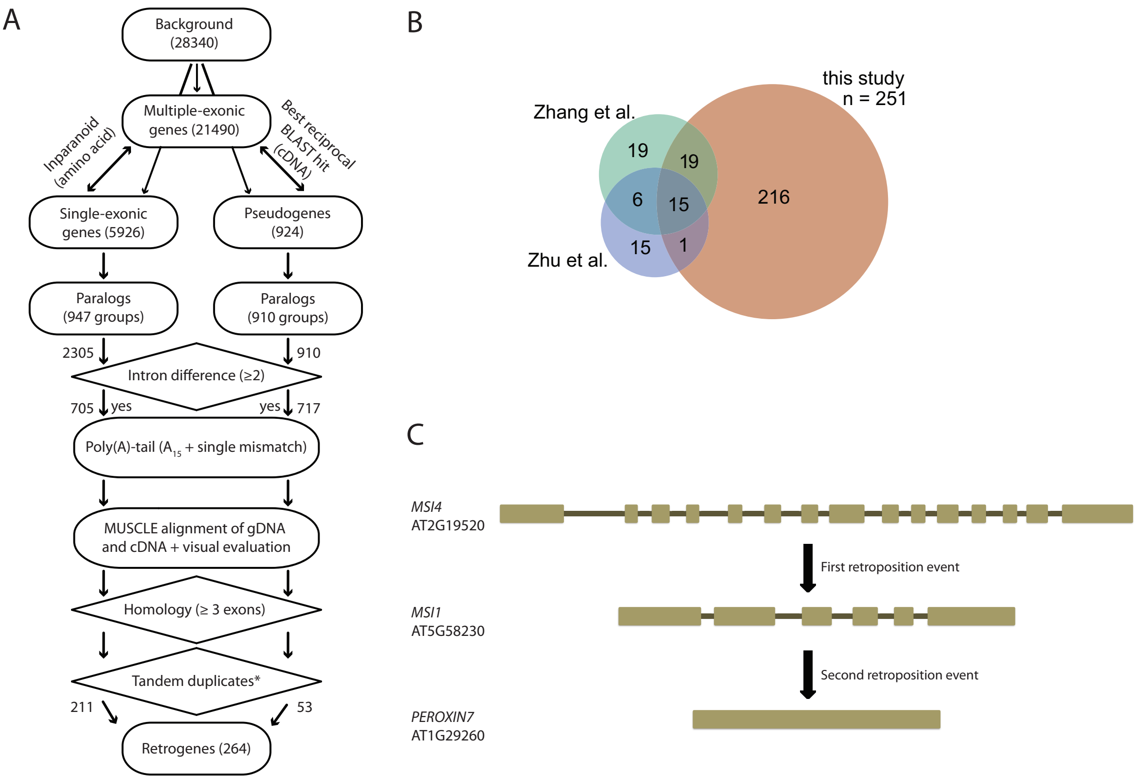

Figure 1. Annotation of *A. thaliana* retrogenes using the RAT

(A) Schematic representation of the retrogene annotation tool. **(B)** Venn diagram indicating the numbers of retrogenes identified in three *A. thaliana* genome-wide searches (Abdelsamad and Pecinka, 2014; Zhang et al., 2005; Zhu et al., 2009). The venn does not include disputable retrogenes from the two previous studies. **(C)** Example of repeated retroposition in *A. thaliana*; the *MSI4 – MSI1 – PEROXIN 7* retroposition series.

Since the absence of introns can also be due to a loss of splicing signals (intron retention), homology of exonic and intronic sequence was visually validated following gDNA and cDNA sequence alignment (Edgar, 2004). A retrogene was accepted when a minimum of three consecutive homologous exons, spanning two lost introns, were observed. If multiple parents were predicted for a retrogene, we accepted the candidate with the highest pairwise alignment score in multiple (cDNA) sequence alignment (Deng et al., 2010; Larkin et al., 2007). When a candidate retrogene overlaps with a list of DNA-based gene duplications (Blanc and Wolfe, 2004), it was excluded. The protocol was executed with customized bioperl and awk scripts (Stajich et al., 2002).

Hence, the RAT identified 251 *A. thaliana* retroposition events; that were used for downstream analysis together with previously identified retrogens.

2.2 *A. thaliana* retrogenes are capable of repeated retroposition and occur in gene-rich genomic regions

The RAT tool combines multiple retrogene searches within intron-less and intronized genes; thus, it allows searching for potential secondary retropositions of retrogene transcripts. This revealed 12 retrogenes that served as templates for another round of retroposition (Figure 1C and Table 1). In these cases, the primary parent gave rise to the primary retrogene, whose mRNA served as the precursor for the secondary retrogene. The model where the primary parent gives rise directly to the secondary retrogene was not supported by the order of protein homologies, and that suggests retroposition of the retrogene transcript. Hence, 4.3% of *A. thaliana* retrogenes underwent repeated retroposition without losing their protein coding potential. In addition, we identified multi-retrogene parents. In total, 22 parents gave rise to 54 retrocopies (17 × 2; 3 × 3; 1 × 4; 1 × 7) and a maximum of seven retrocopies derived from a single parent (Appendix A). The observed frequency of multiple retropositions from the same gene is significantly higher than expected at random (Mann-Whitney-Wilcoxon (MWW) test, $P < 2.2 \times 10^{-16}$) strongly arguing that the selection of parental mRNA is not random in at least some cases.

The machinery that transposes retrogenes and retrotransposons often integrate the later at hereochromatic regions (Tsukahara et al., 2012). To explore whether retroposition of retrogenes occurs at specific genomic regions, we plotted densities of all protein coding genes, transposable elements (TEs), parents and retrogenes over the five *A. thaliana* chromosomes (Figure 2A). In agreement with published data (Arabidopsis Genome Initiative, 2000), TEs were enriched in pericentromeric regions and depleted from chromosome arms, while protein-coding genes showed the opposite pattern. Both retrogenes and parents had a profile similar to that of protein coding genes, showing that they occur preferentially in gene-rich genomic regions (Figure 2A).

Table 1. Repeated retroposition events in *A. thaliana*

Parent		Retroposition 1		Retroposition 2	
Gene ID	Introns	Gene ID	Introns	Gene ID	Introns
AT1G08320	12	*AT1G77920*	8	*AT1G58330*	0
AT1G58520	16	*AT1G32090*	10	*AT1G30360*	5
AT2G19520	14	*AT5G58230*	5	*AT1G29260*	0
AT2G28830	6	*AT3G46510*	3	*AT1G29340*	0
AT3G09100	16	*AT5G01290*	14	*AT5G28210*	0
AT3G09810	6	*AT4G35260*	3	*AT1G32480*	0
AT3G24430	13	*AT4G19540*	7	*AT5G50960*	2
AT4G34480	5	*AT5G24318*	3	*AT3G55430*	1
AT4G40040	4	*AT5G10980*	2	*AT5G10400*	0
AT5G28340	5	*AT3G60960*	2	*AT3G60980*	0
AT5G56890	13	*AT1G70460*	7	*AT3G55950*	0
AT5G67320	13	*AT2G26060*	9	*AT1G24530*	0

We showed that retrogenes integrate preferentially in chromosome arms; however, they may still integrate nearby local TEs. To test for association of retrogenes and/or parents with TEs at local scale, we estimated the frequency of all genes with TEs in 1 kbp intervals up- and down-stream of gene transcription start and termination sites (TSS and TTS, respectively). On average, there were fewer TEs upstream than downstream of genes. The frequency of TEs in TSS-upstream regions for all protein-coding genes and retrogenes (17% and 22%, respectively) was not significantly different (Figure 2B). In contrast, parental genes with TEs in the first two kbp upstream of the TSS were scarce relative to the whole genome (chi-square test, $P < 0.05$). Similarly, 25% of all genes and retrogenes contained TEs in the first two kbp of the TTS-downstream region, while it was only 17% for parents (chi-square test: $P < 0.05$ in the first kbp). This shows that retrogenes are not enriched for close-lying TEs compared to the genomic average, but parents are depleted of TEs in both up- and down-stream intergenic regions.

Hence, the *A. thaliana* genome contains at least 291 retrogenes located predominantly in gene-rich chromosomal regions. About 10% of the parents gave rise to multiple retrogenes and approximately 4.3% of the retrogenes underwent a second retroposition.

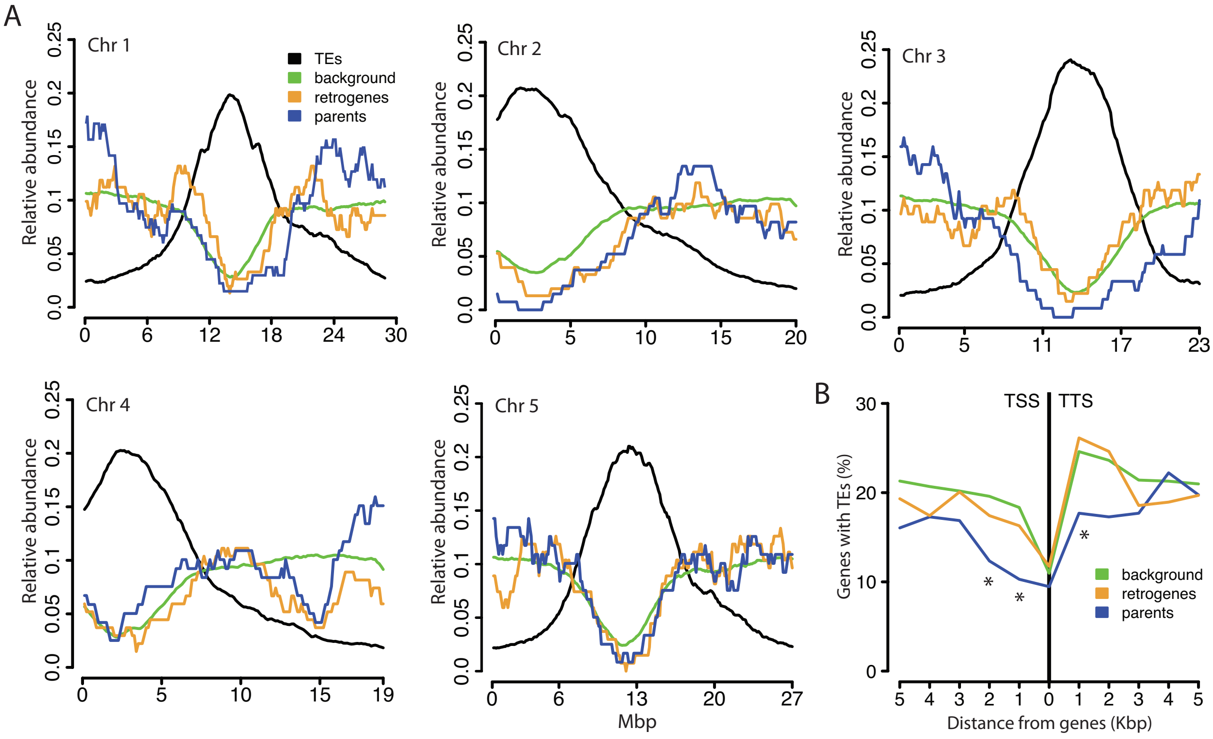

Figure 2. Genomic features of *A. thaliana* retrogenes

(A) Relative abundance (*y*-axis) of transposable elements (TEs, black), all genes and pseudogenes (background, green), retrogenes (red) and parents (blue) over the five *A. thaliana* chromosomes (*x*-axis). **(B)** Percentage of genes containing TEs (*y*-axis) in 1 kbp intervals from the gene transcription start and termination sites (TSS and TTS, respectively) for all protein coding genes (background, green), retrogenes (red) and parents (blue). Significant differences (P < 0.05) in chi-square test relative to background are indicated by asterisk.

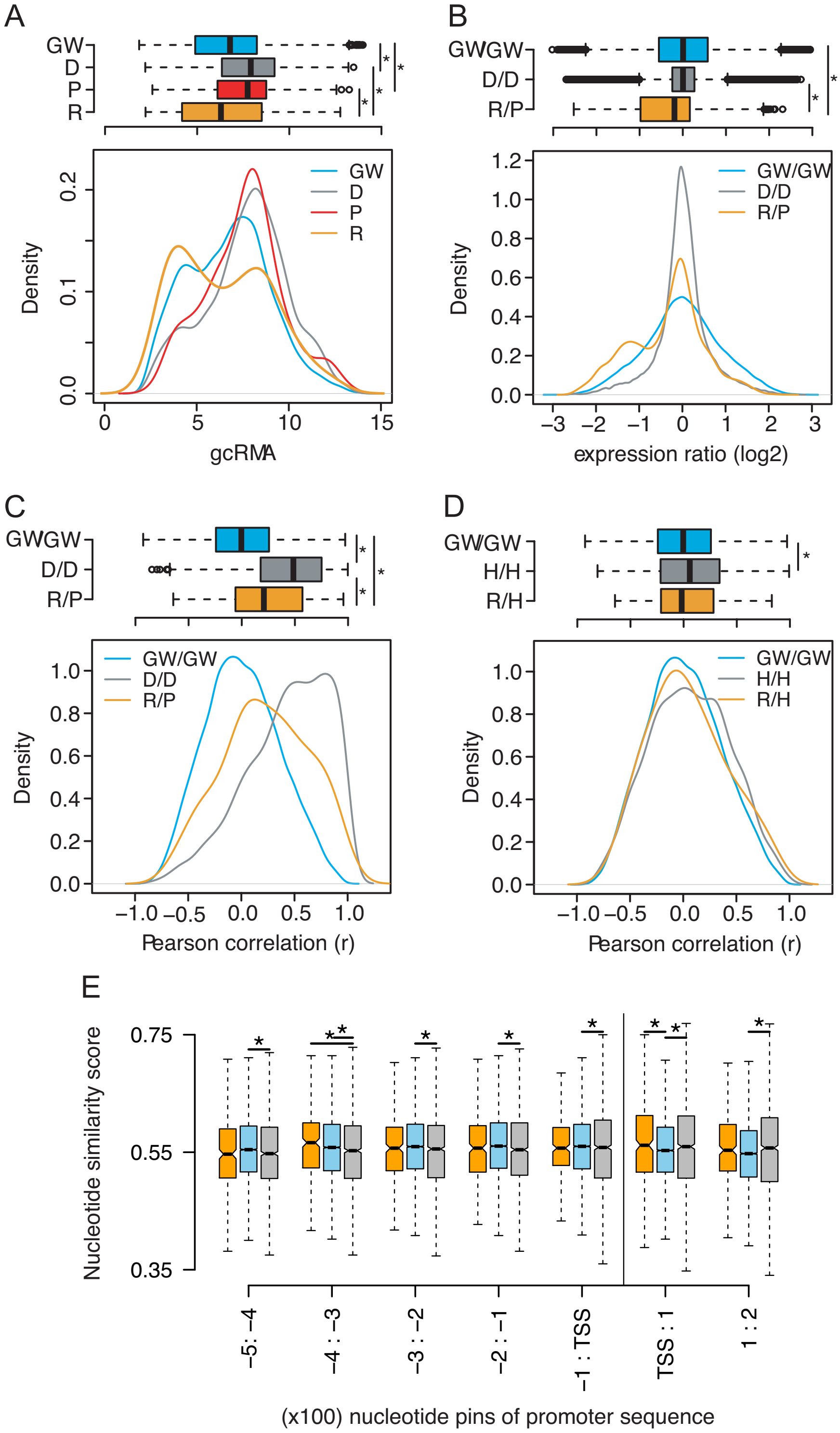

Figure 3. Retrogenes are driven by novel promoters

(A) Box and density plots of $\log_2$ robust microarray averaging (gcRMA) values for genome-wide genes (GW), DNA duplicated genes (D), parents (P) and retrogenes (R) over the 49 *A. thaliana* developmental stages. **(B)** $\log_2$ transcription ratios of the random genome-wide gene pairs (GW/GW), DNA duplicated pairs (D/D) and retrogene/parent pairs (R/P). **(C, D)** Pearson correlation of gene co-transcription between random genome-wide gene pairs (GW/GW), DNA duplicated pairs (D/D), retrogene-parent pairs (R/P), genome-wide head-to-head oriented genes (H/H) and retrogene-head-to-head oriented neighboring genes (R/H) in 49 developmental stages. **(E)** Box plots of nucleotide similarity score for (x100) nucleotide-long pins of promoter sequences. Nucleotide similarity scores for retrogene-parent promoters (orange) are not significantly different from for random gene pairs (sky blue), but usually less than for DNA duplicated gene pairs (grey). Non-significant ($P \geq 0.05$) relationships are not shown.

2.3 Retrogenes are derived from highly transcribed parental genes and are transcribed preferentially by novel promoters

The cDNA origin of retrogenes implies their retroposition without their regulatory sequences (promoters). However, the majority of them retains intact open reading frames (ORFs) and is transcribed. We took advantage of the comprehensive retrogene list assembled in this study (Appendix A) (Abdelsamad and Pecinka, 2014) and explored the patterns of retrogene transcription in *A. thaliana*. The mRNA accumulation was analyzed using microarray data from the 49 *A. thaliana* developmental stages assembled by the AtGenExpress consortium (Schmid et al., 2005) and validated for selected tissues by RNA-sequencing (Loraine et al., 2013). In total, 209 retrogenes and 245 parents are present on the ATH1 cDNA microarray. To compare the effects of RNA- and DNA-based duplications, we also analyzed the set of 3,088 *A. thaliana* DNA duplicated genes (Blanc and Wolfe, 2004). Plotting the mean $\log_2$ Robust Multi-array Averaging (gcRMA; (Irizarry et al., 2003) values of all ATH1 probesets (n = 22,746) revealed a double-peak distribution with the left peak representing genes with poor mRNA levels and/or background signals (Figure 3A). The gcRMA values of some retrogenes and parents overlapped with this region and suggested that some of the candidates may not be transcribed in any of the 49 stages. Therefore, we kept only the genes with gcRMA values of 5 or higher in at least one developmental stage (transcribed genes). In total, 89.4% (n = 20,398) of all genes, 85.2% (n = 178) of retrogenes, 94.7% (n = 232) of parents and 99.3% (n = 3067) of DNA duplicated genes passed these criteria (Figure 3A). This shows that the majority of *A. thaliana* retrogenes are transcribed in at least some developmental stages and their mean gcRMA values did not differ significantly from the genome-wide gene set (MWW test, $P = 0.48$; Figure 3A). The parents were significantly enriched for highly transcribed genes relative to both retrogenes and the whole-genome set (MWW test, $P = 7.64 \times 10^{-06}$ and $P = 1.86 \times 10^{-11}$, respectively; Figure

3A). Similarly, DNA duplicated genes were strongly transcribed and therefore similar to parents, but strongly different from retrogenes (MWW test, $P = 0.16$ and $P = 1.56 \times 10^{-10}$, respectively).

To reveal the transcription relationships between individual retrogene–parent pairs, we compared their developmental stage-specific gcRMA ratios with the transcription of 5,000 randomly selected gene-pairs and the 1,527 DNA duplicated gene-pairs (Figure 3B). Transcript accumulation ratios of random pairs and DNA duplicated genes represented a broad and narrow range of normally distributed values (MWW test, $P = 0.85$). Although many retrogenes have a comparable degree of transcription relative to their parents, there is a specific group of two-to-three-fold less transcribed retrogenes making retrogene–parent pairs significantly different from both the random gene set and DNA duplicated genes (MWW test, both comparisons $P < 2.2 \times 10^{-16}$; Figure 3B). Inspecting the gcRMA values over individual developmental stages for the low-transcribed group revealed that these retrogenes were transcribed above the threshold (gcRMA ≥ 5) in only one or few tissues while their parents frequently showed ubiquitous transcription.

A recent study in rice suggested frequent co-transcription between retrogenes and parents in plants (Sakai et al., 2011). Our retrogene identification criteria and the nature of *A. thaliana* retrogenes (e.g. an absence of retrogenes residing in the introns of other genes) allowed testing three possible mechanisms of retrogene *cis*-regulatory element origin: 1) carry-over of parental promoters, 2) the use of bi-directional promoters at integration sites, and 3) an acquisition of novel *cis*-regulatory elements. First, we tested whether the *A. thaliana* retrogenes inherit the parental transcription pattern. We calculated co- transcription of retrogene–parent pairs as Pearson product-moment correlation coefficients (*r*) across the 49 developmental samples of the AtGenExpress dataset. Indeed, co-transcription in the set of retrogene–parent pairs (n = 179) was significantly higher than in the 20,000 randomly selected gene pairs (MWW test, $P = 2.30 \times 10^{-6}$; Figure 3C). We calculated the frequencies of genes per 0.1 *r* correlation bins for retrogenes and genome background and used this to calculate the number of highly co-transcribed retrogene–parent pairs. In total, 25% of the retrogene–parent pairs (26 out of 102) were correlated more than random gene pairs. However, the co-transcription of DNA duplicated gene pairs, calculated in the same way, was more prominent (MWW test,

$P < 2.2 \times 10^{-16}$; Figure 3C) and 45.6% of them surpassed the random-pairs background.

Second, we tested the possibility for retrogene transcription by bi-directional promoters of head-to-head ("head") oriented neighboring genes. The Pearson correlations of random transcribed gene-pairs (n = 20,000) and the genome-wide set of transcribed "head" oriented genes (n = 2,087) revealed an infrequent but consistent co- transcription between head-oriented gene pairs (MWW test, $P = 2.705 \times 10^{-10}$; Figure 3D). This shows that sharing bi-directional *cis*-elements is not common in *A. thaliana*. Retrogene–head oriented neighbor pairs (n = 63) displayed an intermediate pattern that was not significantly different from either genome-wide or head oriented genes (MWW test, both $P = 0.60$; Figure 3D). Only 2.5% of head oriented retrogenes had higher correlation than random pairs, illustrating negligible effect of promoter sharing (Figure 3D). Consequently, retrogenes seemed to acquire novel *cis*-regulatory sequences at their integration sites. The low nucleotide similarity scores between retrogenes and parental gene promoters supported this hypothesis; that were not significantly different from scores for random gene pairs (GW) but significantly less than for DNA duplicated gene pairs (Figure 3E).

Hence, retrogenes show low transcription, while their parents show high and ubiquitous transcription. The transcription of most of the retrogene–parent pairs is not correlated, due to acquisition of novel regulatory elements at retrogene integration sites.

2.4 *A. thaliana* retrogenes are transcribed in male gametes

In insects and animals, retrogenes show preferential transcription in male germ cells (Kaessmann, 2010). To analyze developmental regulation of *A. thaliana* retrogene transcription, we plotted the mean gcRMA values of genome-wide, parent and retrogene sets for each of the 49 analyzed developmental stages (Figure 4A). The average mRNA level of parents was higher than that of retrogenes and the genome-wide gene set in all stages. The mean transcription per group was relatively constant, except for pollen where there was a dip in transcription in the parents and the genome-wide set that was contrasted with a peak of retrogene transcription (Figure 4A). To identify relationships between developmental stages and retrogenes, we hierarchically clustered both groups and expressed the result as a heat-map of the retrogene transcription z-scores (Figure 4B). This separated stamen and pollen from the rest of the tissues. The highest frequency of retrogenes with positive z-scores (>0) was then found in pollen and seeds (62% and 63%, respectively; Figure 4C). However, with more stringent criteria (z-scores >1 and >3), the pollen peak became more prominent relative to other tissues and corresponded to 50% and 30% of retrogenes, respectively (Figure 4C). This shows that many retrogenes reach their transcription maxima in pollen. The pollen-specific transcription pattern has been confirmed by analysis of individual cases (Figure 4D, Figure 5A).

However, plotting the transcription quantiles (Appendix C) of retrogene $\log_2$ gcRMA revealed that not all retrogenes followed this simple trend; and the retrogenes with a negative z-score (pollen down-regulated) usually derived from the group of developmentally highly transcribed genes (Figure 4E, bottom). Remarkably, this distribution also held true for the genome-wide gene set (Figure 4D, top). The parents and the DNA duplicated genes showed more prominent down-regulation of the highly transcribed genes (quantile 4) and less obvious up-regulation of lowly transcribed genes (quantile 1), while TEs showed up-regulation for all quantiles (Figure 5B). Hence, we found a pollen specific activation of retrogenes that is a part of the global pollen-specific transcriptional reprogramming.

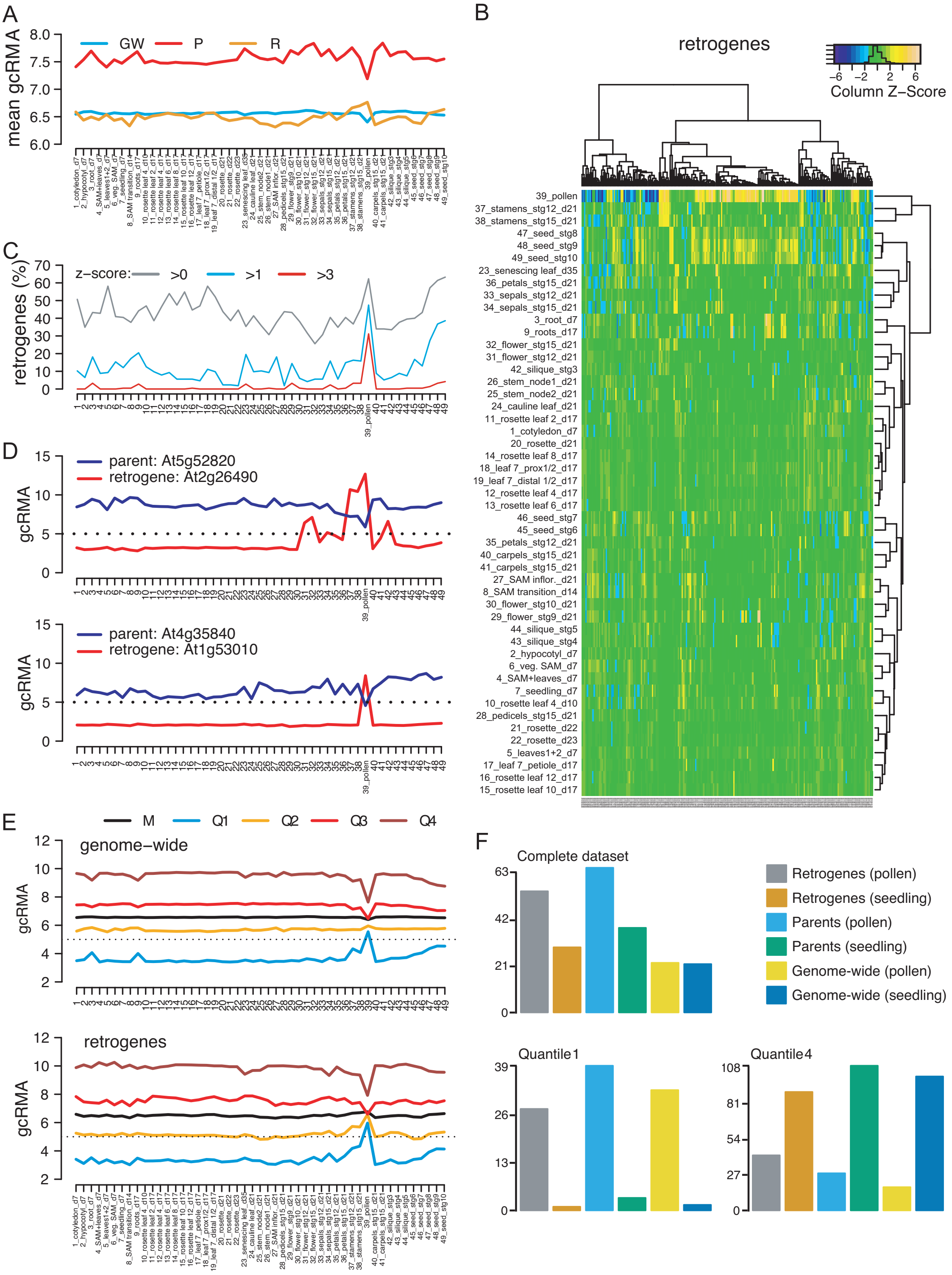

Figure 4. Retrogenes are transcriptionally up-regulated in pollen

(A) The mean $\log_2$ robust microarray averaging (gcRMA) values for genome-wide genes (GW), parents (P) and retrogenes (R) at each of the 49 *A. thaliana* developmental stages. **(B)** Hierarchically clustered heat map of retrogene z-scores (*y*-axis) and developmental stages (*x*-axis). **(C)** The frequency of retrogenes with row z-scores in (**B**) >0, >1 and >3 in individual developmental stages. **(D)** Examples of retrogenes and parents showing tissue-specific and ubiquitous transcription, respectively, with major transcription changes in pollen (stage 39). **(E)** Developmental gcRMA values for genome-wide set of genes and retrogenes.

Transcription is shown for mean (M) and transcription quantiles: low-transcribed/quantile 1 (Q1), mid-low-transcribed/quantile 2 (Q2), mid-high-transcribed/quantile 3 (Q3) and high-transcribed/quantile 4 (Q4). **(F)** Mean RNA-sequencing RPKM values (*y*-axis) for all genes (Genome-wide), parents and retrogenes in vegetative rosettes and pollen as complete datasets, quantile 1 (lowly transcribed genes) and quantile 4 (highly transcribed genes).

Pollen development includes several stages (Honys and Twell, 2003). To find out whether retrogenes are transcribed in specific pollen developmental stages, we compared their transcription in the three final developmental stages; unicellular microspores, bicellular pollen, tricellular pollen and two highly correlated (*r* = 0.92) samples of mature pollen grains (Honys and Twell, 2004; Schmid et al., 2005). This revealed continuous increase of mean retrogene transcription throughout pollen development that contrasted with down-regulation of parental genes in tri-cellular pollen and mature pollen grains (Figure 5C). There are two distinct cell types in mature pollen: vegetative and sperm cells. Thus, we investigated whether there is enrichment for retrogene transcripts in vegetative and sperm cells (Honys and Twell, 2003). We used TEs as the control for vegetative cell specific transcription based on the recently proposed model (Slotkin et al., 2009). Although we observed strong TE up-regulation in pollen relative to leaves (MWW test, $P < 2.2 \times 10^{-16}$), there was a significantly higher amount of TE transcripts in sperm cells relative to the entire pollen (MWW test, *P* = 0.013; Figure 5C). This indicates that there is a higher amount of TE transcripts in both pollen cell types. The parents were significantly more transcribed in sperm cells relative to seedlings (MWW test, *P* = 0.001) and were underrepresented for the low transcribed genes in this tissue relative to entire pollen (Figure 5C). Therefore, retrogene parents are transcribed preferentially in sperm cells. The median of retrogene transcription was higher than that of TEs and increased in both pollen samples relative to seedlings, but only the entire pollen differed significantly (MWW test, *P* = 0.008; Figure 4E). In combination with pollen developmental series data, this shows that retrogenes are transcribed in both pollen cell types.

In order to validate our results by independent experiment, we tested whether our findings hold true in datasets generated by RNA-sequencing. Gene transcription in mature pollen grains was compared with that in seedling tissues (Loraine et al., 2013). Plotting the mean RPKM (reads per one kilobase per one million reads) values for entire set, quantile 1 (lowest transcribed) and quantile 4 (highest transcribed) of all genes, retrogenes and parents confirmed microarray data (Figures

4A, E, F; Figure 5B). The only exception was higher transcription of parents in pollen relative to seedlings in RNA-sequencing (mean and quantile 1 samples; absent in quantile 4 sample) while this was opposite in microarrays (Figure 4A,F). This difference can be attributed to higher sensitivity of RNA-sequencing technology to quantify transcripts from low transcribed genes (Mooney et al., 2013; Zhao et al., 2014). This partially applies also to retrogenes as the up-regulation in pollen versus seedling is more pronounced in RNA-seq compared to microarrays (Figure 4F). From this we conclude that retrogene activation starts prior to pollen maturation and later occurs in both terminal pollen cell types (vegetative and sperm cells).

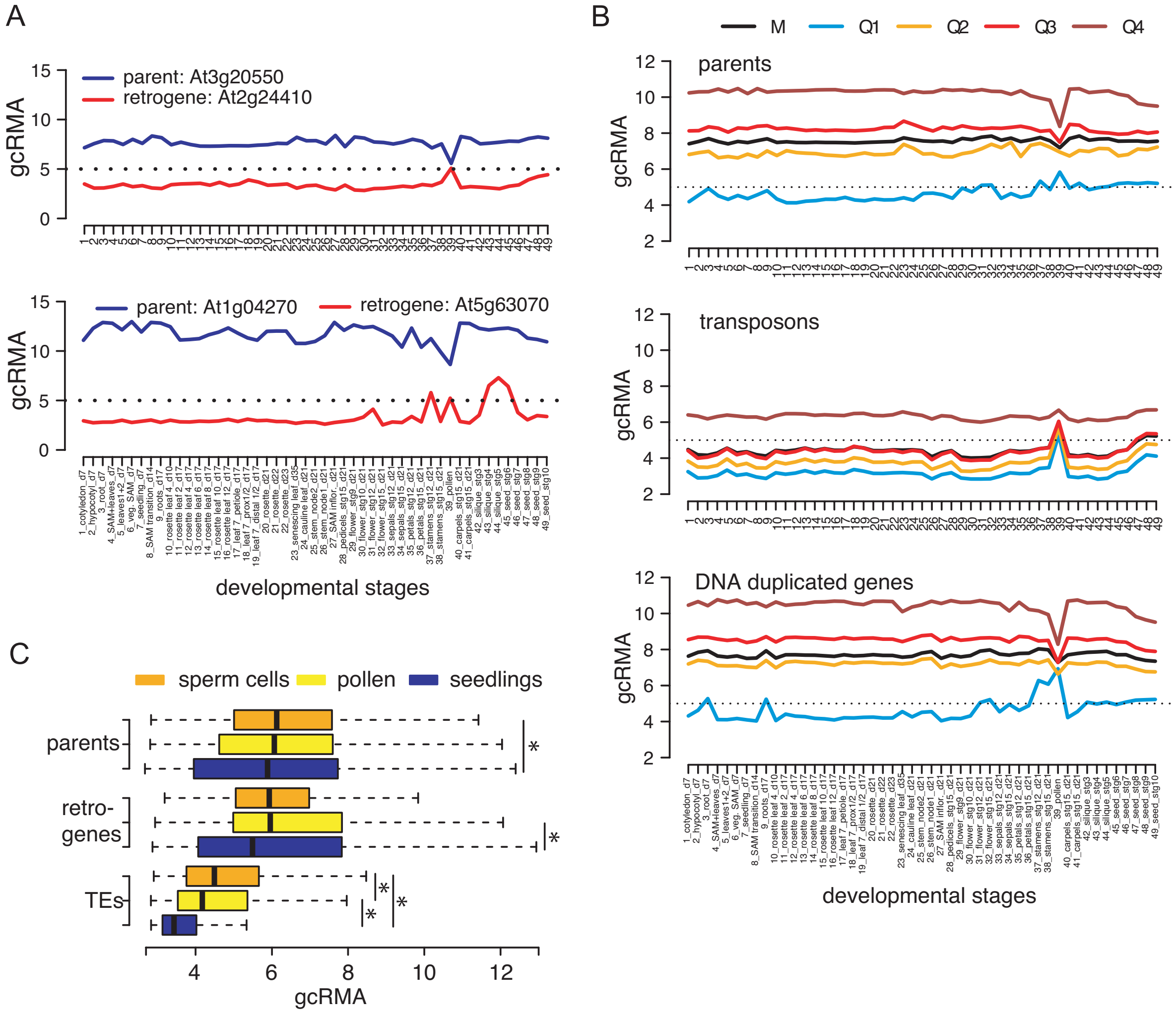

Figure 5. *Arabidopsis* retrogenes are expressed in pollen

(A,B) Log$_2$ robust microarray averaging (gcRMA) values (*y*-axis) of specific groups of genes in 49 *A. thaliana* developmental stages and tissues (*x*-axis). The horizontal dashed line (gcRMA = 5) indicates the threshold of high expression. **(A)** Representative examples of retrogene-parent pairs with ubiquitously expressed parents and tissue-specifically expressed retrogenes. **(B)**. gcRMA values for parents (top), transposons (middle) and DNA duplicated genes (bottom) shown as the mean (M) and expression quantiles from low-expressed (Q1) to high expressed (Q4). **(C)** gcRMA expression values for parents, retrogenes and transposons (TEs) in pollen sperm, entire pollen (sperm cells and vegetative cells) and seedlings. Asterisks show significant differences (P < 0.05) in Mann-Whitney-Wilcoxon test.

2.5 Retrogenes are deficient for transcription-permissive chromatin marks in leaf tissues

Analysis of transcription quantiles suggests that global transcriptional changes in pollen have a major effect on retrogene transcription. This may be achieved by a global chromatin reprogramming (Kaessmann et al., 2009). Therefore, we calculated $\log_2$ fold transcription changes between pollen and 21 day-old rosettes (ATGE_73/ATGE_22; Schmid et al., 2005) and correlated those with transcriptional changes induced by chromatin mutants (mutant rosettes/wild type rosettes). Five groups were compared: all genes (n = 22,746), pollen up-regulated genes (n = 5,171), leaf up-regulated genes (n = 6,057), pollen up-regulated retrogenes (n = 51) and leaf up-regulated retrogenes (n = 53). Tissue up-regulated genes were defined as having $\log_2$-fold change ≥ 1 in one versus the other tissue. First we estimated the effects of the transposon silencing machinery by testing mutants for *DECREASED DNA METHYLATION 1* (*DDM1*), *KRYPTONITE* (*KYP*) and *HISTONE DEACETYLASE 6* (*HDA6*) (Baubec et al., 2010; Inagaki et al., 2010; Popova et al., 2013), which lead to loss of repressive DNA methylation and H3K9me2; and gain of permissive histone-acetylation at heterochromatic loci, respectively. There was no clear correlation (maximum r = 0.040) between transcription in pollen relative to leaves and transcriptional changes induced by *ddm1*, *kyp* and *hda6* for all tested groups (Figures 6A-C). This demonstrates that TE silencing components do not determine the global gene transcription pattern in pollen nor affect retrogenes.

Recently, a connection between pollen-specific genes and H3K27 methylation has been reported in *Arabidopsis* (Hoffmann and Palmgren, 2013). Therefore, we tested the effects of the histone H3K27me3 mark by analyzing mutants of the polycomb group repressive complex factors *CURLY LEAF* (*CLF*) and *SWINGER* (*SWN*) that have been shown to control transcription during development (Farrona et al., 2011; Lafos et al., 2011). The correlation between *clf* and *swn* single mutants, with pollen-specific transcriptional changes was low (r < 0.20; Figures 6D,E). Because CLF and SWN are partially functionally redundant (Lafos et al., 2011), we tested for effects in the *clf*/*swn* double mutant. The correlation between pollen and *clf*/*swn* transcription profiles for the set of all genes was higher (r = 0.277) than for *clf* and *swn* single mutants (Figure 7A; Figure 6D,E). Surprisingly, the high correlation was mainly due to leaf up-regulated genes and retrogenes (r = 0.469 and 0.364, respectively) that were coordinately down-regulated in both pollen and *clf*/*swn* double mutant (Figure 7A). In contrast, pollen up-regulated genes showed generally uncorrelated transcription with *clf*/*swn* (r = -0.047).

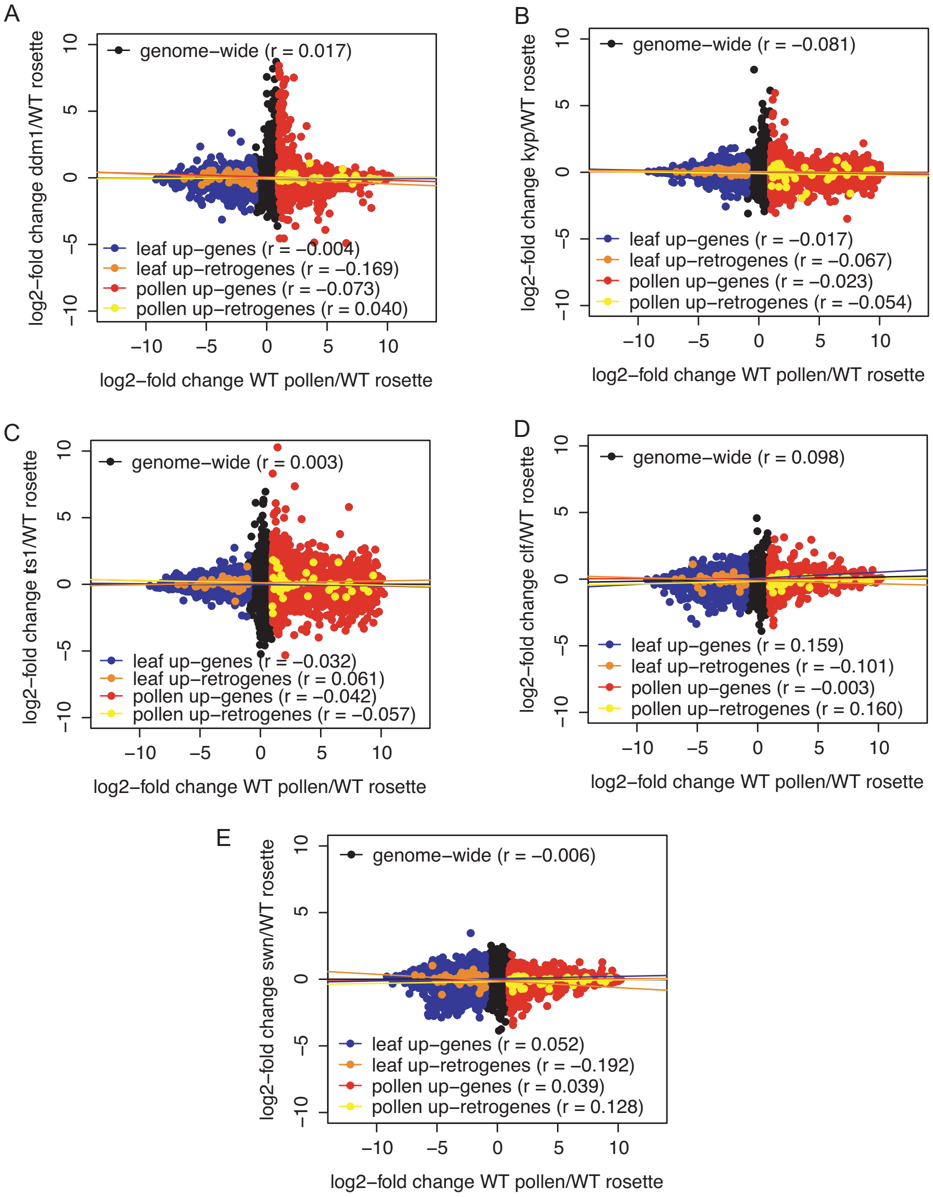

Figure 6. Expression correlations between pollen and chromatin mutants

(A-E) Dot plots of microarray based $\log_2$-fold-changes in wild type pollen (ATGE_73)/rosettes (ATGE_22) (*x*-axis) versus mutant rosettes/wild type rosettes (*y*-axis). Specific gene sets were superimposed on the genome-wide gene set. The lines indicate expression correlation (r) between the *x*- and the *y*-axis gene sets. **(A)** shows comparison of pollen with *ddm1*, **(B)** with *kyp*, **(C)** with *HDA6* mutant allele *rts1-1*, **(D)** with *clf* and **(E)** with *swn*.

To further test the connection between pollen-specific transcription and H3K27me3 changes, we analyzed transcription in a mutant for *FERTILIZATION INDEPENDENT ENDOSPERM* (*FIE*), another key gene of the Polycomb repressive complex (Bouyer et al., 2011). Although the correlations between *fie* and pollen transcription profiles were weaker (*r* = 0.186, 0.366 and 0.268 for all genes, leaf up-regulated genes and retrogenes, respectively; Figure 7B), they perfectly recapitulated trends observed in the comparison between *clf*/*swn* and pollen. Hence, loss of key components of the Polycomb repressive complex correlates with pollen-specific gene down-regulation of leaf-transcribed genes but does not explain pollen-specific gene up-regulation.

To identify chromatin modification(s) associated with retrogenes and pollen-up-regulated genes in somatic tissues, we used publicly available chromatin data from young *A. thaliana* leaves (Roudier et al., 2011). We extracted information on chromatin marks for every gene and compared the full sets of retrogenes, parents and all genes (Figure 7C). In accordance with high and ubiquitous transcription, the parents were enriched for permissive chromatin marks histone H3 lysine 4 di- and tri-methylation (H3K4me2 and me3), histone H3 lysine K36 tri-methylation (H3K36me3) and histone H2B ubiquitination (H2Bub), followed by retrogenes and the genome-wide set. None of these groups was enriched for the repressive H3K27 modifications. The enrichment for gene body DNA methylation in highly expressing genes is consistent with the currently proposed function of this modification (Coleman-Derr and Zilberman, 2012).

The next step was to compare the pattern of chromatin marks distribution for pollen up-regulated and leaf up-regulated genes (Appendix B). The distribution pattern of chromatin marks for each individual group (retrogenes, parents and all genes) was relatively similar (Figure 7D, Figures 8A,B). There were no changes in gene body DNA methylation. While the H3K27 modifications were enriched in pollen up-regulated genes of the genome-wide set, this mark does not seem to play a major role in somatic silencing of pollen up-regulated retrogenes (Figure 8A). In contrast, all analyzed transcription-permissive marks (H3K4me2 and me3, H2Bub and H3K36me3) were underrepresented in pollen up-regulated genes in leaf tissues (Figure 7D, Figures 8A,B). The presence or the absence of these marks was strongly correlated in pair-wise comparisons of individual modifications (Figure 7E, Figure 8C).

This suggests that in leaf tissues, retrogenes and other pollen up-regulated genes are depleted of permissive chromatin marks without enrichment for repressive marks. In contrast, leaf up-regulated genes are down regulated in pollen by a mechanism involving the Polycomb repressive complex components *CLF*, *SWN* and *FIE*.

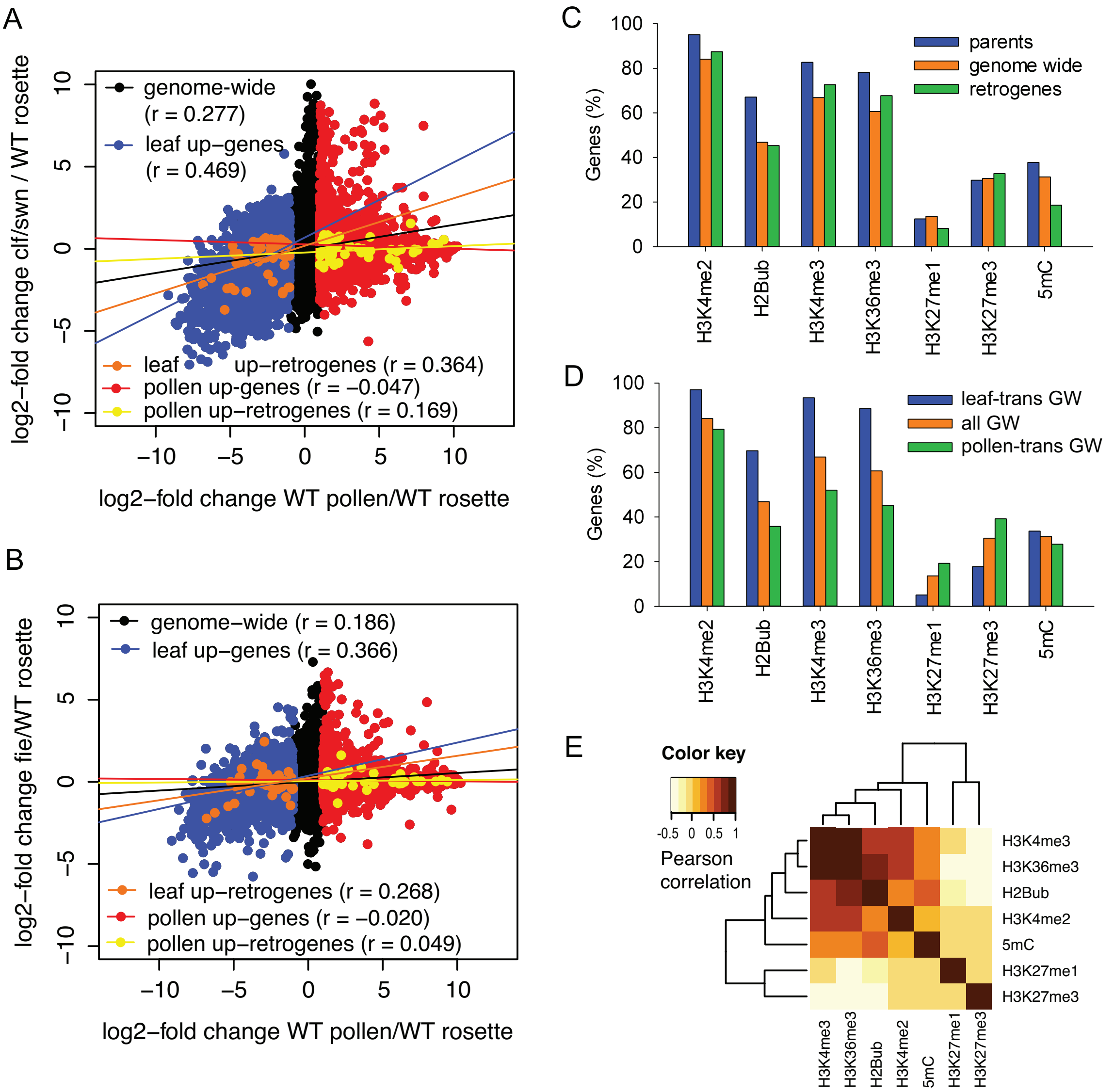

Figure 7. Chromatin control of pollen-specific gene transcription

(A-B) Dot plots of $\log_2$-fold changes in wild type pollen/rosettes (*x*-axis) and **(A)** *clf/swn* double mutants or **(B)** *fie*/wild type rosettes (*y*-axis). Specific gene sets were superimposed on the genome-wide set in different colors. Lines indicate transcription correlation (r) between the *x*- and the *y*-axis for specific gene sets. The r values are given in parentheses. **(C)** The frequency of seven chromatin modifications at gene coding sequences for GW, P and R in young leaf tissues. **(D)** The same as (**C**) for all genes (all GW), leaf-transcribed genes (leaf-trans GW) and pollen-transcribed genes (pollen-trans GW). **(E)** Hierarchical clustering and heat map of Pearson correlation values of co-localization between 7 chromatin modifications for all *A. thaliana* genes.

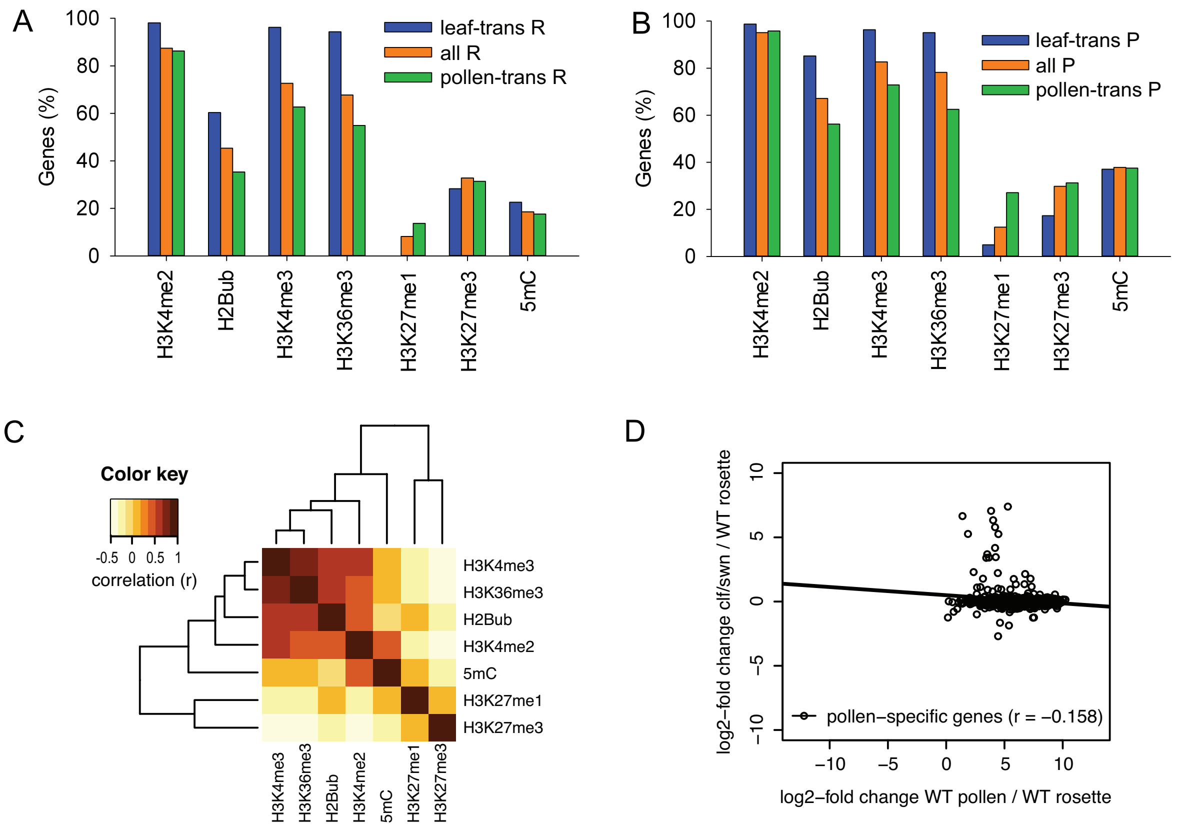

Figure 8. Chromatin control of pollen-specific gene expression

(A) The frequency of seven chromatin modifications at protein coding regions for all retrogenes (all R), leaf-expressed retrogenes (leaf-expr R) and pollen-expressed retroges (pollen-expr R) in young leaf tissues. **(B)** Shows the same as **(A)** but for parents. **(C)** Hierarchical clustering and heat map of Pearson correlations between seven analyzed chromatin modifications for all retrogenes. **(D)** Dot plot of $\log_2$-fold-changes in wild type pollen (ATGE73)/rosettes (ATGE_22) (*x*-axis) versus *clf/swn* doublemutant/wild type rosettes (*y*-axis) for the set of 584 pollen-specific genes defined by Hoffmann and Palmgren, 2013. The black line indicates expression correlation (*r*) between the *x*- and the *y*-axis datasets.

2.6 Gain of transcription factor binding sites facilitates *PCR11* retrogene sperm-specific transcription

Retrogenes showed gamete-specific transcriptional activation. Therefore, we tested whether gamete-specific transcription of retrogenes has evolved into gamete-specific developmental functions. Five retrogenes found in our screen *MULTICOPY SUPRESSOR OF IRA1* (*MSI1*), *PLANT CADMIUM RESISTANCE 11* (*PCR11*), *BETA GLUCOSIDASE 14* (*BGLU14*), *MATERNAL EFFECT EMBRYO ARREST 25* (*MEE25*) and *PEROXIDASE* are associated with pollen development, sperm cell differentiation, pollen tube growth and development (TAIR10). To test whether these retrogenes have evolved parent-independent gamete-specific expression and function, we investigated the relationship between transcription of these retrogenes and their parents. We plotted their mean developmental gcRMA values and calculated transcription Pearson correlations (Figs 9A-D). The parental gene of the *PEROXIDASE* was not included on the ATH1 array and therefore we did not continue its analysis. The transcription of *MSI1* was strongly correlated (*r* = 0.905) with its parent *MSI4* and both were ubiquitously transcribed throughout development (Figure 9A). *BGLU14* and its parent *BGLU15* were both up-regulated in pollen (Figure 9B). The *MEE25* retrogene was lowly transcribed during the entire development and higher transcription was found only in embryonic tissues (Figure 9C). However, its parent, At4g10960, was transcribed mainly in floral tissues, seeds and pollen where it greatly surpassed *MEE25* transcription. Hence, these three retrogenes did not provide evidence for development of parent-independent pollen-specific transcription. In contrast, *PCR11* was lowly transcribed almost throughout entire development, but activated in floral tissues, stamen and pollen. This pattern was opposite to that of its parent, *PCR2*, which was active mainly in the photosynthetically active tissues and down regulated in stamen and pollen (Figure 9D). The *PCR11* gene is transcribed specifically in pollen sperm cells by the MYB-transcription factor DUO1 (Borg et al., 2011). Therefore, we compared promoter regions of *PCR11* and *PCR2* and looked for previously described DUO1 binding motifs (Borg et al., 2011). There are three binding regions in the 500 bp region upstream of the *PCR11* TSS (TAACCGTC at −47 to −54 bp and AAACCG at −153 to −158 and −452 to −457 bp). However, only a single DUO1 binding motif (AAACCGT at −100 to −106 bp from the TSS) is found in the promoter of *PCR2*. To test whether this represents gain of function in *PCR11* or loss of function in *PCR2*, we compared promoter regions of several other *PCR* family members representing both the *PCR2* clade (*PCR1*, *PCR3*) and the out-groups (*PCR4*, *PCR8*, *PCR10*) (Song et al., 2010). None of these genes contained a single DUO1 binding motif in the 500 bp region upstream of the TSS. Furthermore, comparing their transcript levels revealed that only *PCR11* is significantly up regulated in pollen relative to *PCR2* (Figure 9E).

To test these results in an independent experiment, we analyzed retrogene and parent transcription in *A. thaliana* lines carrying somatically inducible DUO1 (Borg et al., 2011). Upon 6, 12 and 24 h DUO1 induction, we observed 36, 131 and

125 significantly up regulated and 47, 124 and 121 significantly down-regulated genes, respectively. The number of up- and down-regulated retrogenes (2 and 1, respectively) was small, showing that DUO1 controls transcription of only few specific retrogenes. Importantly, the set of significantly up-regulated retrogenes included *PCR11* retrogene (log2-fold changes in 6, 12 and 24 h: 0.26, 2.21 and 4.03; t-test *P* values: 0.010, 3.3 × 10^{-5}, 5.4 × 10^{-5}; respectively). This has been reflected by significant down-regulation of its parent *PCR2* in two out of three experimental points (log2-fold changes in 6, 12 and 24 h: -1.18, -1.60 and -0.74; t-test *P* values: 0.003, 0.007, 0.19; respectively). Therefore, we conclude that the *PCR11* retrogene gained sperm cell-specific DUO1-dependent transcription independent of its parent *PCR2*.

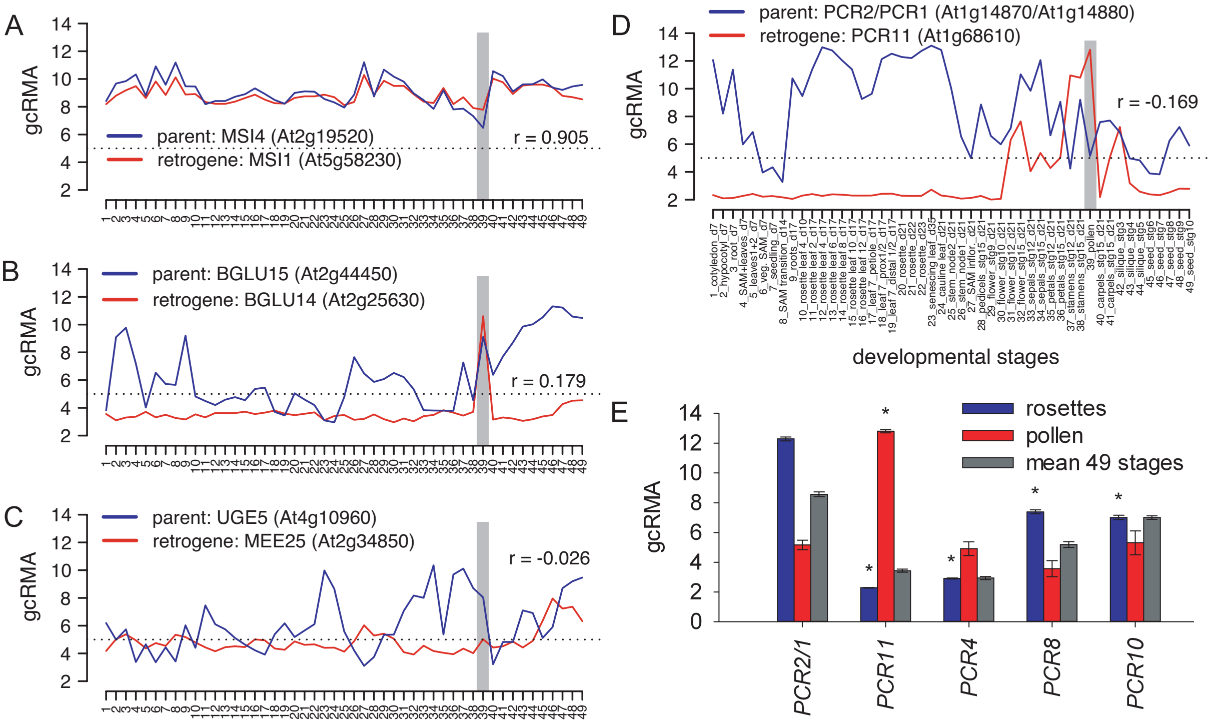

Figure 9. Gain of pollen-specific transcription by *PCR11* retrogene

(A-D) Developmental gcRMA transcription profiles of retrogenes associated with pollen growth and development and their parents. Pollen stage is highlighted by vertical gray bar. **(E)** gcRMA transcription values of *PCR* family genes in rosettes, pollen and mean of 49 developmental stages and tissues. *PCR2* and *PCR1* correspond to single microarray element and therefore are shown together. Transcription values were compared to *PCR2*/*PCR1* transcription in the same tissue and statistically analyzed by t-test. Error bars denote standard deviation of three biological replicates.

2.7 Improving gene structure annotation of *A. lyrata* genome using RNA-seq data

We aimed to identify *A. lyrata* retrogenes to study their genomic features, regulation and evolution in comparison to those of *A. thaliana*. Therefore, we used the RAT to screen for retrogenes in the published genome annotation of *A. lyrata* subsp. *lyrata* accession MN47 (Grigoriev et al., 2012; Hu et al., 2011). However, we noticed major structural differences in many *A. lyrata* genes compared to their orthologs in *A. thaliana*. This was later confirmed to be due to an inaccuracy in *A. lyrata* genome annotation version 1, which was based almost exclusively on *in silico* prediction tools (Hu et al., 2011). Two major observed annotation inaccuracies were: 1) merging two neighbor genes into one gene model (Figure 10A), and 2) splitting a single gene into two gene models (Figure 10B). This suggested that *A. lyrata* genome annotation v1, cannot be used for precise retrogene mapping. To correct the inaccuracies and enhance the overall structural annotation of *A. lyrata* genome, we generated deep transcriptome sequencing (RNA-seq) data to guide our *in silico* prediction of *A. lyrata* gene models (Figure 10). We used RNA samples from rosettes, inflorescences and shoot apical meristem tissues from plants grown under ambient conditions, for deep sequencing (Illumina technology). Additionally, RNA sequencing reads from tissue samples grown under heat and cold stress conditions were provided by Pecinka lab and Weigel lab at Max Planck Institutes. All grouped RNA-seq reads were mapped against the *A. lyrata* reference genome using Bowtie2/Tophat2 (Kim et al., 2013; Langmead and Salzberg, 2012) and used to direct the *in silico* prediction of gene models using AUGUSTUS software (Stanke et al., 2008; Stanke et al., 2006) (Figure 10). Reverse transcription PCR data further confirmed gene structure prediction. Our annotation, version 2, showed improved estimation of annotated coding regions as well as gene and exon length (Table 2). Additionally, genes were named following *A. thaliana* unified gene nomenclature. Gene model annotation was done in collaboration with Schneeberger lab at MPIPZ (Rawat, Abdelsamad et al., submitted for publication). This version of annotation (v2) was then used for identification of retrogenes.

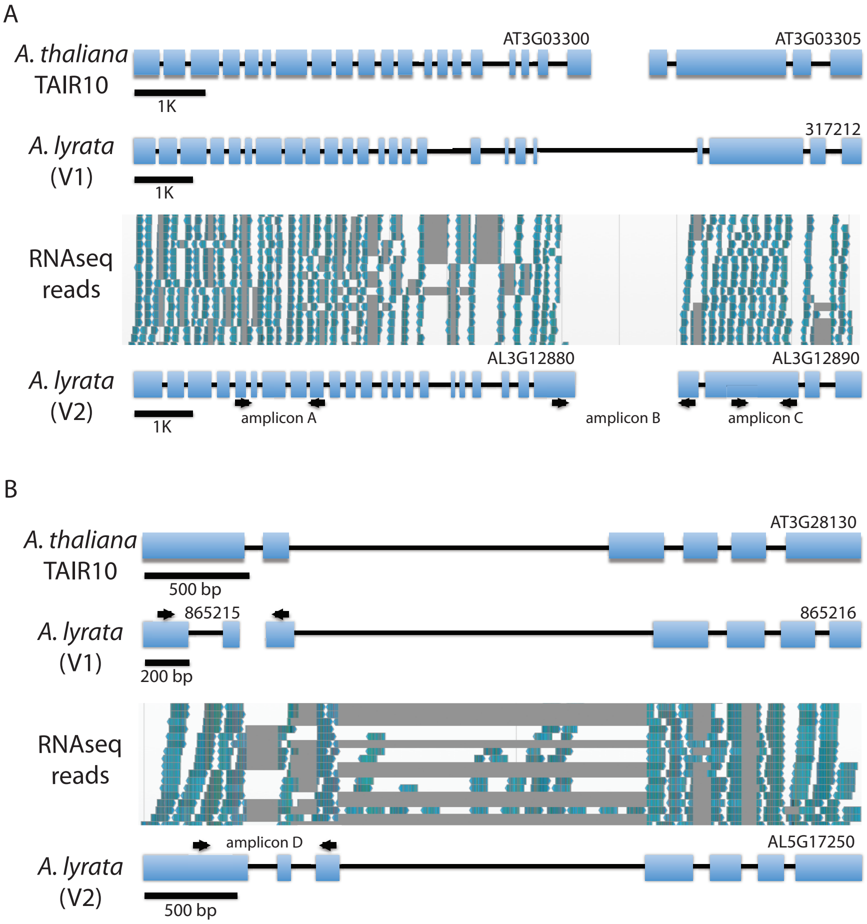

Figure 10. Enhancement of *A. lyrata* gene models using RNA-seq

(A) Example of inaccurate fusion of two neighboring gene models in annotation (V1), and its correction in our improved version (V2) using RNA-seq reads as guide for prediction. **(B)** Example of inaccurate split of a gene model into two neighboring gene models in annotation (V1), and its correction in our improved version (V2) using RNA-seq reads as guide for prediction. Black arrows indicate primer positions for reverse transcription PCR amplicons that support our predicted gene models.

Table 2. Comparison of *A. lyrata* annotation (version 2) to version 1 and TAIR10

	A. thaliana TAIR10	*A. lyrata* Version 1	*A. lyrata* Version 2
Genome size (Mbp)	120	207	207
Coding portion (Mbp)	65 (54%)	39 (19%)	79 (38%)
Protein coding genes	27416	32670	31606
Average gene length (bp)	2122	1192	2496
Average peptide length (bp)	1855	1085	1244
Total exon length (Mbp)	50	39	51
Average exon no. per gene	4.7	5.3	5.5
Average exon size (bp)	328	223	291

2.8 Fast emergence of *Arabidopsis* retrogenes revealed by interspecies comparison

We used RAT to conduct a genome-wide screen for retrogenes in our improved version of *A. lyrata* genome annotation. The screen was based on global gene paralogy search among non-transposable element protein coding genes, followed by search for retrogene-specific features among primary retrogene candidates (Figure 11B). Applying stringent quality criteria, our tool identified 168 putatively functional retrogenes (Appendix D), which represent 0.53% of *A. lyrata* protein coding genes (n = 31,606).

The close taxonomic relationship between *A. thaliana* and *A. lyrata* opened the opportunity to investigate evolution of retrogenes identified by RAT tool in both species (Table 3). Therefore, we established genome-wide gene orthology relationships among *A. thaliana*, *A. lyrata* and outgroup species *Capsella rubella* (Figure 11A). We used Inparanoid v4.01 (Remm et al., 2001) to establish gene orthology, based on protein sequence similarity. Among the three used genome annotations, pseudogenes are only annotated in *A. thaliana* genome hindering the efforts to identify the orthologs of this category in the other genomes. We specifically investigated retrogenes, of which 157 (62%) and 147 (86%) in *A. thaliana* and *A. lyrata*, respectively, had orthologs in the second species and/or in the out-group. This supports their conservation and origin before split from the last common ancestor (Figure 11C). For 51 (20%) *A. thaliana* and 23 (~14%) *A. lyrata* retrogenes, we couldn't find orthologs in the second species or *C. rubella*; i.e. these retrogenes were species-specific. The nucleotide similarity between these species-specific retrogenes and their parental genes was significantly higher than the nucleotide similarity between pre-split retrogenes and their parental genes, as shown by MWW test, P = 1.479e-07 and 0.1053 in *A. lyrata* and *A. thaliana*, respectively (Figure 11C). This supported their retroposition post the split of both species from the last common ancestor. The observed post-split retroposition events estimated an evolutionary rate of retrogene emergence of 5-10 retrogenes per million years.

Interestingly, some of pre-split and post-split retrogenes originate from the same parental gene. Like for *A. thaliana* we observed repeated secondary retroposition events; i.e. when a primary parent gives rise to a primary retrogene whose mRNA serves as the precursor for a secondary retrogene. Table 4 lists six retrogenes whose mRNAs were precursors for seven secondary retrogenes, one of which happened after the split. We also identified parental genes that gave rise to multiple retrogene copies. In total six parents gave rise to fifteen retrocopies (4 × 2, 1 × 3, 1 × 4). While most of them retroposed before the split (pre-split), a single gene (*AL7G18010*) gave rise to three retrocopies post split (Appendix D). The observed frequency of multiple retroposition events from the same parental gene is significantly higher than expected at random (MWW test, $P < 0.05$).

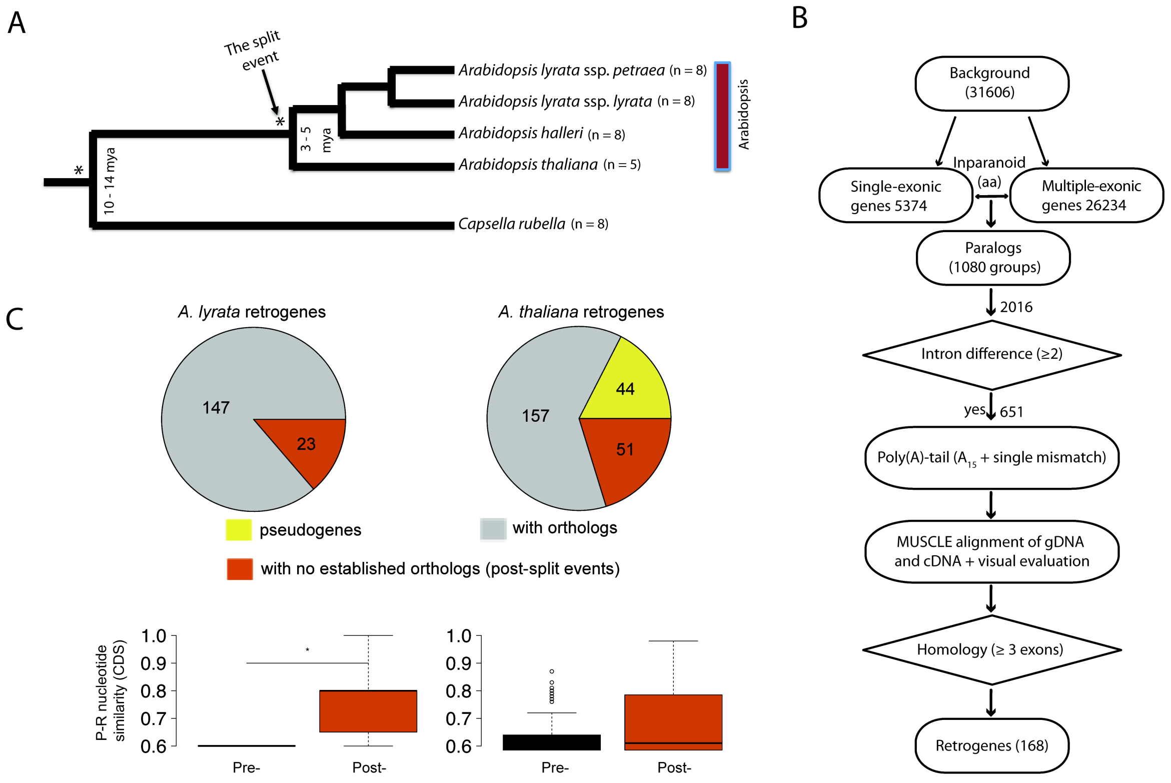

Figure 11. Novel identification of *A. lyrata* retrogenes and the relatively recent emergence

(A) A schematic strict consensus tree of some *Arabidopsis* species. *C. rubella* serves as an outgroup. The tree illustrates the relatively early divergence of *A. thaliana* from other *Arabidopsis* species. Abbreviation: mya, million years ago. Figure modified from (Clauss and Koch, 2006; Yogeeswaran et al., 2005) **(B)** Schematic representation of retrogene identification in *A. lyrata* genome (Abdelsamad and Pecinka, 2014). **(C)** Pie charts present orthology-based classification of *A. lyrata* and *A. thaliana* retrogenes into: 1. Pre-split retrogenes with an ortholog in the other species and/or the outgroup (grey), 2. Post-split retrogenes with no orthologs in the other species or the outgroup (orange), and 3. Pseudogenes (yellow). Graphs show parent-retrogene (P-R) coding region (CDS) nucleotide similarity for pre- and post-split retrogenes.

Table 3. Total (conserved) retrogenes and parental genes identified by RAT tool

	A. lyrata		*A. thaliana*	
	Parents	Retrogenes	Parents	Retrogenes
Has ortholog in the other species and/or *C. rubella*	147	145	195	157
No established orthologs in the other species or in *C. rubella*	12	23	33	51
Pseudogenes	0	0	0	44
Total	**159**	**168**	**228**	**252**

Table 4. Repeated retroposition events in *A. lyrata*

Parents		1st retroposition		2nd retroposition	
Gene ID	Introns	Gene ID	Introns	Gene ID	Introns
AL6G33170	27	*AL1G36880*	17	*AL1G44670*	9
				AL3G30910	0
AL3G33920	9	*AL1G40380*	3	*AL3G35610*	0
AL5G23750	11	*AL2G28410*	6	*AL1G21800*	0
AL3G13250	4	*AL4G19410*	2	*AL1G17240**	0
AL2G21490	9	*AL7G13720*	6	*AL5G23850*	0
AL7G22160	14	*AL7G15950*	5	*AL5G15220*	0

* Post-split event

Hence, we report the first of *A. lyrata* retrogenes (168), 86% of which had orthologs in *A. thaliana* and/or *C. rubella*. About 3.8% of parents gave rise to multiple retrogenes, some of which occurred after the split from common ancestor of both species; and ~ 3.6% of the retrogenes underwent a second retroposition without loss of functionality. About 14% and 20% of *A. lyrata* and *A. thaliana* retroposition events, respectively, occurred after the split between the two species corresponding to 5-10 successful retroposition events per million year.

2.9 *Arabidopsis* retrogenes and transposable elements share amplification mechanism but not chromosomal location and transcriptional regulation

Retrogene precursor mRNA molecules are reverse transcribed and integrated in the genome by the same enzymatic machinery that duplicates retrotransposons and retroviruses (Kaessmann et al., 2009). Many retrotransposons are integrated in repeat-rich regions, which affect their transcription and silencing (Tsukahara et al., 2012). To investigate the pattern of retrogene integration, we plotted the densities of retrogenes, parents, TEs and non-TE protein-coding genes over the eight largest scaffolds of *A. lyrata* genome assembly v1 representing the eight pseudochromosomal molecules of *A. lyrata* genome (Figure 12). In agreement with their distribution over *A. thaliana* chromosomes (Figure 2A), retrogenes and their parents show overall distribution profiles similar to that of protein-coding genes (GW) and different from that of TEs. Unlike for *A. thaliana* chromosomes (Figure 2A), the current assembly of *A. lyrata* scaffolds (Grigoriev et al., 2012; Hu et al., 2011) does not include centromeric regions; and thus, the distinguished enrichment of TEs at the centromere cannot be clearly seen in Figure 12. Hence the overall distribution of parents and retrogenes in both genomes follow that of other non-TE protein coding genes.

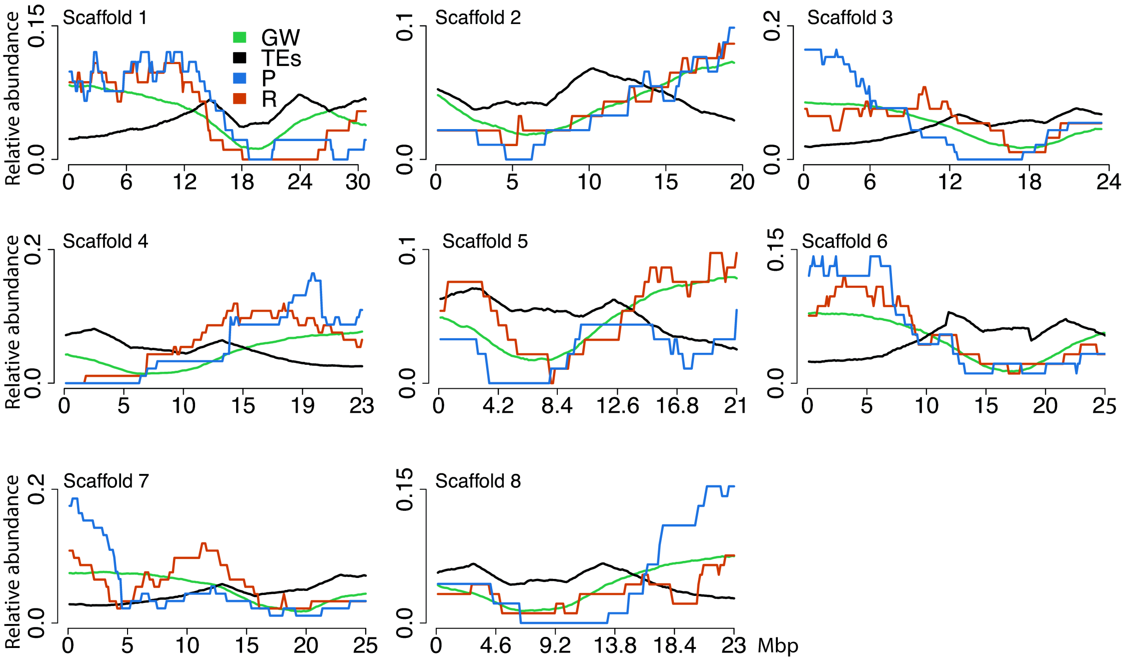

Figure 12. Distinctive chromosomal location of retrogenes and TEs

Retrogenes (R; orange) and parents (P; blue) show relative abundance (*y*-axis) over the eight main scaffolds (*x*-axis) of *A. lyrata* similar to non-TE protein coding genes (GW; green) and different from TEs (black).

To further investigate the local association of parents and retrogenes with TEs at single gene resolution, we calculated the frequency of genes overlapping with TEs, or flanked by TEs in one kilobase (1-kb) intervals upstream and downstream of gene transcription start sites (TSSs) and gene transcription termination sites (TTSs) respectively (Figure 13A). Generally, there were slightly fewer genes with TEs in their downstream regions than genes with TEs in their upstream regions. On average, there is a non-significantly lower frequency of retrogenes with flanking TEs compared to parents and genome wide genes. However, retrogenes that overlap with TEs are significantly scarce compared to genome wide protein coding genes, 4% and 11%, respectively (chi-square test, $P < 0.05$); and that might be explained by the significantly shorter average retrogene length compared to GW genes (Figure 13D). We further investigated the local surroundings of retrogenes by plotting their intergenic distances (Figure 13B). Although *A. lyrata* has longer intergenic distances and lower gene density than *A. thaliana*, retrogenes retain their preference for occurrence in gene-rich genomic regions with relatively short intergenic regions. This is similar to parents and other protein-coding genes. Hence, the local distribution of retrogenes indicates their preferential integration in gene-rich regions that are not enriched for TEs.

Retrotransposable elements transpose together with their regulatory sequences that drive their expression post integration. On the contrary, retrogenes are supposed to transpose, through a mature mRNA intermediate, without upstream regulatory regions. We wanted to explore the expression behavior of retrogenes in comparison to TEs and non-TE protein coding genes. Using our data of deep transcriptome sequencing, we calculated transcription values as a sequencing read per kilobase per million reads (RPKM). The genetic element was considered expressed if RPKM ≥ 1 in at least one tissue type, developmental stage or stress condition under investigation. We plotted transcription values (RPKM) for transcribed genes; i.e. RPKM ≥ 1 (Figure 13C), 84% and 90% of retrogenes and other non-TE protein coding genes (GW) were expressed at non-significantly different levels (MWW test, $P = 0.702$). In contrast, only 28% of TEs were expressed; and their expression was at levels significantly lower than for protein coding genes (MWW test, $P = 0.0017$). On the contrary, 96% of parental genes are expressed at significantly higher levels than genome wide genes, retrogenes and TEs (MWW test, P= 3.573e-06, 0.0067 and 8.149e-07 respectively). Hence, the frequency of expressed retrogenes and their expression levels are higher than for TEs and mirror genome wide genes.

The cell could consider retrogene copies as dispersed repeats, based on their repeated nature and TE-like duplication mechanism. Therefore, silencing small RNA molecules (sRNAs) might regulate retrogenes transcription in a pattern similar to TE regulation. We calculated the number of gene-specific and TE-specific 21nt and 24nt sRNA molecules, respectively, per kilobase of each retrogene, parent, genome-wide genes, retrotransposons and DNA transposons. Retrogenes and genome-wide genes are targeted by 21nt sRNAs at non-significantly different ratios (MWW test, P = 0.489); however, significantly higher than retrotransposons and DNA transposons (MWW test, P < 2.2e-16 and P = 3.107e-16, respectively) (Figure 13E). On the other hand, ratios of 24nt sRNA targeting retrogenes are significantly lower than ratios for retrotransposons and DNA transposons (MWW test, P = 9.914e-14 and P = 1.439e-12, respectively), but non-significantly different from genome-wide genes (MWW test, P = 0.125) (Figure 13F). Interestingly, the category of parental genes is targeted by significantly more 21nt sRNA than retrogenes and genome-wide genes (MWW test, P = 0.00616 and P = 0.00069, respectively); i.e. 21nt sRNA are targeting parent genes at the highest density among all categories of genetic elements (Figure 13E).

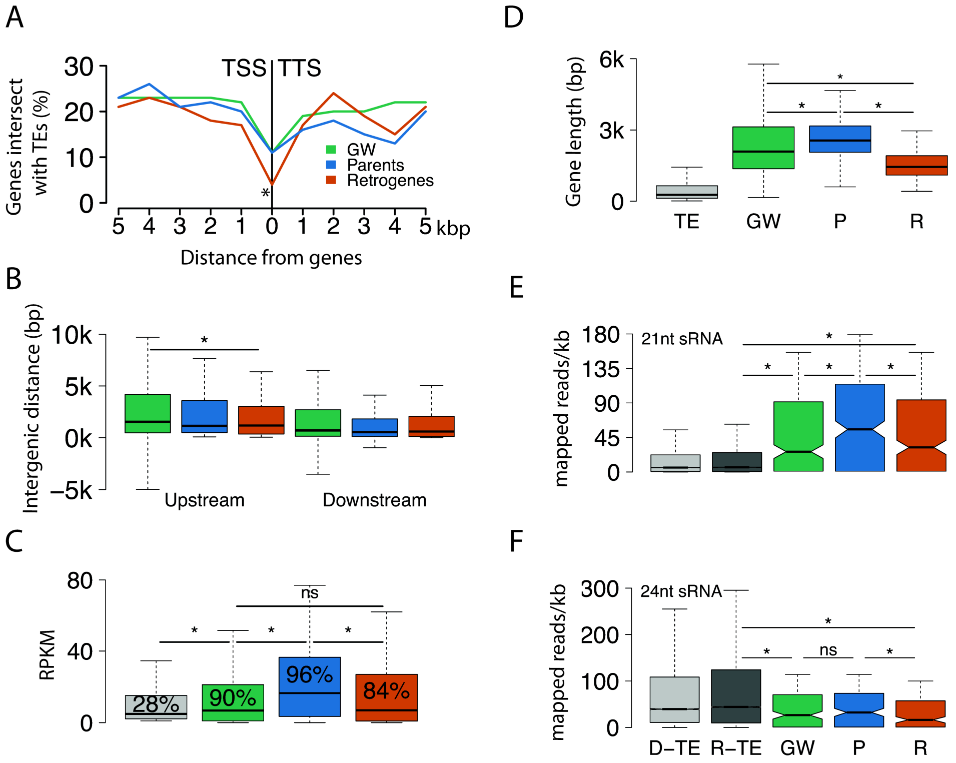

Figure 13. Retrogenes are not integrated, expressed or regulated like TEs

(A) Percentage of retrogenes overlapping with TEs (*y*-axis) in 1 kbp intervals from the gene transcription start and termination sites (TSS and TTS, respectively) is not significantly different from all non-TE protein coding genes (GW) and parents. Significant differences ($P <$ 0.05) in X^2-test relative to GW are indicated by asterisk. **(B)** Retrogenes are preferentially inserted in gene-rich genomic regions flanked by similar or less intergenic distance in bp (*y*-axis) to/than GW and parents. Significant ($P < 0.05$) difference in MWW test is indicated by asterisk. **(C)** Boxplots of mRNA sequencing reads per kilobase per million reads (RPKM) show non-significantly different expression of retrogenes and genome-wide protein coding genes, while parents are expressed at significantly higher levels. Meaningful significant ($P <$ 0.05) and non-significant ($P \geq 0.05$) comparisons in MWW test are indicated. **(D)** Boxplots of gene length indicate that retrogenes are significantly shorter than GW and parents. Meaningful significant ($P < 0.05$) comparisons in MWW test are indicated. **(E, F)** Absolute numbers of mapped regulatory 21nt sRNA reads **(E)** and 24nt sRNA **(F)** per kilobase (kb) (*y*-axis) indicate that retrognes (R, orange), parents (P, blue) and all protein coding genes (GW, green) are regulated by significantly more and significantly less gene-specific and TE-specific regulatory 21nt sRNA and 24nt sRNA, respectively, than retrotransposons (R-TE, dark grey) and DNA transposons (D-TE, light-grey). The most meaningful significant ($P < 0.05$) and non-significant ($P \geq 0.05$) comparisons in MWW test are indicated.

2.10 *NRPD2E2*$^{\text{Aly\_MN47}}$: an unusual retrogene in *A. lyrata* genome

We identified the *A. lyrata* ortholog of *A. thaliana NRPD2E2* gene. The gene encodes for the second largest subunit of *NUCLEAR RNA POLYMERASE IV 2/V 2 (NRPD2E2)*; and is essential for sRNA biogenesis (Kanno et al., 2005; Onodera et al., 2005). Whenever *NRPD2E2*$^{\text{Aly\_MN47}}$ is mentioned throughout the text, we refer to the allele (*AL3G37870*) in the North American *A. lyrata* subsp. *lyrata* accession MN47. It is a plant-specific gene with conserved structure of eight exons and seven introns (Figure 14A). Sequence alignment comparison of *NRPD2E2*$^{\text{Aly\_MN47}}$ gDNA and cDNA to its *A. thaliana* ortholog (*AT3G23780*) *NRPD2E2*$^{\text{At\_Col}}$ revealed total loss of the seven conserved introns from the retrogene *NRPD2E2*$^{\text{Aly\_MN47}}$ (Appendix E), supporting it's evolutionary origin as a retrogene. However, the published *A. lyrata* genome annotation had *in silico* predicted three introns in *NRPD2E2*$^{\text{Aly\_MN47}}$ gene structure (Hu et al., 2011). We further confirmed that by RT-PCR (data not shown) and next generation sequencing (RNA-Seq) (Figure 14D). Mapping the intronic sequences to the orthologous gene in other species has revealed that those introns have emerged through intronization of *NRPD2E2*$^{\text{Aly\_MN47}}$ exonic sequences (Figure 14A). As a result, *NRPD2E2*$^{\text{Aly\_MN47}}$ mRNA transcript was shortened compared to *NRPD2E2*$^{\text{At\_Col}}$*,* although all *NRPD2E2*$^{\text{Aly\_MN47}}$ exons are represented in the genomic sequence. Consequently, the encoded protein was significantly shorter, which would presumably affect protein structure and function.

We encountered another unusual observation when compared *NRPD2E2*$^{\text{Aly\_MN47}}$ gene locus to *A. thaliana* genome (Figure 14C). The retrogene was located in a conserved syntenic block of genes belonging to the parental gene; i.e. the retrogene had the exact genomic position of the parental gene. Interestingly, the parental gene was absent from *A. lyrata* genome; i.e. was replaced by the retrogene copy in a process called gene targeting. The retrogene copy had reduced gene length coinciding with the length of lost introns. To estimate the approximate age of the event and its conservation in *Arabidopsis,* we used PCR to compare gene length among *A. thaliana*, *Arabidopsis arenosa* and several accessions of *A. lyrata* subsp. *lyrata* and *A. lyrata* subsp. *petraea*. The results indicate that the event is specific to *A. lyrata* subsp. *lyrata* accessions (Figure 14B). Hence, it is a relatively recent retroposition event, specific to *A. lyrata* subsp. *lyrata*, where the retrogene copy has replaced the original (parental) gene copy upon integration and underwent intronization of exonic sequences post-integration.

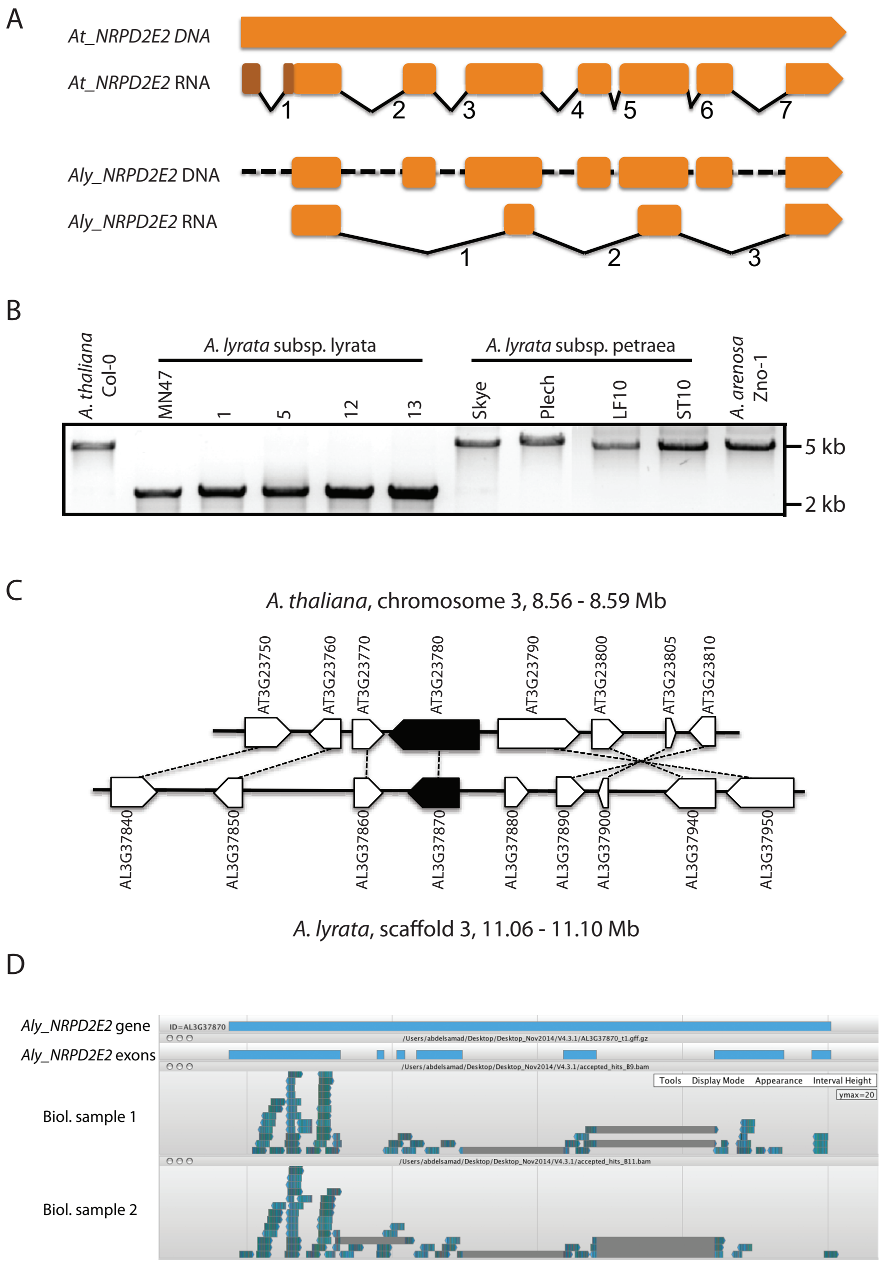

Figure 14. *NRPD2E2*$^{\text{Aly\_MN47}}$ retrogene targeting event

(A) The seven conserved introns of *NRPD2E2*$^{\text{Aly\_MN47}}$ are lost, causing the gene structure to be similar to the cDNA of *A. thaliana* ortholog (*NRPD2E2*$^{\text{At\_Col}}$). Gain of splice sites has led to the intronization of some exonic sequences; and consequently the resulting protein has lost some of its necessary domains. **(B)** PCR amplification of *NRPD2E2* genomic sequence from different *Arabidopsis* species and accessions, using the same primer pair. It confirms the reduction in *NRPD2E2*$^{\text{Aly\_MN47}}$ gene length; and suggests that the targeting is a relatively recent event that happened after the speciation of *A. lyrata* ssp. lyrata. **(C)** Representation of synteny conservation between *NRPD2E2*$^{\text{Aly\_MN47}}$ and *NRPD2E2*$^{\text{At\_Col}}$ loci. Conserved synteny location, yet loss of introns indicate retrogene targeting event in *NRPD2E2*$^{\text{Aly\_MN47}}$. (D) Mapped reads of mRNA deep sequencing (RNA-seq) confirm the expression of *NRPD2E2*$^{\text{Aly\_MN47}}$, and confirm the *in silico* predicted introns shortening its mRNA transcripts.

2.11 *Arabidopsis* retrogene targeting and the origin of their introns

The *NRPD2E2* case of retrogene prompted us to screen the genomes of *A. thaliana* and *A. lyrata* for the frequency of natural retrogene targeting events. To achieve that, we designed a novel bioinformatic tool, Targeted Retrogene Annotation Tool (TRAT) (Figure 15A). The tool runs a genome-wide comparison of protein coding genes between two closely related species to identify retrogene targeting events in both genomes. The gene is considered a primary retrogene targeting candidate in one of the species if has a minimum of three lost introns compared to its syntenic ortholog in the other species. The lost introns should present in both syntenic orthologs in the other species as well as in the out-group. Figure 15A depicts the steps and results of a TRAT tool run on *A. thaliana* and *A. lyrata* genomes. It started by establishing sequence based orthology between 20552 pairs of *A. thaliana* and *A. lyrata* genes; of which 19694 ortholog pair were located in syntenic blocks, and thus identified as syntenic orthologs and considered for further analysis. Only 473 syntenic orthologous gene pairs had difference of at least four homologous exons spanning three absent introns (363 and 110 in *A. thaliana* and *A. lyrata*, respectively). Then, the gDNA and cDNA of these gene pairs were aligned using MUSCLE v3.8.31 (Edgar, 2004) and visually inspected to evaluate exon sequence homology and intron differences. Unexpectedly, only a single and six cases in *A. thaliana* and *A. lyrata*, respectively, have passed our criteria. The high false positive rate was mainly due to discrepancies in annotated intron number per gene. Two main observed scenarios were; first, the same introns were present in both species but annotated as introns in only one of them; second, simple point mutation(s) of splice sites converted intronic sequence in one of the species into exonic or vise versa. The syntenic orthologs of the seven remaining candidate gene pairs were then identified in the out-group genome of *C. rubella*. The ancestral gene structure is crucial to determine whether the difference in intron number among the orthologous genes is a transposition-caused intron loss in one of the species or intron gain in the other. We referred the structure of the ancestral gene from the agreement in structure between the candidate gene in one of the two species and its ortholog in the *C. rubella* (Figure 15B). Of the seven primary candidates none and two retrogene targeting events were considered as likely true events after manual inspection in *A. thaliana* and *A. lyrata* genomes respectively (Figure 15C and D).

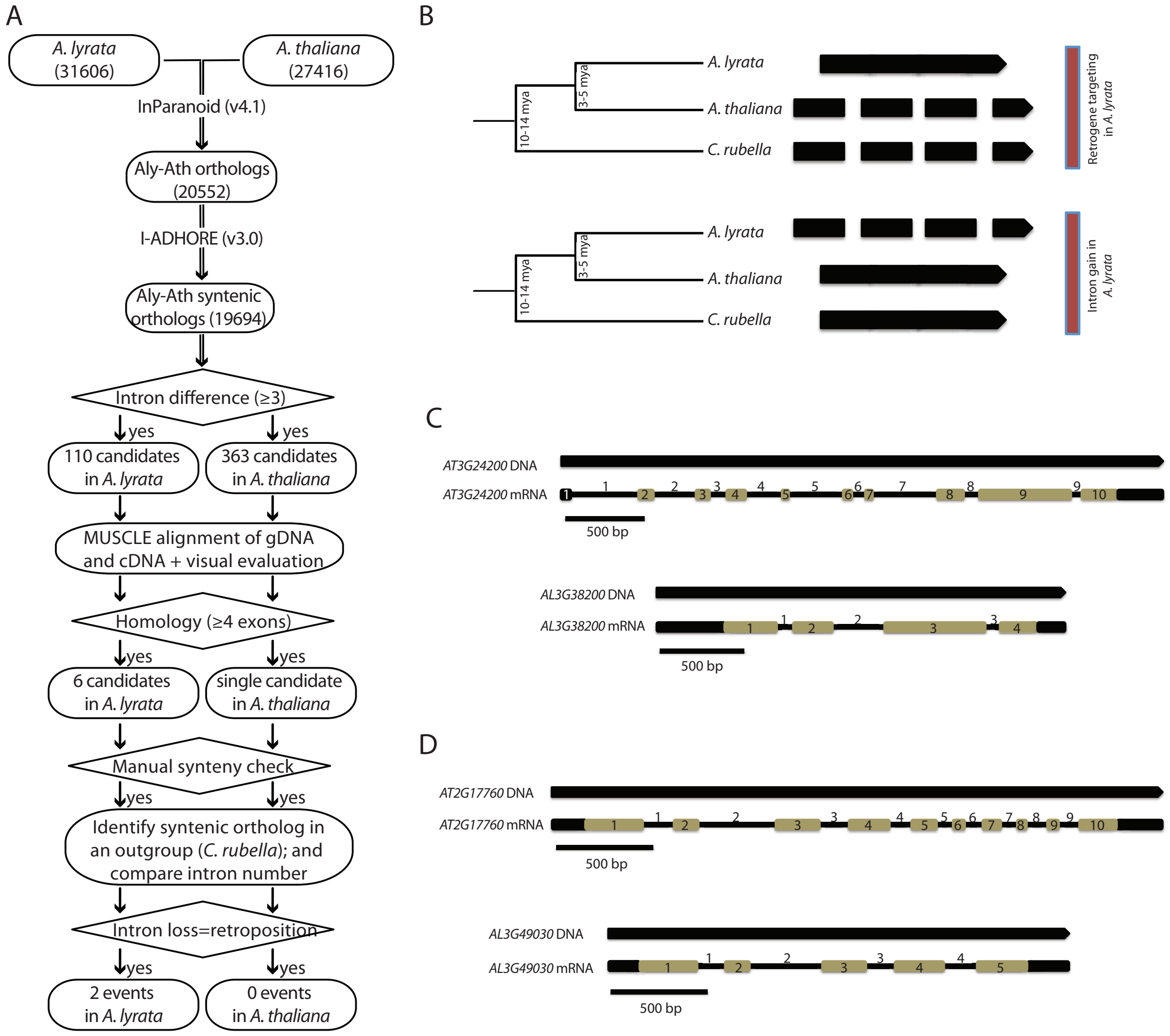

Figure 15. Identification of natural retrogene targeting events using TRAT

(A) Schematic representation of TRAT for identification of retrogene targeting events in *Arabidopsis*. **(B)** Schematic comparison of *A. thaliana, A. lyrata* and *C. rubella* orthologous gene structure to determine whether the cause of intron number difference is intron gain or transposition-based intron loss. **(C, D)** Gene models of the *A. lyrata* retrogene-targeting events and the *A. thaliana* orthologs of their parental genes.

Table 5. Natural retrogene targeting events

A. thaliana ortholog	Intron no.	Gene /CDS length	*A. lyrata* ortholog	Intron no.	Gene /CDS length	*C. rubella* ortholog	Intron no.	Gene length
AT3G24200 monooxygenase	9	3.8 /1.5 kb	*AL3G38200*	3	2.4 /1.3 kb	*Carubv10015259m*	9	3.4 /1.5 kb
AT2G17760 aspartyl protease	9	3.2 /1.2 kb	*AL3G49030*	4	2.4 /1.2 kb	*Carubv10016111m*	9	2.8 /1.5 kb

The two *A. lyrata* targeted retrogenes were *AL3G38200* and *AL3G49030* (Table 5). The first targeted retrogene (*AL3G38200*) was the syntenic ortholog of *A. thaliana* monooxygenase (*AT3G24200*). It was localized in the exact syntenic position of its parental gene; which in turn was absent from the genome. The gene encodes a conserved FAD/NAD(P)-binding oxidoreductase family protein found from yeast to humans. Comparing gene structure in *A. thaliana*, *C. rubella* and *Camelina sativa* showed a relatively conserved structure of ten exons and nine introns in *Brassicaceae* (Figure 15C). However, gDNA and cDNA sequence alignment revealed the loss of eight of the nine conserved introns in the allele of *A. lyrata* subsp. *lyrata*; i.e. one parental intron was retained in the precursor mRNA or acquired post-integration (Table 6). The loss of introns coincided with the reduction in gene length in *A. lyrata* compared to relatively conserved cDNA length (Table 5). However, our RNAseq-supported annotation of *A. lyrata* genome has predicted two more introns in gene structure; in addition to the retained parental intron (Figure 15C). When analyzed, the two introns showed no significant homology to the syntenic orthologs or any of *A. thaliana* or *C. rubella* sequences, but had a minimum of 94% sequence identity to multiple *A. lyrata*-specific intergenic regions. This suggests that the integration of these *A. lyrata*-specific introns in the targeted retrogene was relatively recent and happened post integration of the nascent retrogene copy (Table 6).

The second targeted retrogene (*AL3G49030*) was the syntenic ortholog of *A. thaliana* (*AT2G17760*). It was localized in the conserved syntenic position of its parental gene; which in turn was absent from the genome. The gene encodes a relatively conserved aspartyl protease family protein. Comparing gene structure in *A. thaliana*, *C. rubella, C. sativa, Brassica rapa* and *Eutrema salsugineum* showed a well-conserved structure of ten exons and nine introns in *Brassicaceae* (Figure 15D). However, gDNA and cDNA sequence alignment revealed the loss of six of the nine conserved introns in *A. lyrata* allele; i.e. one parental intron was retained in the precursor mRNA or acquired post-integration. This was supported by the high sequence identity between the three retained introns and the syntenic ortholog introns (Table 6). The loss of introns coincided with a reduction in gene length in *A. lyrata* compared to relatively conserved cDNA length (Table 5). Our RNAseq-supported annotation of *A. lyrata* genome has predicted a fourth intron in gene structure (Figure 15D). When analyzed, the intron did not show homology to *A. thaliana* syntenic ortholog; but showed homology to many other aspartyl protease family proteins. This suggested the retention of this intron from the parental precursor mRNA or post integration acquisition (Table 6).

Table 6. Origin of targeted retrogenes introns

Intron	Coordinates	Length	Parental?	Possible origin	Note
			AL3G38200		
1st	11298541-11298640	100 nt	No	Post-integration	Intergenic in *A. lyrata*
2nd A	11298875-11299100	226 nt	No	Post-integration	Intergenic in *A. lyrata*
2nd B	11299101-11299179	79 nt	Yes (100, 90)*	Intron retention	-
3rd	11299773-11299856	84 nt	Yes (100, 84)*	Intron retention	-
			AL3G49030		
1st	22236470-22236619	150 nt	Yes (100, 90)*	Intron retention	Family specific intron
2nd	22236755-22237127	373 nt	Yes (97, 91)*	Intron retention	Family specific intron
3rd	22237364-22237505	142 nt	Yes (100, 81)*	Intron retention	Family specific intron
4th	22237774-22237922	149 nt	No	Post-integration	Family specific intron

* Intron coverage % and identity % (C, I) to parental intron, respectively.

2.12 Introns increase stability of retrogenes transcripts

Most retrogene copies are expected to be intron-less at the time of integration. However, the three identified targeted retrogenes and approximately one-third of *Arabidopsis* retrogenes contained introns, many of which were acquired post integration. This indicated that retrogene intronization has a functional role. We tested whether intronization plays a role in retrogene mRNA stability. First, we compared the mRNA half-life of transcribed retrogenes in *A. thaliana* (n = 100), parents (n = 147) and the genome-wide set of transcribed genes (n = 13,012) included in the publicly available mRNA decay dataset (Narsai et al., 2007). The mRNA half-life of the parents and the genome-wide gene set was similar (MWW test, *P* = 0.21) and significantly longer than that of the retrogene mRNA (MWW test, *P* = 3.56×10^{-5} and 2.54×10^{-5}, respectively; Figure 16A). Furthermore, mRNA of intron-containing retrogenes (29%) had a slightly but significantly longer half-life compared to that of intron-less retrogenes (MWW test, *P* = 0.04; Figure 16B). Hence, acquisition of introns increases retrogene mRNA half-life.

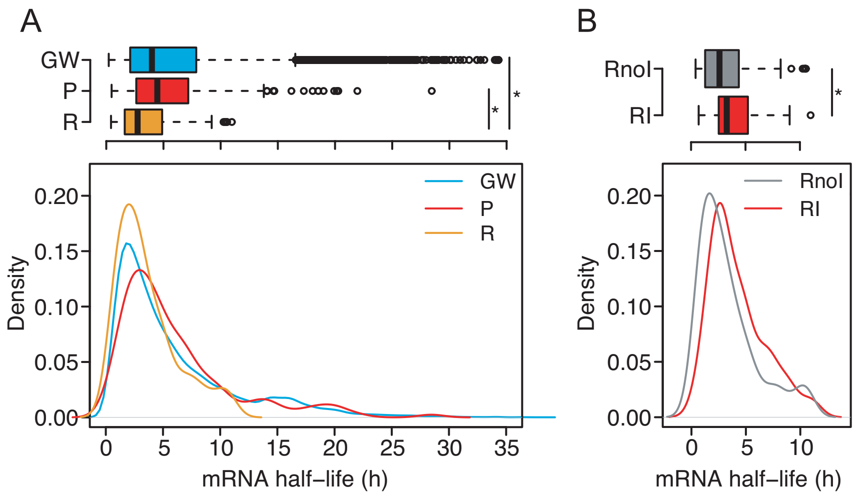

Figure 16. Introns increase half-life of retrogenes transcripts

(A, B) mRNA half-lives of genome-wide genes (GW), parents (P), retrogenes (R), intron-less retrogenes (RnoI) and intronized retrogenes (RI). Significances were calculated using the Mann-Whitney-Wilcoxon test for all group combinations within each graph, and asterisk in box plots indicate significant differences (*P* < 0.05). Non-significant (*P* ≥ 0.05) relationships are not shown.

3. DISCUSSION

3.1 The novel bioinformatic retrogene annotation tool (RAT) proved useful for identification of retrogenes across plant genomes

Gene duplication is a major force in genome evolution and adaptation. RNA-mediated gene duplication (retroposition) is capable of generating evolutionary innovations and providing fast responses to environmental challenges at single gene level. Single gene duplications have been also linked to human diseases including Parkinson's and cancer (Chartier-Harlin et al., 2004; Cooke et al., 2014) RNA-mediated duplicated genes (retrogenes) are reverse-transcribed from mature mRNA transcripts of protein coding genes (parent genes) and integrate randomly in the genome. Therefore, retrogenes are usually intron-less, retropose without their regulatory sequences and might have a poly(A)-tail in their downstream regions. We used these retrogene-specific characters to annotate them in *Arabidopsis* genomes. We developed a novel bioinformatic Retrogene Annotation Tool (RAT) to screen the genomes of *A. thaliana* and *A. lyrata* for retrogenes (Figure 1A and 11B); and consequently studied their genomic features, expression, epigenetic regulation and evolution.

3.1.1 Annotation of retrogenes in *A. thaliana* genome (TAIR10)

We annotated 251 retrogenes in *A. thaliana* genome (Appendix A), 216 of which had not previously been identified (Zhang et al., 2005; Zhu et al., 2009). The limited overlap of our set with the previous *A. thaliana* retrogene lists was most likely due to partly different search criteria and thresholds of individual methods. We detected approximately 50% of the retrogenes found in the study of Zhang et al., (2005). A specific subset of the remaining retrogenes was not accepted by RAT tool, owing to different thresholds for selection or lack of positive evidence for retroposition such as missing information on parental gene or insufficient difference in introns number. The smaller (43.2%) overlap with the set identified by Zhu et al. (2009) is due to their use of very specific criteria to identify chimeric retrogenes. These criteria apparently hamper identification of structurally simple retrogenes; while, conversely, RAT tool does not allow identification of chimeric retrogenes. The higher number of retrogenes detected with our analysis is most likely due to several factors: (i) search among *A. thaliana* pseudogenes; (ii) allowing intronized retrogenes; and (iii) accepting multiple retrocopies derived from a single parent (applied also in Zhang et al., 2005). Although we increased the number of retrogenes in *A. thaliana* three-fold, our selection criteria were conservative and the current number is most likely an underestimate based on two facts. First, we omitted several hundred candidates that had at least one paralog within the *A. thaliana* genome but did not show evidence of retroposition (*i.e*. did not differ by ≥2 introns nor had a polyA-tail). Second, none of the plant genome-wide retrogene screens detected retrogenes of the SET-domain protein group (Abdelsamad and Pecinka, 2014; Zhang et al., 2005; Zhu et al., 2009), which were identified in studies focusing specifically on the evolution of this gene family (Baumbusch et al., 2001; Borg et al., 2011; Zhu et al., 2011). Hence, 1% of *A. thaliana* genes estimated to be retrogenes is most likely an underestimation.

3.1.2 Identification of *A. lyrata* retrogenes using the newly developed genome annotation

The close taxonomic relationship of *A. lyrata* to *A. thaliana* and the recent release of its genome sequence made it a promising model organism to study gene and genome evolution in comparative studies (Clauss and Koch, 2006; Hu et al., 2011). Northern rock-cress, *A. lyrata*, is a perennial outcrosser that has two main subspecies; the eurasian *A. lyrata* subsp. *petraea* and the North American *A. lyrata* subsp. *lyrata*. The sequenced accession, MN47, belongs to the latter subspecies, and is referred to throughout this thesis (Clauss and Koch, 2006; Hu et al., 2011). The published gene models of *A. lyrata* genome were almost solely based on *in silico* prediction tools. Although usually sufficient for general annotation of genic versus intergenic regions, but incomprehensive annotation of exon-intron boundaries and alternative splicing isoforms hinders any genome-wide search for retrogenes. Therefore, an enhanced version of *A. lyrata* gene models annotation, developed by a collaboration of several laboratories including my input, was a prerequisite for successful retrogene identification. The work is currently submitted for publication. Using this resource in combination with our novel bioinformatic RAT, we revealed the first of retrogenes in *A. lyrata* genome. In total we identified 168 retrogenes passing stringent selection criteria; representing 0.53% of *A. lyrata* protein coding genes.

Although we have found approximately double and three-fold more retrogenes in *A. lyrata* and *A. thaliana*, respectively, than previously found in rice (Sakai et al., 2011), the number of conservatively estimated retrogenes per plant genome is much smaller compared to metazoans, e.g. 19.1% in humans (Marques et al., 2005; Pennisi, 2012). This difference may have several reasons. Since most of the retrogenes are identified based on intron loss, greater intron numbers in metazoan parents would simplify retrogene identification. This may partially explain the difference between *A. thaliana* and human genomes, which have average numbers of 4.2 and 7.8 introns per gene, respectively (Arabidopsis Genome Initiative, 2000; Sakharkar et al., 2004). Another possibility, which is not mutually exclusive, builds on the scarcity of WGDs in many groups of higher animals compared to plants (Gregory and Mable, 2005). This may favor local gene duplication mechanisms, including retroposition, in metazoa versus plants. Finally, higher activity of *LINE* element reverse transcriptases may be responsible for an increased retroposition rate in animals (Beck et al., 2010).

In contrast to animals, where 82% of retrocopies contain premature stop codons (Marques et al., 2005), only 17.4% of *A. thaliana* retrogenes are annotated as pseudogenes. This suggests a higher retrogene success rate in plants relative to the total number of retrocopies (Abdelsamad and Pecinka, 2014). Further support comes from our observation that several retrogenes served as parents and produced secondary retrocopies. Therefore, retroposition contributes to the functional plant genome evolution.

3.2 Fast evolutionary emergence of *Arabidopsis* retrogenes

Retroposition-based gene duplication contributes to plant genome evolution. To evaluate retrogene role in *Arabidopsis* genome evolution, we aimed to calculate their evolutionary emergence rate through comparing their conservation in *A. thaliana* and *A. lyrata* genomes, with *C. rubella* as an out-group. Gene duplication may elevate the selection pressure put on the parent gene; allowing the parent and/or the retrogene to accumulate DNA sequence polymorphisms, possibly leading to their sub- or neofunctionalization. This is an evolutionary advantage; however, from a technical point of view, post-transposition sequence polymorphisms represent a challenge towards retrogene identification screens and orthology searches. We established orthology among *A. thaliana*, *A. lyrata* and *C. rubella* protein coding genes to study retrogene conservation among them. About 86% (145) of the identified *A. lyrata* functional retrogenes (n = 168) were conserved genes in *A. thaliana* and/or *C. rubella* (Figure 11C). Similarly, 75% (157) of *A. thaliana* functional retrogenes, identified solely by our RAT tool (n = 208), were conserved genes in *A. lyrata* and/or *C. rubella*. This strongly argues for the retroposition preceding the split of *A. thaliana* and *A. lyrata* lineages, about 3-5 million year ago. Nevertheless, only small number of these pre-split retrogenes (61) is conserved as retrogenes in both species. The rest of retrogenes had orthologous genes in the other species that were not identified as retrogenes by RAT tool. A possible reason is that during the evolutionary past of these retrogenes, they might have lost retrogene-specific features in one of the species; and thus they didn't pass our stringent selection criteria. The average half-life of eukaryotic duplicated gene is 4.0 million years (Lynch and Conery, 2003); however, the existence of orthologs for many of the identifed retrogenes in *C. rubella* genome indicates a longer half-life, which is often associated to neofunctionalization (Konrad et al., 2011).

The remaining 14% (23) and 25% (51) of *A. lyrata* and *A. thaliana* functional retrogenes, respectively, were species-specific; i.e. have transposed after the split of the two species from the last common ancestor 3-5 million years ago. An alternative explanation of species-specific retrogenes would be the loss of the orthologs from the other species. However, we couldn't find the orthologs of these retrogenes in the out-group species as well. This supports the hypothesis of post-split retroposition in one of the species rather than loss of the orthologous retrogene from the other. This hypothesis is further supported by significantly higher nucleotide similarity between species-specific retrogenes and their parental genes in *A. thaliana* and *A. lyrata* (Figure 11). With 23 and 51 post-split transposition events in *A. lyrata* and *A. thaliana*, respectively, we calculated the evolutionary rate of retrogene emergence in *A. thaliana* genome. With 5 MYA as the high limit of divergence between the two species (Clauss and Koch, 2006), we estimated a minimal rate of 5-10 successful retroposition events per one million of years. That is at least ten times faster rate than previously calculated for *Arabidopsis* retrogene duplication (Zhang et al., 2005). Our stringent criteria used for retrogene identification have excluded many false negative events that would have even further increased the estimated evolutionary rate of retrogene emergence.

In general, retrogene identification is a complex process, and possibly all retrogene identification methods suffer from a specific false positive discovery rate, that is currently difficult to estimate. Additionally, retrogene annotation methods are faced by multiple challenges that increase their false negatives. Three of these challenges are; 1) post-retroposition evolution of retrogenes and their parental genes causes the loss of retrogene specific features, hindering their identification; 2) alternative and trans-splicing of the precursor mRNA may result in chimeric retrogenes which do not match standard selection criteria; 3) the quality of genome annotation is a limiting factor in retrogene identification. Consequently, many true retrogenes might get excluded due to the lack of positive evidence.

3.3 Multiple and repeated retropositions in *Arabidopsis*

One of the unresolved questions in retrogene biology is how transcripts are selected for retroposition. Although retroposition in animals has been associated with *LINE* element amplification machinery, this link has not been firmly proven in plants (Ohshima, 2013). Our data show that parent genes gave rise to multiple retrogenes before and after the split of *A. thaliana* and *A. lyrata* from the last common ancestor; and some underwent secondary retropositions. We describe twenty-two parents that produced up to seven retrogenes each in *A. thaliana*. Similarly, six parents produced up to four retrogene each in *A. lyrata*. This indicates non-random selection of parental mRNA for retroposition at least in some cases. The highly non-random pattern strongly suggests one or more common features or a signal for retroposition in *Arabidopsis*. Another support for non-random selection of precursor transcripts comes from the six and thirteen cases where a repeated retroposition has been found in *A. lyrata* and *A. thaliana* respectively (Table 1 and 5). Repeated retroposition occurs when the mRNA transcript of a retrogene serves as a precursor for a secondary retroposition. Since retrotransposon reverse transcriptases favor specific sequences in combination with transcript folding (Ohshima, 2013), it is possible that such structures exist also in transcripts of some protein coding genes. However, our preliminary efforts to identify the most common sequence motifs among those parents as potential signal for transposition were inconclusive. Therefore, among our planned future analyses is the search for potential transposition signals in the primary sequence and the secondary structure of parent transcripts as suggested by *in silico* tools (Ohshima, 2013). Similarly to other plant and animal studies (Marques et al., 2005; Potrzebowski et al., 2008; Sakai et al., 2011), we have confirmed that parents are generally strongly and ubiquitously transcribed (Figure 3A and 13C), indicating that higher amounts of transcript may increase the probability of retroposition. Although produced by the retrotransposon amplification machinery, retrogenes are located in gene rich chromosome arms (euchromatin) in *A. thaliana* (Figure 2) and *A. lyrata*,(Figure 12) and thus fundamentally differ in their genomic distribution from repetitive elements. This also holds true for their up- and down-stream intergenic regions that are not enriched for repetitive DNA (Figure 3A).

Hence, multiple and repeated retroposition indicate non-random selection of retrogene precursor transcripts from strongly and ubiquitously transcribed genes. And retrogenes integrate in gene-rich regions (open chromatin), and that facilitates their transcription to meet cellular requirements. However, they usually lack promoter sequences, as they are reverse transcribed from mature mRNA transcripts.

3.4 *Arabidopsis* retrogenes are transcribed via newly acquired promoters

One of the major limitations to the establishment of retrogenes as functional genes is the loss of *cis*-regulatory sequences (Kaessmann et al., 2009). Therefore, transposed retrocopies that cannot acquire regulatory sequences to be expressed often accumulate mutations and turn into processed pseudogenes (Hirotsune et al., 2003; Pink et al., 2011). Hence, we analyzed the retrogene transcription in *A. thaliana* using genome-wide transcription data of 49 different *A. thaliana* developmental stages by microarrays. In agreement with the observations in rice (Sakai et al., 2011), we found that retrogenes are transcribed less compared to their parents (Figure 3A). However, retrogene transcription resembles the whole genome average, suggesting that they are not 'dead on arrival' in *A. thaliana*. We saw similar pattern for *A. lyrata* retrogenes using next generation sequencing (RNA-seq) data (Figure 13C).

In humans, it has been shown that retrogenes and parents may share promoter sequences, implying a carry-over of the parental promoter by retroposition of transcripts from an upstream TSS (Okamura and Nakai, 2008). Furthermore, a recent study in rice revealed a number of retrogene–parent pairs with positively correlated transcription profiles among seven developmental stages (Sakai et al., 2011). However, this analysis did not include correction for co-transcription of random gene pairs (Sakai et al., 2011) and therefore the extent of correlation may be partially overestimated. Our data show that approximately 25% of retrogene–parent pairs and 3% of retrogene head-to-head oriented neighboring genes are co-transcribed beyond genome background in *A. thaliana*. Hence, rice and *Arabidopsis* data support the mechanism of *cis*-regulatory element carry-over in plants. However, DNA sequence analysis of parent and retrogene promoters did not reveal significant homology in rice (Sakai et al., 2011). We show similar results for *A. thaliana* (Figure 3E). Therefore, it remains unclear whether retrogenes retropose including parental upstream regulatory sequences that mutate rapidly afterwards, or they carry cryptic exonic regulatory sequences. In *A. thaliana*, majority (72%) of retrogenes are transcribed in a pattern that is not correlated to that of parents and neighboring genes, suggesting acquisition of novel *cis*-regulatory elements in most cases. Currently it is unknown whether this pattern is the result of post-integration selection or whether the compact *A. thaliana* genome offers a sufficient density of cryptic promoters.

3.5 Retrogenes are preferentially up regulated in pollen

In flies and mammals, many retrogenes are transcribed specifically in male germ cells (Bai et al., 2008; Marques et al., 2005; Vinckenbosch et al., 2006). The separation of gametes from somatic cells is very much delayed in plant compared to animal development (Wang and Ma, 2011). Therefore, somatic retroposition events in the shoot apical meristems may also be transmitted to the next generations. Therefore, we tested for tissue specific transcription of retrogenes in *A. thaliana* using a developmental transcription data series (Schmid et al., 2005) and validated our findings using RNA-sequencing datasets (Loraine et al., 2013). Surprisingly, this revealed that retrogenes are over-transcribed in pollen while overall transcription was not increased at this stage (Figure 4A). However, the pollen-specific up-regulation of retrogenes was not uniform for the whole group, as lowly transcribed retrogenes became up regulated in pollen while highly-transcribed ones were down-regulated. In addition, the set of all *A. thaliana* genes showed a similar trend. Hence, this transcription pattern is not restricted to retrogenes. More likely, many retrogenes are part of global cellular reprogramming in male gametes. So far, chromatin changes in male gametes have been associated mainly with DNA methylation changes (Ibarra et al., 2012; Slotkin et al., 2009), but there is emerging evidence that histone modifications may also contribute to pollen-specific gene reprogramming (Borges et al., 2012; Hoffmann and Palmgren, 2013). In order to identify possible causes of the observed pollen-specific transcription, we explored available data on tissue- and mutant-specific transcription and distribution of chromatin modifications. By comparing transcriptional profiles of pollen and mutants defective in transcriptional gene silencing, we excluded loss of DNA methylation and repressive H3K9me2 or heterochromatin-specific histone hyper-acetylation as the factors leading to global transcription changes in pollen. The analysis of chromatin profiles in leaves revealed that pollen up-regulated genes (and retrogenes) are depleted of transcription permissive marks (H2Bub, H3K4me3 and H3K36me3) in these tissues. Recently, it has been reported that pollen-specific genes are controlled by H3K27 methylation in *Arabidopsis* (Hoffmann and Palmgren, 2013), but this trend was much less pronounced in our dataset. This is due to different selection criteria of candidate genes in both studies. Our set of pollen up-regulated genes (n = 5,171) included the entire (99.1%) set of pollen-specific genes (n = 584; Hoffmann and Palmgren, 2013). This is most likely masking the enrichment for H3K27me modifications of specific-subset of pollen-transcribed genes in leaves. However, it has to be noted that H3K27me3 modification may control pollen-specific transcription indirectly, as suggested by our transcription analysis of the *CLF*/*SWN* and *FIE* mutants. This also holds true for the group of pollen-specific genes associated with H3K27me1 and me3 in leaf tissues (Hoffmann and Palmgren, 2013), as only a few of those genes are up-regulated in *clf*/*swn* (Figure 8D). Unexpectedly, we found correlated down-regulation of similar sets of genes (and retrogenes) in pollen and leaves of *clf*/*swn* or *fie* (*r* = 0.462 and 0.366, respectively). This indicates down-regulation of genes (and retrogenes) in response to lack of repressive chromatin marks in mutants of the polycomb group repressive complex factors. Gene down-regulation in response to the loss of repressive mark is counterintuitive and suggests that the effect is indirect, and may be achieved by an activation of specific H3K27me3 controlled suppressors such as miRNAs (Lafos et al., 2011). Based on this, we suggest that it is most likely temporary absence of permissive marks (without strong enrichment for repressive marks) that causes up-regulation of specific genes in pollen relative to somatic tissues.

Pollen-specific transcription of *A. thaliana* retrogenes was unanticipated and is analogous to retrogene transcription in animal spermatocytes (Marques et al., 2005; Vinckenbosch et al., 2006; Bai et al., 2008). Although the molecular nature of this specific transcription is so far unknown, two explanatory models have been proposed in animals (Kaessmann et al., 2009). The first suggests sperm-specific retroposition and integration into open (and thus more likely to be transcribed) chromatin that allows transcription and perpetuates this behavior. However, our data do not support this model in two aspects. First, integration into active chromatin would most likely be reflected by co-transcription between neighboring genes, which was rare in *A. thaliana*. Second, we observed many non-retrogene–genes with pollen-specific transcription. The second model proposes spermatocyte-specific transcriptional reprogramming by global chromatin changes and transcriptional activation of retrogenes and their subsequent functionalization specific to spermatocytes (Marques et al., 2005; Potrzebowski et al., 2008). In plants, pollen have been identified as the hot spot of chromatin reprogramming (Slotkin et al., 2009; Ibarra et al., 2012; Hoffmann and Palmgren, 2013), and we have shown that pollen up-regulated genes are depleted from transcription permissive chromatin marks in somatic tissues. Furthermore, we found several retrogenes that are associated with pollen growth and development and the *PCR11* retrogene that is transcribed in pollen, contrary to its parent. This is due to the presence of multiple pollen-specific DUO1 transcription factor binding motifs in its promoter. Hence, our data support the second model, and suggest that a small number of retrogenes has developed or retained male gamete-specific functions in *A. thaliana*.

The activation of many normally lowly transcribed genes and subsequent down-regulation of highly transcribed genes just prior to the onset of the next generation is an intriguing pattern with no known molecular function. However, it seems to be present in both plant and animal lineages and suggests evolutionarily conserved or analogous mechanisms that control gene transcription during this critical stage of development.

3.6 *Arabidopsis* retrogenes and retrotransposons share retroposition mechanism but not transcriptional regulation.

In mammals, LINE1 (long interspersed nuclear element 1) is a very active retrotransposon that reverse transcribes precursor transcripts of cytosolic mRNA molecules, generating retrogenes (Ding et al., 2006). LINEs have been proposed to catalyze retrogene transposition in plants; however, this has not been experimentally supported so far (Ohshima, 2013). On average, retrogenes have significantly higher transcription levels than that of TEs (Figure 13C), suggesting a different regulatory mechanism of retrogenes and TEs. Plant cells use regulatory small RNA (sRNA) to orchestrate the transcription levels of genes and TEs (Chen, 2009). Transposition and duplication of TEs signal the cell to exert a tight epigenetic transcriptional silencing preferentially through targeting by 24 nt sRNAs. In contrast, 21 nt sRNA molecules preferentially orchestrate transcription of protein-coding genes (Creasey et al., 2014; Slotkin et al., 2009). We found that the pool of sRNA targeting retrogenes is enriched for gene-specific 21 nt sRNAs and depleted of TE-specific 24 nt sRNAs, similar to other genes genome wide and in contrast to DNA and retrotransposons. Additionally, parental genes show significantly higher transcription levels than other genes in the genome, which might explain the high level of 21nt sRNA targeting (Figure 13C, E). This also suggests that those 21 nt sRNA are rather regulatory miRNAs. Hence, retrogenes share the same machinery of TEs, yet expressed and regulated at different pattern, which mirrors that of other genes.

3.7 *Arabidopsis* natural *in planta* retrogene targeting

During our manual inspection of conserved retrogenes between *A. thaliana* and *A. lyrata*, we identified *NRPD2E2*$^{\text{Aly-MN47}}$ as retrogene targeting event in *A. lyrata* genome. *NRPD2E2* is an indispensable component of small RNA biogenesis and transcriptional gene silencing in *A. thaliana* (Kanno et al., 2005; Onodera et al., 2005; Ream et al., 2009). The gene itself has emerged as a retrogene at the onset of land plant evolution (Tucker et al., 2010). After the split of *A. thaliana* and *A. lyrata*, 3-5 million years ago, another event of retroposition has occurred, where another cDNA copy of *NRPD2E2* was generated in *A. lyrata* subsp. *lyrata* (Figure 17). However, the generated copy didn’t integrate in the genome randomly as most retrogenes; but instead replaced the original (parental) gene copy. Although there is no direct experimental evidence for this model, it is strongly supported by the fact that *NRPD2E2*$^{\text{Aly-MN47}}$ is in the exact syntenic position to its *A. thaliana* ortholog (*NRPD2E2*$^{\text{At-Col}}$); yet the structure and sequences of its genomic DNA matches the cDNA structure of *NRPD2E2*$^{\text{At-Col}}$ (Figure 14). The targeting event is specific to accessions of *A. lyrata* subsp. *lyrata*, where changes in DNA sequence has created functional splice sites leading to intronization of exonic sequences, which resulted in shorter mature transcript and protein sequence. To our knowledge, this is the first reported case of *in planta* gene targeting.

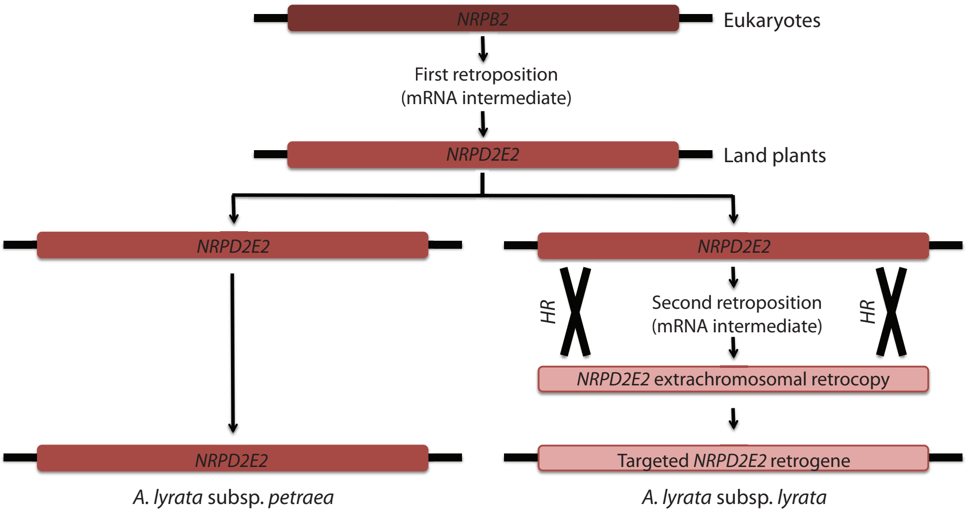

Figure 17. Evolution of *NRPD2E2* gene in *A. lyrata*.

NRPD2E2 emerged as a retrogene from the second largest subunit of RNA polymerase II (*NRPB2*) at the onset of plant evolution. The gene was replaced, through a homologous recombination (HR), with a retrocpoy generated by a second retroposition event in *A. lyrata* subsp. *lyrata*.

Gene targeting is a genetic process that requires homologous recombination to exchange two genetic elements with adequate sequence homology (Ishizaki et al., 2013). Large scale retrogene targeting has been reported in the yeast *Saccharomyces cerevisiae*, where targeted integration of retrogenes replace the original parent genes leaving intron-less copies in the exact chromosomal loci (Fink, 1987). Our results show that *Arabidopsis* retrogenes often integrate in gene-rich regions independently of the position of their parental genes. However, for retrogene targeting events, the absence of the parent gene hinders the identification of targeted retrogenes in screens that depend on intraspecies parent-retrogene paralogy; e.g.

RAT tool. Therefore, we designed a novel bioinformatic Targeted Retrogene Annotation Tool (TRAT tool) to screen for retrogene targeting events in *Arabidopsis* genomes. TRAT tool identified zero targeted retrogenes in *A. thaliana* and only two retrogene targeting events in *A. lyrata*, not including. *NRPD2E2*$^{\text{Aly-MN47}}$. It was not surprising that TRAT tool didn't identify *NRPD2E2*$^{\text{Aly-MN47}}$ retrogene targeting case. The intronization of *NRPD2E2* exonic regions has significantly altered protein sequence. That in turn, hindered establishing protein sequence based orthology between *A. thaliana* and *A. lyrata NRPD2E2* orthologs, a crucial step in TRAT tool to define retrogene targeting events.

In total, we identified three targeted retrogenes in *A. lyrata* genome and none in *A. thaliana*. Based on our calculation, retrogenes are generated at a rate of five to ten events per million year per species. Then the machinery of homologous recombination selects targeting candidates out of this pool. Unlike in yeast and mammals, homologous recombination is a minor DNA repair pathway in plants, causing gene-targeting rate of (~3 x 10^{-6}) in *Arabidopsis* (Jelesko et al., 1999). Therefore, the low rate of retroposition and homologous recombination in *Arabidopsis* explain the very little numbers of targeting events in *Arabidopsis*.

Both targeted retrogenes identified by TRAT were functional and transcribed in our analyzed tissues, indicating that targeting has not affected gene transcription. Most of retrogenes are expected to be intron-less at the time of integration; however, considerable number of retrogenes found in this study contained introns, including the targeted retrogenes (Table 5). There are multiple described mechanisms of intron gain for genes and retrogenes (Fablet et al., 2009; Irimia et al., 2008; Roy and Irimia, 2009; Szczesniak et al., 2011; Yenerall et al., 2011). This includes; *Intron Transfer*, in which an intron of a paralog is transferred to an intron-absent position in the other paralog; *Tandem Genomic Duplications*, in which the tandem DNA-based duplication of a gene segment creates an intronic sequence; *Intron Transposition*, in which a noncoding sequence transposes or gets spliced into an intron-less position in DNA sequence or in a transcript that then reverse transcribed and integrated in the genome; *Intron Retention*, in which an intron of the parent gene is not spliced out during transcript processing and gets transposed with the retrogene; and *Intronization*, in which polymorphism in exonic sequences creates functional splice sites converting exonic sequences into introns. Some of the introns in the targeted retrogenes originated by post-integration transposition process; and thus have no sequence homology to parental sequences. However, have high sequence homology to multiple *A. lyrata*-specific intergenic regions. The rest of the introns had high sequence homology to parental introns with existence in the exact order arguing for intron retention rather than intron transfer. Interestingly, retained introns would provide longer region of homology between the extrachromosomal retrogene copy and the parental gene favoring homologous recombination. Intron-less genes were shown to respond rapidly to abiotic stress (Jeffares et al., 2008), but their transcripts have relatively short half-life (Narsai et al., 2007). Retrogenes tend to acquire introns, which significantly increase their mRNA half-life (Figure 16).

Hence, we developed targeted retrogene annotation tool (TRAT); and to our knowledge, we report the first natural *in planta* gene targeting events. Retrogenes acquisition of introns increases their mRNA half-life.

4. Materials and methods

4.1 Defining the minimum length of poly(A)-tail in *Arabidopsis* genome.

We define the poly(A)-tail to be the minimum length of non-random consecutive adenine stretches down-stream of protein coding genes. We calculate the length of consecutive adenine stretches in the 150 and 250 bp downstream of stop codon for genes with and without 3' untranslated regions (UTR), respectively (Figure 18). About 99% of TAIR10 genes had adenine stretches with a length <15 nt in their downstream regions, allowing a single non-adenine nucleotide per stretch. Therefore, we considered a ≥15 nt long adenine stretch as a poly(A)-tail. Consequently, genes with such poly(A)-tail in their downstream regions were accepted as retrogene candidates.

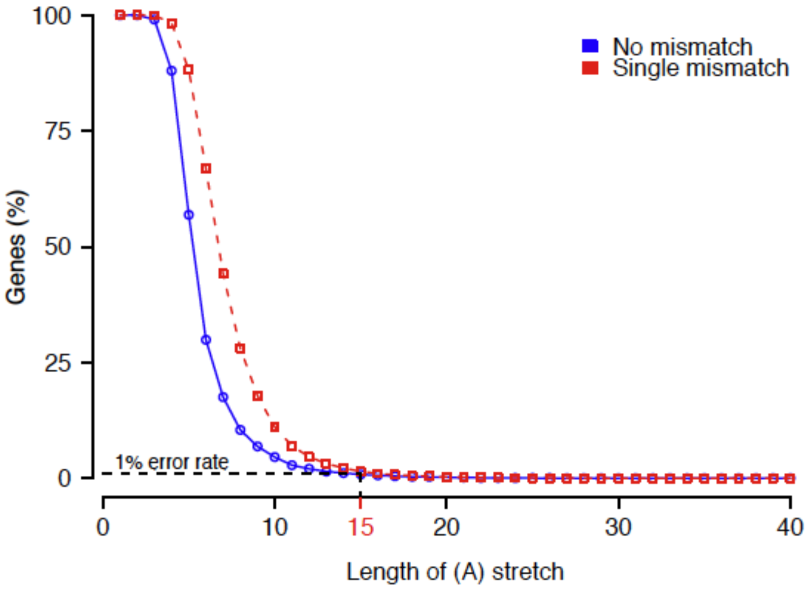

Figure 18. Defining the minimum length of non-random poly(A)-tail in *Arabidopsis* genome.

The length of consecutive adenine stretches (x-axis) in 150 or 250 bp downstream regions of the stop codons for genes with or without 3'-UTR, respectively (y-axis). Multiple adenine-stretches per gene were calculated. The 1% error rate and a single non-adenine mismatch were accepted.

4.2 Genome-wide transcription and mRNA half-life analysis

All microarray analyses were based on the publicly available datasets. Throughout the study, we used the following ATH1 cDNA microarrays (Affymetrix): wild type *A. thaliana* development produced by the AtGenExpress consortium (Schmid et al., 2005), *A. thaliana* pollen development and sperm cells datasets NASCARRAYS-48 (Honys and Twell, 2003, 2004), *ddm1-12* dataset deposited at the Gene Expression Omnibus (GEO) as the GSE18977 (Baubec et al., 2010), *kyp* dataset GEO GSE22957 (Inagaki et al., 2010), *clf*, *swn* and *clf*/*swn* dataset GEO GSE20256 and the *hda6* (*rts1-1*) dataset NASCARRAYS-538 (Popova et al., 2013). The raw data were processed and normalized using the Robust Multi-array Averaging (RMA) method (Irizarry et al., 2003) in R software (www.R-project.org) using Bioconductor (www.bioconductor.org) and the affy package (Gautier et al., 2004). The *fie* transcription values were retrieved from the GEO dataset GSE19851 (Bouyer et al., 2011) as the normalized transcription values. Retrogene and parent probes that corresponded to multiple gene models were excluded from genome-wide analysis. The transcription borderline for transcribed genes (gcRMA ≥ 5) was based on the minimal density of genes between peaks indicating absent or background signals versus high transcription signals (Figure 3). The *A. thaliana* mRNA half-life data and rosette- and pollen-specific RNA sequencing data were extracted from previously published datasets (Narsai et al., 2007; Loraine et al., 2013). Randomized sets of genes or gene pairs were generated, plots drawn and statistical tests calculated in R. Significance of density distributions was tested using the Mann-Whitney-Wilcoxon (MWW) rank sum test with correction and co-transcription correlation by the Pearson product-moment correlation coefficient (*r*).

4.3 Chromatin analysis

Chromatin data of 10-day-old *A. thaliana* seedlings were retrieved from the publicly available genome-wide atlas of chromatin modifications (Roudier et al., 2011). The frequencies for individual groups were compared. Pearson correlations were calculated in Excel (Microsoft) and heat maps were built in R.

4.4 Targeted retrogene annotation tool (TRAT)

Pairwise interspecies gene orthology was established between 27416, 31606 and 26521 annotated protein coding genes of *A. thaliana*, *A. lyrata* and *C. rubella*, respectively depending on protein sequence homology using InParanoid Version 4.1 with default parameters (Remm et al., 2001). Among the 20552 established orthology groups between *A. thaliana* and *A. lyrata*, 19694 gene pair were identified as syntenic orthologs i-AdHoRe v3.0 (Simillion et al., 2008). In total 473 orthologous pairs have a minimum differential intron number of three, and were considered for further analysis. Genomic (gDNA) and complimentary (cDNA) DNA of the candidate gene pairs were aligned using MUSCLE v3.8.31 (Edgar, 2004); and the similarity of their exon-intron structure was visually evaluated. We then manually confirmed the conserved synteny between the orthologous pair. Finally, we compared their exon-intron structure to the syntenic ortholog of the out-group *C. rubella* (Slotte et al., 2013). The protocol was executed with customized bioperl and awk scripts (Stajich et al., 2002).

4.5 Nucleotide similarity

The coding sequences (CDS) of all retrogenes and their parents genes were aligned using MUSCLE v3.8.31 (Edgar, 2004). Nucleotide diversity (NuclDiv) between each aligned pair was calculated using R software (www.R-project.org) and library pegas v0.5-1 in R (Paradis, 2010). Nucleotide similarity was calculated as (1 - NuclDiv). Data retrieval, alignment and parsing were done using customized bioperl scripts.

4.6 Small RNA data

Small RNA deep sequencing reads were retrieved from (Ma et al., 2010) and mapped against *A. lyrata* reference genome using Bowtie 2 v2.1.0 (Langmead and Salzberg, 2012). The data from three biological samples (2 flowers and one rosette) were averaged and plotted as calculated number of mapped reads per kbp of genetic elements for each category of interest.

4.7 RNA deep sequencing experiment

RNA samples from whole rosettes, floral tissues and shoot apical meristem were harvested from *A. lyrata* MN47 plants grown under controlled ambient conditions for 2 weeks, 4 weeks and 2 days post sowing, respectively. RNA-seq libraries were prepared and indexed from isolated mRNA using Illumina TruSeq RNA Sample preparation kit v2. Libraries were then sequenced using Illumina HiSeq 2000 platform generating on average 17.4 million single-end 100 nt-long read per sample. Sequencing reads were mapped against the reference genome using Tophat2 (Kim et al., 2013) and Bowtie 2 (Langmead and Salzberg, 2012). RNA sequencing reads from tissues grown under heat and cold stress growth conditions were also used [provided by B. Pietzenuk, unpublished data, and D. Koenig (Seymour et al., 2014)].

4.8 Overlap between genes and TEs

The overlap between repeatmasker-identified TEs and genes was performed with the BEDtools suite (Quinlan and Hall, 2010); as well as the overlap with five 1-kilobase pins upstream the transcription start sites and downstream transcription termination sites.

5. APPENDICES

5.1 Appendix A. Comprehensive list of *A. thaliana* retrogenes

Gene ID		Introns			CDS nucleotide similarity (%)	PolyA-tail	Retrocopies per parent	Study		
retrogene	parent	retrogene	parent	difference				Zhang (2005)	Zhu (2009)	This study
AT1G01300	AT5G10770	0	2	2	59,8	0	1	0	0	1
AT1G02000	AT4G10960	0	11	11	59,4	0	7	0	0	1
AT1G03020	AT5G63030	0	3	3	58,9	0	2	0	0	1
AT1G03300	AT2G47230	1	7	6	75,8	0	1	1	1	0
AT1G03390	AT5G41040	0	3	3	57,7	0	1	0	0	1
AT1G05020	AT4G02650	0	2	2	57,2	0	1	0	0	1
AT1G06370	AT5G14850	pseudogene	8	NA	73,6	1	1	0	0	1
AT1G08120	AT5G26110	0	6	6	92,9	0	1	0	1	1
AT1G08250	AT3G07630	0	11	11	62,1	0	4	1	0	1
AT1G11050	AT1G70520	0	6	6	59,8	0	1	0	0	1
AT1G11090	AT2G39420	0	7	7	59,3	0	1	0	0	1
AT1G11980	AT3G52590	0	4	4	61,9	0	1	0	0	1
AT1G12310	AT2G27030	0	2	2	62,3	0	1	0	0	1
AT1G14430	AT5G19580	0	1	1	57,3	1	1	0	0	1
AT1G14680	AT4G09060	0	5	5	79,5	0	1	1	1	1
AT1G15000	AT5G22980	0	8	8	57,8	0	1	0	0	1
AT1G15040	AT1G66860	1	4	3	69,5	0	1	1	1	0
AT1G15700	AT2G33040	0	7	7	58,6	0	1	0	0	1
AT1G15720	AT5G58340	0	5	5	78,8	0	1	1	1	1
AT1G16390	AT1G73220	0	1	1	59,7	1	1	0	0	1
AT1G16550	AT3G01610	pseudogene	8	NA	90,1	0	2	0	0	1
AT1G18480	AT1G07010	0	9	9	60,1	0	1	0	0	1
AT1G18970	AT1G09560	0	1	1	60,4	1	1	0	0	1
AT1G19810	AT3G01610	pseudogene	8	NA	68,1	0	2	1	0	1
AT1G20000	AT4G20280	1	4	3	80,3	0	1	0	1	0
AT1G24530	AT2G26060	0	9	9	57	0	1	0	0	1
AT1G25390	AT1G66880	3	8	5	65,1	0	1	0	0	1
AT1G26220	AT1G32070	0	7	7	59,4	0	1	0	0	1
AT1G27190	AT5G48380	0	2	2	58,9	0	1	0	0	1
AT1G28760	AT5G67610	0	9	9	57,7	0	1	0	0	1
AT1G29260	AT5G58230	0	5	5	58,8	0	1	0	0	1
AT1G29340	AT3G46510	0	3	3	58,6	0	1	0	0	1
AT1G29780	AT5G11860	0	4	4	56,9	0	1	0	0	1
AT1G30360	AT1G32090	5	10	5	58,1	0	1	0	0	1
AT1G30455	AT2G45100	1	14	13	92,8	0	1	0	1	0
AT1G31814	AT5G48385	0	4	4	57,4	0	1	0	0	1
AT1G31900	AT1G56000	pseudogene	9	NA	73,6	0	1	0	0	1
AT1G32090	AT1G58520	10	16	6	58,9	0	1	0	0	1
AT1G32480	AT4G35260	0	3	3	77,5	0	1	0	0	1
AT1G33280	AT4G17980	2	4	2	62,5	0	3	0	0	1
AT1G33612	AT5G49750	0	13	13	57,2	0	1	0	0	1
AT1G33740	AT5G56500	pseudogene	14	NA	80,3	0	1	0	0	1

AT1G34130	AT5G19690	4	22	18	63,2	0	1	0	1	0
AT1G35440	AT4G19560	0	6	6	62,7	0	1	1	0	1
AT1G43895	AT1G43890	pseudogene	7	NA	64,8	0	1	0	0	1
AT1G45100	AT3G10845	1	16	15	78,4	0	1	1	1	0
AT1G50060	AT1G50050	1	3	2	80,7	0	1	0	0	1
AT1G47056	AT4G15475	0	7	7	58,2	0	2	0	0	1
AT1G53010	AT4G35840	0	4	4	58,1	0	1	0	0	1
AT1G53170	AT5G07310	0	1	1	58	1	1	0	0	1
AT1G53345	AT5G09580	0	7	7	58,9	0	1	1	0	1
AT1G54660	AT5G23960	pseudogene	6	NA	70,7	0	1	0	0	1
AT1G54985	AT1G74190	pseudogene	4	NA	67	0	1	0	0	1
AT1G55035	AT3G06720	pseudogene	10	NA	82,2	0	1	1	0	1
AT1G55390	AT5G42280	0	2	2	82	0	1	0	0	1
AT1G55928	AT2G27285	0	3	3	79	0	1	0	0	1
AT1G56070	AT1G06220	3	11	8	58	0	1	0	0	1
AT1G57740	AT4G18465	pseudogene	22	NA	69,5	0	1	0	0	1
AT1G58330	AT1G77920	0	8	8	57,1	0	1	0	0	1
AT1G60480	AT1G10630	pseudogene	6	NA	76,7	0	1	0	0	1
AT1G60660	AT1G26340	0	2	2	64,2	0	1	0	0	1
AT1G61410	AT3G23100	0	5	5	94,1	0	1	1	0	0
AT1G63210	AT1G65440	1	19	18	81,2	0	1	1	0	0
AT1G63760	AT1G05890	pseudogene	15	NA	94,2	0	1	1	0	1
AT1G64780	AT2G38290	0	4	4	55,6	0	1	0	0	1
AT1G65210	AT4G38030	1	12	11	88,1	0	1	1	0	0
AT1G66770	AT4G10850	0	4	4	81	0	1	1	1	1
AT1G68610	AT1G14870	0	3	3	59,9	0	1	1	1	1
AT1G70430	AT5G14720	17	20	3	65,7	0	2	0	0	1
AT1G70460	AT5G56890	7	13	6	61	0	1	0	0	1
AT1G71090	AT5G01990	0	8	8	59,6	0	1	0	0	1
AT1G72820	AT5G15640	0	5	5	60,3	0	1	0	0	1
AT1G73050	AT1G72970	2	6	4	59	0	1	0	1	0
AT1G73500	AT5G56580	0	7	7	57,3	0	2	0	0	1
AT1G74550	AT2G40890	0	2	2	64,1	0	1	0	0	1
AT1G76090	AT5G13710	0	13	13	59,8	0	1	0	0	1
AT1G77130	AT3G18660	2	4	2	66,9	0	1	0	1	0
AT1G77920	AT1G08320	8	12	4	62,8	0	1	0	0	1
AT1G80510	AT5G38820	0	4	4	61,6	0	1	0	0	1
AT2G01180	AT3G02600	1	8	7	64,2	0	1	1	1	0
AT2G01372	AT3G13062	pseudogene	6	NA	75,8	0	1	0	0	1
AT2G03410	AT5G47540	0	10	10	67,9	0	1	1	0	1
AT2G04120	AT5G59150	pseudogene	2	NA	70,2	0	1	0	0	1
AT2G04280	AT3G56750	0	5	5	58,9	0	1	0	0	1
AT2G10735	AT5G45573	pseudogene	4	NA	65,4	0	1	0	0	1
AT2G11280	AT1G63640	pseudogene	20	NA	64,2	0	1	0	0	1
AT2G16830	AT2G16850	pseudogene	3	NA	65,2	0	1	0	0	1
AT2G18940	AT1G74850	0	3	3	54,7	0	1	0	0	1
AT2G19250	AT1G08520	pseudogene	14	NA	60,3	0	1	0	0	1
AT2G19550	AT3G47590	0	4	4	77,7	0	1	1	1	1
AT2G21060	AT1G75560	0	3	3	56,8	1	2	0	0	1
AT2G22760	AT4G37850	2	4	2	62,9	1	1	0	0	1

AT2G24410	AT3G20550	0	10	10	84,8	0	1	1	0	1
AT2G24748	AT4G00260	pseudogene	6	NA	78,7	0	1	0	0	1
AT2G25500	AT5G66550	0	8	8	88,9	0	1	1	0	0
AT2G25630	AT2G44450	1	11	10	86,6	0	3	1	0	0
AT2G26060	AT5G67320	9	13	4	58,2	0	1	0	0	1
AT2G26490	AT5G52820	0	11	11	58,6	0	1	0	0	1
AT2G27820	AT3G07630	0	11	11	63	0	4	1	0	0
AT2G28420	AT1G80160	0	2	2	64,6	0	1	0	0	1
AT2G28850	AT2G32440	0	8	8	57,1	0	2	0	0	1
AT2G29160	AT2G29350	pseudogene	4	NA	72	0	1	0	0	1
AT2G31230	AT1G43160	0	1	1	62,3	1	1	0	0	1
AT2G32050	AT1G17130	0	7	7	71,8	0	2	1	0	1
AT2G32510	AT1G09000	0	16	16	60,6	0	2	0	0	1
AT2G33580	AT3G21630	0	11	11	54,8	0	1	0	0	1
AT2G34760	AT2G34770	pseudogene	6	NA	61,2	0	1	0	0	1
AT2G34850	AT4G10960	5	11	6	60,7	0	7	0	0	1
AT2G34960	AT4G21120	0	2	2	62,1	0	1	0	0	1
AT2G35180	AT3G10950	pseudogene	3	NA	78,6	0	1	0	0	1
AT2G36500	AT5G50530	1	14	13	65,6	0	2	1	0	0
AT2G37970	AT3G10130	0	7	7	57,6	0	1	0	0	1
AT2G38310	AT1G01360	0	2	2	60,6	0	1	0	0	1
AT2G38980	AT3G03720	pseudogene	14	NA	62,3	0	1	0	0	1
AT2G40925	AT2G40910	0	2	2	79,1	1	1	0	0	1
AT2G42850	AT2G32440	2	8	6	57,2	0	2	0	0	1
AT2G43030	AT3G17465	0	4	4	60	0	1	0	0	1
AT2G44630	AT2G41360	0	4	4	62,1	0	1	0	0	1
AT2G45310	AT4G10960	0	11	11	56,6	0	7	0	0	1
AT3G01630	AT1G31470	0	2	2	69,8	0	1	0	0	1
AT3G02270	AT5G19485	0	12	12	58,4	0	1	0	0	1
AT3G03160	AT3G17780	0	3	3	55,9	0	2	1	0	1
AT3G04700	AT2G31560	1	3	2	62,3	0	1	0	0	1
AT3G04790	AT5G44520	0	8	8	61,5	0	1	0	0	1
AT3G07730	AT1G77270	0	4	4	62,8	0	1	0	0	1
AT3G09375	AT3G19760	pseudogene	6	NA	88,2	0	1	0	0	1
AT3G10400	AT5G64200	0	8	8	60,9	0	1	0	0	1
AT3G11810	AT2G03330	0	2	2	91,8	1	1	1	0	0
AT3G12630	AT1G51200	0	2	2	60,7	0	1	0	0	1
AT3G12910	AT4G17980	2	4	2	57,5	0	3	0	0	1
AT3G14370	AT2G26700	0	2	2	60	1	1	0	0	1
AT3G14440	AT3G63520	0	13	13	57,7	0	1	0	0	1
AT3G16510	AT4G34150	0	6	6	56,9	0	1	0	0	1
AT3G18190	AT1G24510	0	8	8	55,8	0	1	0	0	1
AT3G18420	AT4G39470	0	3	3	57,3	0	1	0	0	1
AT3G21220	AT5G56580	0	7	7	56,7	0	2	0	0	1
AT3G21933	AT3G22010	pseudogene	1	NA	72,5	0	1	0	0	1
AT3G22060	AT3G21960	1	4	3	64,9	0	1	0	0	1
AT3G23780, AT3G18090[1]	AT4G21710	7	24	17	56.9, 58.1	0	2	0	0	1
AT3G23820	AT4G10960	0	11	11	56,4	0	7	0	0	1
AT3G24330	AT2G19440	0	2	2	63,6	0	1	0	0	1

AT3G24500	AT3G58680	0	3	3	64,5	0	1	1	0	0
AT3G24927	AT2G32670	pseudogene	4	NA	57	0	1	0	0	1
AT3G25210	AT2G27800	0	2	2	59,2	0	1	0	0	1
AT3G25495	AT5G24280	pseudogene	37	NA	60,6	0	1	0	0	1
AT3G27710	AT4G34370	0	5	5	78,5	0	2	1	1	1
AT3G27720	AT4G34370	3	5	2	73,5	0	2	1	0	0
AT3G27750	AT5G09320	0	6	6	63,1	0	1	0	0	1
AT3G28720	AT5G58100	0	24	24	56,9	0	1	0	0	1
AT3G29380	AT3G10330	1	6	5	57,3	0	1	0	0	1
AT3G32316	AT1G01530	pseudogene	1	NA	67,7	0	1	0	0	1
AT3G43250	AT1G17130	0	7	7	70,7	0	2	1	0	1
AT3G43251	AT5G26880	pseudogene	6	NA	77,3	0	1	0	0	1
AT3G44717	AT5G03495	pseudogene	6	NA	76,5	0	1	0	0	1
AT3G44720	AT3G07630	0	11	11	60,4	0	4	1	0	0
AT3G45950	AT1G65660	0	9	9	82,8	0	1	1	0	1
AT3G46510	AT2G28830	3	6	3	70,2	0	1	0	0	1
AT3G46730	AT1G58410	0	2	2	57,5	0	1	0	0	1
AT3G47180	AT3G63530	0	8	8	60,2	0	1	0	0	1
AT3G47520	AT5G09660	1	7	6	63,6	0	1	0	1	0
AT3G49162	AT2G23900	pseudogene	4	NA	74,7	0	1	0	0	1
AT3G49480	AT1G05080	0	2	2	64,6	0	1	0	0	1
AT3G49750	AT5G21090	0	5	5	59	0	1	0	0	1
AT3G51110	AT5G41770	0	6	6	82,5	0	1	0	0	1
AT3G52350	AT5G08535	0	4	4	82,6	0	1	0	0	1
AT3G52950	AT5G50530	1	14	13	64,5	0	2	1	0	0
AT3G52960	AT1G65980	0	3	3	62	0	1	0	0	1
AT3G53640	AT1G13350	0	10	10	89,7	0	1	1	1	1
AT3G54900	AT3G15660	0	6	6	61,2	0	1	0	0	1
AT3G55430	AT5G24318	1	3	2	56,1	0	1	0	0	1
AT3G55950	AT1G70460	0	7	7	58,7	0	1	0	0	1
AT3G57210	AT3G25080	1	4	3	69,6	0	1	0	0	1
AT3G58330	AT3G58380	1	4	3	63,9	0	1	0	0	1
AT3G58390	AT4G27650	0	16	16	86,9	0	1	1	0	1
AT3G60610	AT1G60170	pseudogene	7	NA	95,3	0	1	1	0	1
AT3G60955	AT3G60950	pseudogene	21	NA	59,2	0	1	0	0	1
AT3G60960	AT5G28340	2	5	3	77	0	1	0	0	1
AT3G60980	AT3G60960	0	2	2	61,4	0	1	0	0	1
AT3G62350	AT1G71320	0	2	2	73,3	0	1	0	1	0
AT3G63060	AT5G15440	0	2	2	56,5	0	1	0	0	1
AT3G63380	AT4G29900	0	34	34	60,4	0	1	1	0	1
AT4G00110	AT4G10960	0	11	11	57,9	0	7	0	0	1
AT4G01170	AT1G21560	0	4	4	59,7	0	1	0	0	1
AT4G02630	AT1G01540	0	6	6	69,2	0	1	1	1	1
AT4G04693	AT4G04695	pseudogene	7	NA	56,9	0	1	0	0	1
AT4G05053	AT2G26430	pseudogene	6	NA	91,3	0	1	0	0	1
AT4G08136	AT2G18130	pseudogene	6	NA	84	0	1	0	0	1
AT4G09466	AT4G05430	0	2	2	59	0	1	0	0	1
AT4G12250	AT4G10960	0	11	11	59,7	0	7	0	0	1
AT4G14250	AT1G14570	3	9	6	65,9	0	1	0	1	0
AT4G14480	AT5G14720	0	20	20	63,4	0	2	0	0	1

AT4G15975	AT1G33480	0	2	2	62,1	0	1	0	0	1
AT4G16210	AT1G60550	2	8	6	56,6	0	1	0	0	1
AT4G16580	AT5G66720	0	4	4	66	0	1	1	1	1
AT4G16680	AT1G32490	3	27	24	72	0	1	0	1	0
AT4G17160	AT1G02130	4	7	3	59,1	0	1	0	0	1
AT4G17690	AT4G37520	0	3	3	57,8	0	2	0	0	1
AT4G17905	AT4G10150	0	2	2	57,5	0	1	0	0	1
AT4G19540	AT3G24430	7	13	6	64,9	0	1	0	0	1
AT4G20100	AT4G36850	0	10	10	64,7	0	1	0	0	1
AT4G20360	AT4G02930	0	11	11	67,8	0	1	0	0	1
AT4G20860	AT5G44400	0	2	2	60,2	0	1	0	0	1
AT4G26890	AT1G09000	0	16	16	60,6	0	2	0	0	1
AT4G29050	AT1G70110	0	1	1	74,1	1	1	0	0	1
AT4G29120	AT1G17650	0	11	11	55,6	0	1	0	0	1
AT4G30300	AT4G19210	0	12	12	71,5	0	1	1	0	0
AT4G30440	AT4G10960	0	11	11	59,8	0	7	0	0	1
AT4G33460	AT1G65410	4	9	5	60,6	0	1	0	0	1
AT4G34470	AT1G75950	0	1	1	78,5	1	1	0	0	1
AT4G35260	AT3G09810	3	6	3	60,6	0	1	0	0	1
AT4G35490	AT1G32990	0	3	3	59	0	1	0	0	1
AT4G35680	AT4G01590	2	7	5	94,5	0	1	0	1	0
AT4G36020	AT1G75560	0	3	3	57,8	0	2	0	0	1
AT4G39670	AT2G34690	0	3	3	59,3	0	1	0	0	1
AT5G01290	AT3G09100	14	16	2	80,5	0	1	0	0	1
AT5G01715	AT5G01720	pseudogene	7	NA	56,9	0	1	0	0	1
AT5G02460	AT3G55370	1	3	2	63,8	0	1	0	0	1
AT5G03980	AT1G28580	1	4	3	58,6	1	1	0	1	0
AT5G04610	AT2G31740	0	15	15	61,1	0	1	0	0	1
AT5G07225	AT5G52140	4	8	4	61,7	0	1	0	0	1
AT5G10400	AT5G10980	0	2	2	80	0	2	0	0	1
AT5G10600	AT4G37340	0	2	2	58,7	0	1	0	0	1
AT5G10880	AT3G62120	2	11	9	75,2	0	1	0	1	0
AT5G10980	AT4G40040	2	4	2	79,8	0	1	0	0	1
AT5G12030	AT1G54050	0	2	2	60,8	0	1	0	0	1
AT5G14900	AT3G62310	0	6	6	80,8	0	1	1	0	0
AT5G15870	AT1G18310	0	4	4	80,2	0	1	1	0	1
AT5G16080	AT5G06570	0	2	2	59,3	1	1	0	0	1
AT5G16510	AT3G02230	1	3	2	61,1	0	1	0	1	0
AT5G16760	AT4G08170	0	9	9	59,4	0	1	0	0	1
AT5G17190	AT3G17780	0	3	3	62,1	0	2	0	0	1
AT5G17630	AT5G54800	0	4	4	59,8	0	1	1	1	1
AT5G18202	AT3G03960	pseudogene	12	NA	81	0	1	0	0	1
AT5G18560	AT3G14230	0	2	2	57,9	0	1	0	0	1
AT5G18600	AT5G63030	0	3	3	57,7	0	2	0	0	1
AT5G22630	AT3G07630	0	11	11	63	0	4	1	0	0
AT5G22680	AT5G22720	0	8	8	76,9	0	1	0	0	1
AT5G23070	AT3G07800	0	3	3	66,2	0	1	0	0	1
AT5G23600	AT2G45330	0	7	7	96,2	0	1	1	1	1
AT5G24318	AT4G34480	3	5	2	57,9	0	1	0	0	1
AT5G25350	AT4G15475	1	7	6	55,8	0	2	0	0	1

AT5G26900	AT4G33270	0	4	4	83,4	0	3	1	0	0
AT5G27080	AT4G33270	1	4	3	83,1	0	3	1	0	0
AT5G27570	AT4G33270	0	4	4	84,6	0	3	1	0	0
AT5G28210	AT5G01290	0	14	14	76,5	0	1	1	0	1
AT5G36270	AT1G75270	pseudogene	2	NA	75,1	0	1	0	0	1
AT5G37445	AT2G32600	pseudogene	6	NA	83,6	0	1	1	0	1
AT5G39840	AT4G14790	1	15	14	60,6	0	1	0	1	0
AT5G40040	AT2G27710	0	4	4	69,7	0	1	0	0	1
AT5G40140	AT2G23140	0	4	4	59,3	0	2	0	0	1
AT5G40250	AT3G48030	0	2	2	63,4	0	1	0	0	1
AT5G40942	AT3G27060	pseudogene	1	NA	82,5	0	1	0	0	1
AT5G42090	AT2G01070	0	4	4	56,7	0	1	0	0	1
AT5G42130	AT4G39460	0	12	12	57,4	0	1	0	0	1
AT5G42260	AT2G44450	0	11	11	86,8	0	3	1	0	0
AT5G42910	AT4G34000	2	5	3	60,8	0	1	0	0	1
AT5G44170	AT1G08125	0	10	10	57,4	0	1	0	0	1
AT5G44640	AT2G44450	0	11	11	87,2	1	3	1	0	1
AT5G46100	AT4G01400	0	3	3	65	0	1	0	0	1
AT5G47000	AT4G37520	0	3	3	60	1	2	0	0	1
AT5G47640	AT3G53340	0	6	6	67,5	0	1	0	0	1
AT5G49050	AT2G47710	0	3	3	62,2	0	1	0	0	1
AT5G49200	AT5G40880	0	3	3	89,7	0	1	1	1	1
AT5G50820	AT4G17980	2	4	2	59,1	0	3	0	0	1
AT5G50960	AT4G19540	2	7	5	59,3	1	1	0	0	1
AT5G52090	AT5G37150	0	4	4	97,7	0	1	1	1	1
AT5G52415	AT2G15710	pseudogene	6	NA	68,1	0	1	0	0	1
AT5G52940	AT1G05540	0	2	2	60,9	0	2	0	0	1
AT5G54480	AT1G21740	0	3	3	59,6	0	1	0	0	1
AT5G54550	AT1G05540	0	2	2	61,8	0	2	0	0	1
AT5G54661	AT5G54660	pseudogene	1	NA	68,3	0	1	0	0	1
AT5G54940	AT4G27130	2	4	2	74,7	0	1	0	1	0
AT5G54960	AT4G33070	0	4	4	75,6	0	1	1	1	1
AT5G56450	AT3G08580	0	3	3	58,6	0	1	0	0	1
AT5G56720	AT1G04410	1	6	5	66,6	0	1	1	1	0
AT5G58010	AT4G02590	4	6	2	64,2	0	1	0	0	1
AT5G58230	AT2G19520	5	14	9	58,9	0	1	0	0	1
AT5G59630	AT1G61210	pseudogene	18	NA	58,7	0	1	0	0	1
AT5G63070	AT1G04270	0	3	3	71,1	0	1	1	1	1
AT5G63100	AT5G44600	0	7	7	58	0	1	0	0	1
AT5G63250	AT5G35740	0	2	2	56,5	0	1	0	0	1
AT5G63370	AT1G67580	0	6	6	65,7	0	1	1	1	0
AT5G63900	AT5G58610	0	8	8	63,1	0	1	0	0	1
AT5G65200	AT2G23140	0	4	4	61,7	0	2	0	0	1
AT5G65360	AT5G10980	0	2	2	80,3	0	2	0	0	1
ATCG00190	AT4G21710	0	24	24	57,1	0	2	0	0	1
ATCG00480	AT5G08670	0	8	8	64,7	0	1	0	0	1
ATCG00810	AT4G28360	0	5	5	57,9	0	1	0	0	1
ATCG01050	ATMG00580	0	3	3	57,5	0	1	0	0	1
ATCG01090	AT1G16700	0	8	8	56	0	1	0	0	1
ATCG01110	ATMG00510	0	4	4	56,7	0	1	0	0	1

[1] Retroposition followed by DNA-based duplication

5.2 Appendix B. Association of genes with epigenetic marks

Percentages of retrogenes (R), parents (P) and all genes (GW) with histone modifications and gene body DNA methylation.

	Total	**H3K4me2**		**H2Bub**		**H3K4me3**		**H3K36me3**		**H3K27me1**		**H3K27me3**		**5mC**	
	n	n	%	n	%	n	%	n	%	n	%	n	%	n	%
retrogenes (R)	183	160	87,4	83	45,4	133	72,7	124	67,8	15	8,2	60	32,8	34	18,6
genome (GW)	22616	19018	84,1	10596	46,9	15126	66,9	13720	60,7	3087	13,6	6909	30,5	7071	31,3
parents (P)	225	214	95,1	151	67,1	186	82,7	176	78,2	28	12,4	67	29,8	85	37,8
R_leaf-specific	53	52	98,1	32	60,4	51	96,2	50	94,3	1	0,0	15	28,3	12	22,6
R_all	183	160	87,4	83	45,4	133	72,7	124	67,8	15	8,2	60	32,8	34	18,6
R_pollen-specific	51	44	86,3	18	35,3	32	62,7	28	54,9	7	13,7	16	31,4	9	17,6
GW_leaf-specific	5978	5799	97,0	4167	69,7	5583	93,4	5292	88,5	307	5,1	1063	17,8	2014	33,7
GW_all	22616	19018	84,1	10596	46,9	15126	66,9	13720	60,7	3087	13,6	6909	30,5	7071	31,3
GW_pollen-specific	5156	4089	79,3	1847	35,8	2681	52,0	2333	45,2	994	19,3	2022	39,2	1435	27,8
P_leaf-specific	81	80	98,8	69	85,2	78	96,3	77	95,1	4	4,9	14	17,3	30	37,0
P_all	225	214	95,1	151	67,1	186	82,7	176	78,2	28	12,4	67	29,8	85	37,8
P_pollen-specific	48	46	95,8	27	56,3	35	72,9	30	62,5	13	27,1	15	31,3	18	37,5

5.3 Appendix C. Robust Multiarray Averaging (gcRMA) values

Transcription quantiles (Q1 to Q4; Q1 - lowly transcribed genes and Q4 - highly transcribed genes) and the group average (M) for genome wide expressed genes, DNA-duplicated genes, retrogenes, parental genes and transposable elements across 49 *A. thaliana* developmental stages.

ATGE ID	Developmental	Genome-wide expressed																								
	stage/tissue	genes					DNA duplicated genes					retrogens					parents					transposable elements (TEs)				
		M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4	M	Q1	Q2	Q3	Q4
ATGE_1	cotyledon_d7	6,6	3,5	5,6	7,4	9,7	7,6	4,3	7,2	8,6	10,5	6,6	3,4	5,3	7,8	9,9	7,4	4,2	6,8	8,1	10,2	4,5	3,3	3,8	4,4	6,4
ATGE_2	hypocotyl_d7	6,6	3,6	5,8	7,5	9,6	7,8	4,6	7,4	8,7	10,7	6,4	3,1	5,1	7,5	10,1	7,5	4,6	6,9	8,1	10,3	4,2	2,9	3,5	4	6,3
ATGE_3	root_d7	6,6	4,1	5,8	7,3	9,2	7,9	5,3	7,4	8,7	10,4	6,5	3,5	5,2	7,4	9,9	7,7	4,9	7	8,4	10,3	4,2	2,9	3,5	4,1	6,2
ATGE_4	SAM+leaves_d7	6,6	3,4	5,7	7,5	9,7	7,6	4,1	7,1	8,6	10,8	6,5	3,1	5,1	7,4	10,2	7,5	4,5	6,6	8,3	10,5	4,3	3	3,6	4,2	6,3
ATGE_5	leaves1+2_d7	6,5	3,5	5,5	7,4	9,7	7,6	4,1	7,1	8,5	10,6	6,5	3,4	5,2	7,6	10,1	7,4	4,3	6,7	8,1	10,3	4,6	3,3	4	4,5	6,4
ATGE_6	veg. SAM_d7	6,6	3,5	5,8	7,5	9,6	7,6	4,2	7,1	8,6	10,8	6,4	3,2	5,1	7,2	10,3	7,5	4,5	6,6	8,3	10,5	4,3	3	3,6	4,2	6,3
ATGE_7	seedling_d7	6,6	3,4	5,7	7,5	9,6	7,5	4,1	7	8,4	10,5	6,5	3,3	5,2	7,5	9,9	7,5	4,4	6,8	8,3	10,2	4,4	3,1	3,7	4,3	6,3
ATGE_8	SAM transition_d14	6,6	3,4	5,8	7,5	9,5	7,5	4	7	8,5	10,6	6,3	3	5,1	7,2	10,1	7,6	4,6	6,7	8,4	10,5	4,2	3	3,5	4,1	6,3
ATGE_9	roots_d17	6,6	4	5,8	7,3	9,2	7,9	5,2	7,4	8,7	10,4	6,5	3,5	5,2	7,6	9,8	7,7	4,8	7	8,4	10,3	4,1	2,9	3,5	4	6,2
ATGE_10	rosette leaf 4_d10	6,5	3,5	5,6	7,5	9,7	7,5	4,1	7	8,4	10,6	6,5	3,3	5,2	7,5	10	7,5	4,3	6,8	8,2	10,3	4,4	3,2	3,8	4,4	6,4
ATGE_12	rosette leaf 2_d17	6,6	3,4	5,6	7,5	9,7	7,7	4,4	7,3	8,7	10,5	6,5	3,3	5,1	7,9	9,8	7,5	4,1	7	8,3	10,3	4,4	3,1	3,7	4,3	6,5
ATGE_13	rosette leaf 4_d17	6,6	3,5	5,6	7,5	9,7	7,7	4,3	7,3	8,7	10,5	6,5	3,3	5,1	7,8	9,9	7,5	4,1	6,9	8,2	10,4	4,5	3,2	3,8	4,4	6,5
ATGE_14	rosette leaf 6_d17	6,6	3,5	5,6	7,5	9,7	7,7	4,3	7,3	8,7	10,5	6,6	3,4	5,2	7,8	10	7,5	4,2	6,9	8,2	10,4	4,6	3,3	3,9	4,5	6,5
ATGE_15	rosette leaf 8_d17	6,6	3,4	5,6	7,5	9,7	7,7	4,2	7,3	8,7	10,6	6,5	3,3	5,1	7,7	10,1	7,5	4,3	6,9	8,2	10,4	4,4	3,2	3,8	4,4	6,4
ATGE_16	rosette leaf 10_d17	6,6	3,5	5,6	7,5	9,7	7,7	4,2	7,2	8,6	10,7	6,5	3,3	5,1	7,6	10,1	7,5	4,3	6,8	8,1	10,4	4,5	3,3	3,9	4,5	6,4
ATGE_17	rosette leaf 12_d17	6,6	3,4	5,6	7,5	9,7	7,6	4,1	7,1	8,6	10,7	6,5	3,2	5,1	7,5	10,1	7,5	4,3	6,8	8,2	10,4	4,4	3,2	3,8	4,4	6,4
ATGE_19	leaf 7_petiole_d17	6,6	3,4	5,6	7,5	9,7	7,7	4,2	7,3	8,7	10,7	6,5	3,3	5,1	7,6	10,1	7,5	4,4	6,7	8,2	10,3	4,4	3,2	3,8	4,4	6,4
ATGE_20	leaf 7_proximal 1/2_d17	6,6	3,5	5,6	7,5	9,7	7,6	4,2	7,2	8,6	10,6	6,6	3,4	5,1	7,7	10	7,5	4,3	6,8	8,1	10,3	4,6	3,4	4	4,7	6,5
ATGE_21	leaf 7_distal 1/2_d17	6,6	3,5	5,6	7,5	9,7	7,7	4,2	7,3	8,6	10,5	6,6	3,4	5,1	7,8	10	7,5	4,2	6,9	8,2	10,3	4,6	3,3	4	4,6	6,5
ATGE_22	rosette_d21	6,6	3,4	5,6	7,5	9,7	7,7	4,3	7,3	8,6	10,7	6,5	3,2	5,1	7,7	10	7,5	4,3	6,8	8,2	10,4	4,4	3,2	3,8	4,4	6,4
ATGE_23	rosette_d22	6,6	3,4	5,6	7,5	9,8	7,7	4,2	7,2	8,6	10,7	6,4	3,3	5	7,5	10	7,5	4,3	6,8	8,3	10,4	4,4	3,2	3,8	4,4	6,4
ATGE_24	rosette_d23	6,6	3,4	5,6	7,6	9,7	7,7	4,2	7,2	8,6	10,7	6,5	3,3	5	7,6	10	7,5	4,3	6,9	8,3	10,4	4,5	3,2	3,8	4,4	6,4
ATGE_25	senescing leaf_d35	6,5	3,5	5,7	7,5	9,4	7,6	4,6	7,2	8,4	10,1	6,5	3,5	5,2	7,9	9,4	7,7	4,4	7,4	8,7	10,2	4,5	3,2	3,7	4,3	6,6
ATGE_26	cauline leaf_d21	6,6	3,4	5,6	7,5	9,6	7,6	4,3	7,3	8,6	10,4	6,5	3,3	5,1	7,9	9,7	7,6	4,3	7,2	8,5	10,3	4,5	3,2	3,8	4,4	6,5
ATGE_27	stem_2nd internode_d21	6,6	3,5	5,7	7,5	9,6	7,8	4,5	7,5	8,8	10,5	6,4	3,3	4,8	7,7	9,7	7,6	4,7	6,8	8,3	10,3	4,4	3,1	3,8	4,4	6,4
ATGE_28	stem_1st internode_d21	6,6	3,4	5,7	7,5	9,7	7,9	4,6	7,5	8,8	10,7	6,4	3,1	4,8	7,6	10	7,5	4,7	6,9	8,1	10,3	4,1	2,9	3,4	3,9	6,2
ATGE_29	SAM inflorescence_d21	6,6	3,5	5,8	7,5	9,4	7,5	4	7	8,5	10,6	6,3	3,1	5	7,2	10	7,6	4,6	6,7	8,3	10,4	4,3	3,1	3,6	4,2	6,4
ATGE_31	pedicels_stage15_d21	6,6	3,5	5,6	7,5	9,7	7,7	4,2	7,2	8,7	10,7	6,4	3,3	4,9	7,4	10	7,5	4,4	6,7	8,2	10,4	4,4	3,2	3,8	4,4	6,3
ATGE_32	flower_stage9_d21	6,6	3,5	5,8	7,5	9,6	7,6	4,2	7,1	8,5	10,6	6,4	3,3	5	7,3	10	7,7	5	6,9	8,4	10,4	4,1	2,9	3,3	3,9	6,1

ATGE_33	flower_stage10_d21	6,6	3,5	5,8	7,5	9,6	7,7	4,3	7,1	8,6	10,6	6,3	3,1	5,1	7,4	9,9	7,6	4,7	7	8,3	10,3	4	2,8	3,3	3,9	6,1
ATGE_34	flower_stage12_d21	6,6	3,6	5,7	7,5	9,6	7,9	5,1	7,2	8,6	10,7	6,4	3,3	5,2	7,3	10	7,8	5,1	7,1	8,2	10,4	4	2,8	3,3	3,9	6
ATGE_35	flower_stage15_d21	6,6	3,6	5,8	7,5	9,6	8	5,2	7,4	8,7	10,6	6,5	3,6	5,2	7,5	9,8	7,8	5,1	7,4	8,3	10,3	4,1	2,9	3,4	3,9	6,1
ATGE_36	sepals_stage12_d21	6,6	3,4	5,7	7,5	9,7	7,7	4,5	7,3	8,6	10,6	6,5	3,3	5,2	7,7	9,8	7,6	4,4	7,2	8,3	10,3	4,2	3	3,6	4,2	6,2
ATGE_37	sepals_stage15_d21	6,6	3,7	5,7	7,4	9,4	7,7	5	7,2	8,4	10,1	6,6	3,8	5,4	7,8	9,3	7,7	4,6	7,5	8,3	10,2	4,4	3,2	3,8	4,4	6,3
ATGE_39	petals_stage12_d21	6,6	3,5	5,7	7,5	9,6	7,8	4,6	7,2	8,7	10,6	6,5	3,3	5,1	7,6	9,9	7,5	4,4	6,7	8,4	10,4	4,4	3,1	3,7	4,3	6,3
ATGE_40	petals_stage15_d21	6,6	3,6	5,7	7,5	9,5	7,8	4,9	7,4	8,7	10,2	6,5	3,6	5,2	7,7	9,6	7,6	4,6	7,3	8,3	10,1	4,4	3,2	3,8	4,4	6,3
ATGE_41	stamens_stage12_d21	6,6	4,3	5,7	7,1	9,1	8	6,3	7,3	8,5	10,2	6,7	4,2	5,7	7,3	9,4	7,8	5,3	7,4	8,1	9,9	4,5	3,4	4	4,5	6,2
ATGE_42	stamens_stage15_d21	6,6	4,1	5,7	7,2	9,2	8	6,1	7,4	8,5	9,9	6,7	4,3	5,7	7,4	9,4	7,6	4,9	7,2	8,2	9,8	4,6	3,5	4	4,6	6,3
ATGE_43	pollen	6,4	5,5	6	6,5	7,6	7,3	6,9	6,6	7,3	8,3	6,8	6	6,6	6,6	7,9	7,2	5,8	7	7,5	8,4	5,9	5,3	5,7	6	6,7
ATGE_45	carpels_stage12_d21	6,6	3,4	5,8	7,5	9,6	7,7	4,2	7,3	8,6	10,7	6,4	3	5	7,4	10	7,7	4,9	6,7	8,5	10,4	4,2	3	3,5	4,1	6,2
ATGE_73	carpels_stage15_d21	6,6	3,5	5,7	7,5	9,6	7,8	4,6	7,2	8,6	10,8	6,4	3,2	5,1	7,4	10	7,8	5,2	7	8,4	10,5	4,1	2,9	3,4	4,1	6
ATGE_76	silique_stage3	6,6	3,7	5,7	7,4	9,6	7,8	5,1	7,2	8,5	10,6	6,5	3,4	5,3	7,3	9,9	7,6	4,9	7	8,1	10,3	4,2	3	3,5	4,1	6,2
ATGE_77	silique_stage4	6,6	3,7	5,8	7,4	9,6	7,9	5	7,3	8,6	10,6	6,5	3,1	5,3	7,5	10,1	7,7	5	7,2	8	10,3	4,1	2,8	3,4	4	6,1
ATGE_78	silique_stage5	6,6	3,7	5,7	7,4	9,6	7,9	5,1	7,3	8,6	10,6	6,5	3,2	5,1	7,6	10,1	7,7	5	7,1	8	10,4	4,1	2,9	3,4	4,1	6,1
ATGE_79	seed_stage6	6,6	3,9	5,8	7,3	9,3	7,7	4,9	7,1	8,4	10,4	6,4	3,4	4,8	7,4	10	7,6	5,2	6,7	7,9	10,2	4,4	3,2	3,7	4,4	6,2
ATGE_81	seed_stage7	6,6	4,1	5,7	7,3	9,3	7,7	5,1	7,1	8,4	10,3	6,4	3,5	4,9	7,3	9,9	7,6	5,2	6,8	8	10,1	4,5	3,3	3,9	4,5	6,3
ATGE_82	seed_stage8	6,6	4,4	5,8	7,1	9	7,5	5,2	6,9	8,1	9,8	6,6	3,9	5,2	7,5	9,6	7,6	5,2	7,1	8,1	9,7	5	3,8	4,4	5,1	6,6
ATGE_83	seed_stage9	6,5	4,5	5,7	7	8,8	7,4	5,2	6,8	7,9	9,7	6,6	4,1	5,3	7,4	9,6	7,5	5,3	7,1	8	9,6	5,3	4,2	4,8	5,4	6,7
ATGE_84	seed_stage10	6,5	4,5	5,8	7	8,8	7,4	5,2	6,8	7,9	9,5	6,6	4,1	5,3	7,5	9,6	7,6	5,2	7,2	8,1	9,5	5,2	4,1	4,8	5,4	6,7

5.4 Appendix D: RAT-generated list of *A. lyrata* retrogenes

A detailed list of all parent and retrogene identified by RAT using our enhanced version of *A. lyrata* genome.

Retrogene	Parent	Introns							Has *A. thaliana* ortholog		Has *C. rubella* ortholog	
ID	ID	Retrogene	Parent	P-R	Poly(A)Tail	NuclSimil	Retrocopies per parent	After split	Retrogene	Parent	Retrogene	Parent
AL4G22540	AL4G26720	0	8	8	0	0.6	4	0	1	1	1	1
AL4G39710	AL4G26720	2	8	6	0	0.6	4	0	1	1	1	1
AL7G43910	AL4G26720	3	8	5	0	0.6	4	0	1	1	1	1
AL7G34050	AL4G26720	6	8	2	0	0.6	4	0	1	1	1	1
AL6G39240	AL7G18010	0	4	4	0	0.8	3	1	0	1	0	1
AL5G22670	AL7G18010	1	4	3	0	0.8	3	1	0	1	0	1
AL4G18580	AL7G18010	1	4	3	0	0.8	3	1	0	1	0	1
AL7G22860	AL1G29580	0	11	11	0	0.6	2	0	1	1	1	1
AL5G19130	AL3G21330	1	6	5	1	0.6	2	0	1	1	0	1
AL4G21160	AL8G17110	0	11	11	0	0.6	2	0	1	1	1	1
AL1G44670	AL1G36880	9	17	8	0	0.6	2	0	1	1	1	1
AL3G40740	AL1G29580	7	11	4	0	0.6	2	0	1	1	1	1
AL3G30910	AL1G36880	0	17	17	0	0.6	2	0	0	1	1	1
AL5G19050	AL3G21330	1	6	5	1	0.6	2	0	0	1	1	1
AL5G41340	AL8G17110	9	11	2	0	0.7	2	1	0	1	0	1
AL6G14140	AL4G25900	0	15	15	0	0.6	1	0	1	1	1	1
AL6G39580	AL6G10840	0	14	14	0	0.8	1	0	1	1	1	1
AL2G34550	AL6G23910	0	13	13	0	0.6	1	0	1	1	1	1
AL8G19170	AL7G11440	0	12	12	0	0.6	1	0	1	1	1	1
AL7G32990	AL4G37380	0	12	12	0	0.9	1	0	1	1	1	1
AL4G21650	AL3G18710	0	11	11	1	0.6	1	0	1	1	1	1
AL7G32790	AL6G48800	0	11	11	0	0.7	1	0	1	1	1	1
AL8G14680	AL4G41660	0	11	11	0	0.9	1	0	1	1	1	1
AL8G39780	AL8G33780	0	9	9	0	0.6	1	0	1	1	1	1
AL1G41580	AL8G42360	0	9	9	0	0.6	1	0	1	1	1	1
AL8G15510	AL1G18350	0	9	9	0	0.6	1	0	1	1	1	1
AL1G30500	AL1G16930	0	8	8	0	0.6	1	0	1	1	1	1
AL5G25560	AL5G44960	0	8	8	0	0.6	1	0	1	1	1	1
AL2G29120	AL6G10040	0	8	8	0	0.6	1	0	1	1	1	1
AL3G14820	AL8G14980	0	8	8	0	0.6	1	0	1	1	1	1
AL3G30780	AL1G39550	0	8	8	0	0.6	1	0	1	1	1	1
AL8G38860	AL8G14740	0	8	8	0	0.6	1	0	1	1	1	1
AL1G26720	AL3G21440	0	8	8	0	0.6	1	0	1	1	1	1
AL6G27270	AL7G17430	0	8	8	0	0.6	1	0	1	1	1	1
AL1G38670	AL1G45000	0	7	7	0	0.6	1	0	1	1	1	1
AL4G26320	AL1G28970	0	7	7	0	0.7	1	0	1	1	0	1
AL1G21840	AL4G35250	0	7	7	0	0.6	1	0	1	1	1	1
AL3G21400	AL8G40130	0	7	7	0	0.6	1	0	1	1	1	1
AL1G12730	AL1G38350	0	7	7	0	0.8	1	0	1	1	1	1

AL5G23850	AL7G13720	0	6	6	0	0.9	1	0	1	1	0	0
AL2G36150	AL7G33810	2	7	5	0	0.6	1	0	1	1	1	1
AL2G30940	AL6G26030	0	5	5	0	0.6	1	0	1	1	1	1
AL5G16810	AL7G16670	0	5	5	0	0.8	1	0	1	1	1	1
AL5G34890	AL3G27820	0	5	5	0	0.6	1	0	1	1	1	1
AL2G23570	AL8G38550	0	5	5	0	0.6	1	0	1	1	1	1
AL3G47110	AL1G42630	0	5	5	0	0.8	1	0	1	1	1	0
AL1G42570	AL6G22240	0	4	4	0	0.6	1	0	1	1	1	1
AL2G39280	AL7G50180	0	4	4	0	0.6	1	0	1	1	1	1
AL1G26320	AL6G40550	0	4	4	0	0.8	1	0	1	1	1	1
AL3G31090	AL7G11450	0	4	4	0	0.6	1	0	1	1	1	1
AL4G39920	AL3G29980	0	4	4	0	0.6	1	0	1	1	1	1
AL6G13490	AL1G41240	0	4	4	1	0.6	1	0	1	1	1	0
AL6G28200	AL8G29320	0	4	4	0	0.6	1	0	1	1	1	1
AL7G37270	AL8G43400	0	4	4	0	0.7	1	0	1	1	1	1
AL7G37950	AL7G36970	2	6	4	0	0.6	1	0	1	1	1	0
AL1G12840	AL7G47230	0	3	3	0	0.6	1	0	1	1	1	1
AL1G27490	AL8G33340	0	3	3	0	0.8	1	0	1	1	1	1
AL8G38800	AL1G13790	0	3	3	0	0.7	1	0	1	1	1	1
AL1G42150	AL5G24530	0	3	3	0	0.6	1	0	1	1	1	1
AL2G26130	AL1G26570	0	3	3	0	0.6	1	0	1	1	1	1
AL3G39070	AL5G39180	0	3	3	0	0.6	1	0	1	1	1	1
AL6G33630	AL3G18890	0	3	3	0	0.7	1	0	1	1	1	1
AL7G11790	AL4G29490	0	3	3	0	0.6	1	0	1	1	1	1
AL7G15460	AL1G46120	0	3	3	0	0.6	1	0	1	1	1	1
AL8G21750	AL4G45430	0	3	3	0	0.6	1	0	1	1	1	1
AL8G28930	AL1G34340	0	3	3	0	0.6	1	0	1	1	1	1
AL8G31200	AL6G23700	0	3	3	0	0.6	1	0	1	1	1	1
AL6G27720	AL1G53600	0	3	3	0	0.6	1	0	1	1	1	1
AL1G12380	AL8G38750	0	3	3	0	0.6	1	0	1	0	1	0
AL1G14680	AL6G49170	0	2	2	0	0.6	1	0	1	1	1	1
AL2G32880	AL4G37200	0	2	2	0	0.6	1	0	1	1	0	1
AL3G10530	AL1G48290	0	2	2	0	0.7	1	0	1	1	1	1
AL4G29790	AL7G32250	0	2	2	0	0.6	1	0	1	1	1	1
AL5G32720	AL2G22290	0	2	2	0	0.6	1	0	1	1	1	1
AL4G34020	AL8G26980	0	2	2	0	0.6	1	0	1	1	1	1
AL7G36050	AL6G24390	0	2	2	0	0.6	1	0	1	1	1	1
AL7G34120	AL5G44950	0	13	13	1	0.6	1	0	1	1	1	1
AL3G47000	AL3G49220	0	13	13	0	0.6	1	0	1	1	1	1
AL1G36880	AL6G33170	17	27	10	0	0.6	1	0	1	1	1	1
AL3G21240	AL8G31900	0	10	10	0	0.6	1	0	1	1	1	1
AL1G29950	AL3G20310	0	9	9	0	0.6	1	0	1	1	1	1
AL6G36340	AL1G53350	0	9	9	0	0.6	1	0	1	1	1	1
AL4G25820	AL4G26770	0	8	8	0	0.5	1	0	1	1	1	1
AL7G22170	AL6G23900	0	8	8	0	0.6	1	0	1	1	1	1
AL8G42100	AL1G18760	0	8	8	0	0.6	1	0	1	1	1	1

AL3G45420	AL5G25210	0	8	8	0	0.7	1	0	1	0	0	0
AL6G46890	AL4G27460	0	7	7	0	0.6	1	0	1	1	1	1
AL8G42760	AL7G39130	0	7	7	0	0.6	1	0	1	1	1	1
AL3G19350	AL4G33420	0	7	7	0	0.6	1	0	1	1	1	1
AL1G19350	AL7G16970	0	6	6	0	0.6	1	0	1	1	1	1
AL1G40380	AL3G33920	3	9	6	0	0.6	1	0	1	1	1	1
AL1G61870	AL3G22570	0	6	6	0	0.6	1	0	1	1	0	0
AL6G29420	AL7G15710	0	6	6	0	0.6	1	0	1	1	1	1
AL8G15620	AL7G17130	0	5	5	0	0.6	1	0	1	1	1	1
AL6G42340	AL6G23660	0	5	5	0	0.6	1	0	1	1	1	1
AL5G29410	AL1G30410	0	5	5	0	0.6	1	0	1	1	1	1
AL7G40600	AL4G34690	0	5	5	0	0.7	1	0	1	1	1	1
AL5G15220	AL7G15950	0	5	5	0	0.6	1	0	1	0	1	1
AL6G16340	AL7G15040	0	4	4	0	0.6	1	0	1	1	1	1
AL5G44970	AL2G35660	0	4	4	0	0.6	1	0	1	1	1	1
AL3G24430	AL8G31600	0	4	4	1	0.6	1	0	1	1	1	1
AL6G19840	AL3G45050	0	4	4	0	0.6	1	0	1	1	1	1
AL6G33810	AL3G28350	0	4	4	0	0.6	1	0	1	1	0	1
AL6G52050	AL1G11610	0	4	4	0	0.6	1	0	1	1	1	1
AL3G34150	AL2G39750	0	4	4	0	0.6	1	0	1	0	0	0
AL5G14500	AL1G65470	0	4	4	0	0.7	1	0	1	0	1	1
AL7G13720	AL2G21490	6	9	3	0	0.9	1	0	1	1	0	1
AL1G64270	AL7G12280	0	3	3	0	0.8	1	0	1	1	1	1
AL3G48250	AL6G12540	3	6	3	0	0.6	1	0	1	1	1	0
AL3G15000	AL3G16740	0	3	3	0	0.6	1	0	1	1	1	1
AL3G35610	AL1G40380	0	3	3	0	0.6	1	0	1	1	1	1
AL4G33340	AL8G39940	0	3	3	1	0.9	1	0	1	1	0	1
AL7G10720	AL7G41710	0	3	3	0	0.7	1	0	1	1	1	1
AL5G41390	AL4G20040	0	3	3	0	0.7	1	0	1	0	1	1
AL7G45940	AL6G14700	0	3	3	0	0.6	1	0	1	0	0	0
AL304U10010	AL3G15770	0	3	3	0	0.7	1	0	1	0	1	0
AL7G30850	AL2G13410	9	11	2	0	0.7	1	0	1	1	1	1
AL2G38270	AL2G38230	1	3	2	0	0.7	1	0	1	1	1	1
AL4G37190	AL6G14770	1	3	2	0	0.7	1	0	1	1	1	1
AL1G36430	AL1G48400	0	2	2	0	0.7	1	0	1	1	1	1
AL1G62060	AL6G28130	0	2	2	0	0.6	1	0	1	1	1	1
AL2G17100	AL1G20700	0	2	2	1	0.7	1	0	1	1	1	1
AL3G40210	AL4G34620	0	2	2	0	0.6	1	0	1	1	1	1
AL4G23460	AL2G34100	0	2	2	0	0.6	1	0	1	1	1	1
AL5G43590	AL8G29350	0	2	2	0	0.6	1	0	1	1	1	1
AL6G20800	AL7G43970	0	2	2	0	0.6	1	0	1	1	1	1
AL7G13410	AL6G21320	0	2	2	0	0.6	1	0	1	1	1	1
AL4G19410	AL3G13250	2	4	2	1	0.6	1	0	1	1	1	1
AL5G38810	AL3G35480	1	3	2	0	0.7	1	0	1	0	1	0
AL6G26760	AL1G44770	0	2	2	0	0.7	1	0	1	1	1	1
AL7G40660	AL2G28320	0	16	16	0	0.6	1	0	0	0	1	1

AL7G15950	AL7G22160	5	14	9	0	0.6	1	0	0	0	1	1
AL5G29500	AL6G27260	0	13	13	0	0.6	1	0	0	0	1	1
AL4G27160	AL6G44850	0	11	11	0	0.6	1	0	0	0	1	1
AL4G28070	AL3G35110	0	11	11	0	0.6	1	0	0	0	1	1
AL5G33540	AL5G14280	0	11	11	0	0.8	1	0	0	0	1	1
AL2G28410	AL5G23750	6	11	5	0	0.6	1	0	0	0	1	1
AL3G34620	AL8G31410	0	7	7	0	0.6	1	0	0	0	1	1
AL6G25230	AL4G44970	0	6	6	0	0.8	1	0	0	1	1	0
AL1G21800	AL2G28410	0	6	6	0	0.6	1	0	0	0	1	1
AL6G49190	AL1G10820	0	6	6	0	0.7	1	0	0	0	1	1
AL8G39120	AL2G24870	0	6	6	0	0.7	1	0	0	0	1	1
AL7G17240	AL3G46010	0	5	5	0	0.6	1	0	0	1	1	1
AL1G41940	AL8G33230	0	5	5	1	0.6	1	0	0	0	1	1
AL1G47600	AL6G34770	0	4	4	0	0.6	1	0	0	0	1	1
AL22U10070	AL1G30400	0	4	4	0	0.6	1	0	0	0	1	1
AL4G23820	AL1G17130	1	4	3	0	0.7	1	0	0	0	1	1
AL6G22580	AL8G33950	0	3	3	0	0.6	1	0	0	0	1	0
AL1G60550	AL4G19070	0	2	2	0	0.6	1	0	0	0	1	1
AL4G31820	AL1G13440	0	12	12	0	0.9	1	1	0	1	0	1
AL1G63780	AL3G17590	0	10	10	0	0.9	1	1	0	1	0	1
AL1G29240	AL6G12550	0	10	10	1	1	1	1	0	1	0	1
AL701U10020	AL239U10010	0	9	9	0	0.7	1	1	0	0	0	0
AL126U10030	AL5G22300	3	7	4	0	0.7	1	1	0	0	0	1
AL8G23490	AL2G38760	0	6	6	0	0.9	1	1	0	1	0	1
AL3G42150	AL4G37910	0	6	6	0	0.9	1	1	0	0	0	1
AL5G13900	AL8G39460	0	5	5	0	0.8	1	1	0	1	0	0
AL9U11510	AL1G63720	2	5	3	0	0.6	1	1	0	0	0	0
AL2G15400	AL6G39910	0	4	4	0	0.8	1	1	0	1	0	1
AL4G45600	AL7G23280	0	4	4	0	0.6	1	1	0	1	0	1
AL6G22560	AL5G13210	0	4	4	0	0.8	1	1	0	0	0	0
AL2G20810	AL5G41680	2	4	2	0	0.6	1	1	0	1	0	1
AL6G41920	AL2G30020	0	3	3	0	0.6	1	1	0	1	0	1
AL5G31460	AL6G33790	0	3	3	0	0.6	1	1	0	0	0	0
AL7G44060	AL2G33620	0	2	2	0	0.8	1	1	0	1	0	1
AL1G17240	AL4G19410	0	2	2	0	0.7	1	1	0	1	0	1
AL5G16160	AL5G36700	0	2	2	1	0.8	1	1	0	1	0	0
AL3G40250	AL1G43450	0	2	2	0	0.6	1	1	0	0	0	0

5.5 Appendix E: Interspecies comparison of *NRPD2E2* DNA sequence

The alignment shows loss of *NRPD2E2* introns causing the structure of *NRPD2E2*$^{\mathrm{Aly\_MN47}}$ genomic DNA (Aly_NRPD2E2_gDNA) to match that of *NRPD2E2*$^{\mathrm{At\_Col}}$ cDNA (Ath_NRPD2E2_cDNA).

```
Ath_NRPD2E2_gDNA        --------------AATTTCTTCACTTCTCTTTGACTGCTTCG------CTTAACCACTGAAAAAGTGTGCCAAGGGTTTTCTACGTCGAATCT------------------CTCCGCAT 82
Ath_NRPD2E2_cDNA        --------------AATTTCTTCACTTCTCTTTGACTGCTTCG------CTTAACCACTGAAAAAGTGTGCCAAGGGTTTTCTACGTCGAATCT------------------CTCCGCAT 82
Aly_NRPD2E2_gDNA        CGTTTACTCTGCCTTCCTCCAACACCGCCGTTTTACTCCATCGTGCCAGCTTAAGCAATCAAGGTACCCATTTTAGGTATTACGCTTTGATTCTGCTTTTAAGCATTGGAAATTCCGGAG 120
                                      :. * *::***  *  *** *** *:***      ***** **.* **..:.   .  ::.***:**. .* * **:***                   **** *

Ath_NRPD2E2_gDNA        TCTCAG-----------------CGATTTCCGGCGACGTTAC---------TCTGCACTCCTCCGACACCG--CCGTTTTACTCCATCGTGCCAGCT--TTAAGCAATCAAGGTACCT 172
Ath_NRPD2E2_cDNA        TCTCAG-----------------CGATTTCCGGCGACGTTAC---------TCTGCACTCCTCCGACACCG--CCGTTTTACTCCATCGTGCCAGCT--TTAAGCAATCAAG------ 166
Aly_NRPD2E2_gDNA        ACTATATGCTTTAGAGAATGATTCGGTTCTAGGGGAAAGTTTTTGATTGCGTGTTTGTATTCGTATGATGCATTTTCGTGGTTCATGATTTTCACGGCTTCTTAATCTTTGTTTG----- 235
                        :**.:.                 **.** *. ** .*.****:          * ** * ** *. ** .*.    ***  *:*:  **  * .*.***  **** *::* ::

Ath_NRPD2E2_gDNA        ATTTTAGGTAATACGCTTTGATTCTGCTTTTAAGCATCGGAGAATATGTTATGGAGAATGATTCGGTTCTAAGGGAAAGTTGTTGATTTCGTGTTTGTATTCGCATGATTGCATTTTCGT 292
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------

Ath_NRPD2E2_gDNA        GGTTCATGATTTTCACAGCTTTTTAATCAATTTCTCTGTCTTTGTTTAGGGTTTTTGTTCG--TACAGTGTGTTTTGAGGTATGCCAGATAT---------------------------- 382
Ath_NRPD2E2_cDNA        -------------------------------------------------GGTTTTTGTTCG--TACAGTGTGTTTTGAGGTATGCCAGATAT---------------------------- 207
Aly_NRPD2E2_gDNA        ------------------------------------------------GGGTTTTTTTTTTGTTTCAGTGTGTTTTGAGGTATACCAGAAAAGATGGACTATATTGTTGAACGGAATTAA 307
                                                                            ******* ** *:*****************.*****:*:

Ath_NRPD2E2_gDNA        ---------------------GGACATTGATGTGAAGGATCTTGAAGAGTTCGAGGCTACTACTGGGGAGATCAATCTATCTGAGCTAGGAGAAGGTTTTCTGCAGAGTTTCTGCAAAAA 481
Ath_NRPD2E2_cDNA        ---------------------GGACATTGATGTGAAGGATCTTGAAGAGTTCGAGGCTACTACTGGGGAGATCAATCTATCTGAGCTAGGAGAAGGTTTTCTGCAGAGTTTCTGCAAAAA 306
Aly_NRPD2E2_gDNA        TTTTCTGTTACCAGAAAAGATGGACATTGATGAGATGGATATTGAAGAGATCGAGGCTACTGCG---GAGATCAATCTATCTGAGCTAGGAGAAAGTTTTCTCCAGAGTTTCTGCAAGAA 424
                                             ***********:**:****.********:***********.*    **************************.******* **************.**

Ath_NRPD2E2_gDNA        AGCTGCAACTTCTTTCTTTGATAAGTATGGACTTATAAGTCATCAGCTCAACTCCTACAACTACTTCATTGAACACGGGCTTCAGAATGTGTTTCAATCCTTTGGTGAGATGCTTGTGGA 601
Ath_NRPD2E2_cDNA        AGCTGCAACTTCTTTCTTTGATAAGTATGGACTTATAAGTCATCAGCTCAACTCCTACAACTACTTCATTGAACACGGGCTTCAGAATGTGTTTCAATCCTTTGGTGAGATGCTTGTGGA 426
Aly_NRPD2E2_gDNA        AGCTGCAACTTCCTTCTTTGATAAGTATGGACTTATAAGTCATCAGCTCAATTCCTACAACTTCTTCATTCAACACGGGCTTCAGGATGTGTTTGAATCCTTTGGTGATATGCTTGTGGA 544
                        **:******* **************.******** ************* ***********

Ath_NRPD2E2_gDNA        ACCGTCTTTTGATGTTGTAAAGAAGAAGGATAATGATTGGAGATACGCAACGGTGAAGTTCGGAGAAGTCACTGTGGAGAAGCCTACTTTCTTTTCGGATGACAAGGAGCTTGAGTTTCT 721
Ath_NRPD2E2_cDNA        ACCGTCTTTTGATGTTGTAAAGAAGAAGGATAATGATTGGAGATACGCAACGGTGAAGTTCGGAGAAGTCACTGTGGAGAAGCCTACTTTCTTTTCGGATGACAAGGAGCTTGAGTTTCT 546
Aly_NRPD2E2_gDNA        ACCGTCGTTTGATGTGATAAAGAAGAAGGATAACGATTGGAGATACGCTACGGTGAAATTCGGAAAAGTCACTGTGGAGAAGCCCACTTTCTTTTCCGATGACAAGGAGCTTGAGTTTCT 664
                        ****** ******** .*************** **************:********.******.****************** *********** **********************

Ath_NRPD2E2_gDNA        CCCATGGCATGCTAGGCTTCAGAACATGACATACTCTGCAAGGATCAAAGTCAATGTCCAAGTTGAGGTAACAGAAATTCTTTGTCGAAATTAAGTAACCTTGTCTGGATTTGATGAATG 841
Ath_NRPD2E2_cDNA        CCCATGGCATGCTAGGCTTCAGAACATGACATACTCTGCAAGGATCAAAGTCAATGTCCAAGTTGAG----------------------------------------------------- 613
Aly_NRPD2E2_gDNA        CCCATGGCATGCCAGGCTTCAGAACATGACATATTCAGCAAGGATCAAAGTCAATGTCCAAGTTGAG----------------------------------------------------- 731
                        ************ ******************* **:*****************************

Ath_NRPD2E2_gDNA        ATAAAGAACACATGGTATAAGCTTATTTCTTGATGTTTCTACTAGACTCTTTCTGACACATATATGAAGATGTTGACATACACTGAGGTTCCTGTCATAGATTTCTCAATTTAACTTGCC 961
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        ATCAAATAATTTACTAAGGGTTAAGGAACATATTTGTCTGAAACTGGTTTCACTCTTTTTGGCTTTACAAGTTTTCTGTAATTGGATTTGGTTCCTTATTTGCATTCGCTGGATTTCTTA 1081
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        CCTGAGCAAAATATCTAGTAAAAGAGATTTATTACAGTTACATGTTCGTGTGAAGTAGAGGTGTATTTCAAGCTTGGTTGTGTTTAAGATTGATGATTTTGTCTGCTCCCAATCTTTAGA 1201
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        TGTTTCTTGTTTAAAATTTGAATTGTGATTACTTTTCCTTGTAGTGGTGGGTATTCAAACGAAATAAGCTTTAGTTTGTTTCATTTTAAAGTTTGGATGCAATAAAAGAAAAACATCTTC 1321
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        AGCTTTTTTTTTTATTTAGTTCTTCCCCACTGCCTCACTGTGCTTTAGTTTGAGTGTTTTATGCTTGTGTGCAATGACTCTTGTACTGTCAAACTTTTGATGATGTTTCTGTTTTGCTGT 1441
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        CCATGTATCTTATTCTTATAAATGTAGTTTATTGTCTAACTGCTTCTTCACTCTATAAATTGACTAGGTGTTCAAGAATACTGTTGTTAAAAGCGACAAATTCAAGACAGGACAAGACAA 1561
Ath_NRPD2E2_cDNA        -------------------------------------------------------------------GTGTTCAAGAATACTGTTGTTAAAAGCGACAAATTCAAGACAGGACAAGACAA 666
Aly_NRPD2E2_gDNA        ----------------------------------------------------------------------GTAACAAAATCTTTGTCGAAAAATTAAGTAAGCTTGTCTGGATTTGATAA 781
                                                                                               *.*..**::** *    ****.  *..:*: *::*:*:*** ::** **

Ath_NRPD2E2_gDNA        CT-ATGTCGAG---AAGAAGATACTGGATGTCAA--------------AAAGCAGGACATTCTAATTGGTAGCATTCCTGTCATGGTGAAATCTATCCTTTGCAAAACAAGCGAGAAAG 1662
Ath_NRPD2E2_cDNA        CT-ATGTCGAG---AAGAAGATACTGGATGTCAA--------------AAAGCAGGACATTCTAATTGGTAGCATTCCTGTCATGGTGAAATCTATCCTTTGCAAAACAAGCGAGAAAG 767
Aly_NRPD2E2_gDNA        ATGATTTCCCTTGCTTGAAAACTCAGAAAGACCAGTTAACTATCACTTTTTAGTTCAACATTATGCAATATGTCTATGTAGTCGAGAGTAAGCTCATTTTTTGATGTTTCTACTAGACTC 901
                        .* ** **  .   ::***.* :*:*.*:*:*.*               ::** : .*****.*..:: .*. *::*  :***.:*.  **.   **  ****.:.:: .:.* ***.:

Ath_NRPD2E2_gDNA        GGAAAGAAAACTGCAAAAAGGG---GGATTGTGCCTTTGATCAGG--GTGGATATTTCGTGATAAAGGGGGCTGAGAAGGTGAGTTTAACTAATACATACATATATGCATATTGCCATTC 1777
Ath_NRPD2E2_cDNA        GGAAAGAAAACTGCAAAAAGGG---GGATTGTGCCTTTGATCAGG--GTGGATATTTCGTGATAAAGGGGGCTGAGAAG----------------------------------------- 841
Aly_NRPD2E2_gDNA        TTGCTGACACATATATGAAGATGTTGACATACACTGAGGTTCCTGTCATAGATTTCTCAAACTTATCAAAACCTTTAACTTG-------------------------------------- 983
                        ..:**.*..*. *:.***.    *..:*. .*  : *:**. *  .*.***:* **.:  ..*:*: ....*  : **

Ath_NRPD2E2_gDNA        AATACGTACAATAACTTTATTTTCTATGCCAAAACGGATTTTGTTTGTCAGCAATCCATAAAACGGATTATAGTTACAATTTTCTATCATCAGATAATAGTGTTTATCAGCAAAAAGATG 1897
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        TTACAAATTAGACAATAACCTATTTGGTTCATTTTATTTTCCTAATGGAGATGAAGTAAGAAATTAAGAAACTTAACTTATTTATGACTTTGTATACTTCGTTAGCATCAAAGATATAAA 2017
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        CTTTTTCTTTCCTTTCTGGCTTGACCATGAGGCCATGAGTATTCAAATCTTACAGGAAGCGTTCTTTGCAATCTTAGGCTCTGGGACAGATGATTTGACTCTAATATATTCTGGAAAAAA 2137
Ath_NRPD2E2_cDNA        ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA        ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA        TATTTTCAAGTTGTTATACTTCTCCCTAACGTTATTATATTGTTTTGCAGGTGTTTATAGCTCAAGAACAGATGTGCACAAAGAGACTGTGGATTTCTAATTCACCATGGACAGTCTCCT 2257
Ath_NRPD2E2_cDNA        --------------------------------------------------GTGTTTATAGCTCAAGAACAGATGTGCACAAAGAGACTGTGGATTTCTAATTCACCATGGACAGTCTCCT 911
Aly_NRPD2E2_gDNA        -------------------------------------------------CCATAAAATAATATATTAAGGGTTATGGCACATATATGTCTGGAAACTGGTTTCACTCTTTTTGG--CTTT 1052
                                                                           : :::***. : *: ** .*:*.** ...*:.:.: * ****::   .:***** .*  : .*     *

Ath_NRPD2E2_gDNA        TCAGGTCCGAAAATAAAAGAAATAGATTCATTGTGCGCCTCTCGGAGAATGAGAAAGCAGAAGACTATAAGAGAAGGGAGAAAGTACTGACAGTGTACTTCTTGTCGACTG-AGATTCCA 2376
Ath_NRPD2E2_cDNA        TCAGGTCCGAAAATAAAAGAAATAGATTCATTGTGCGCCTCTCGGAGAATGAGAAAGCAGAAGACTATAAGAGAAGGGAGAAAGTACTGACAGTGTACTTCTTGTCGACTG-AGATTCCA 1030
Aly_NRPD2E2_gDNA        ACAAGTTTTCTATTCTTGGATTTGGTTCCTTATTTGCATTCGCTGGATTTCTTACGTGAGCAAAATATCTAGTAAAAGAGATTTATTACATTTACATTTTCGTGTGAAGTAGAGGTATGT 1172
                        :**.**   .:*:*.::.**::*.*:* *:*: *   . ** * *..::* : *..  **.*.*.***.:.. **..****:: :: : * : : :: *** *** .* *. **.*:  :

Ath_NRPD2E2_gDNA        GTCTGGCTCCTCT-TCTTTGCGCTAGGTGTTTCGTCAGACAAAGAAGCCATGGAT-CTAATTGCTTTTGATGGTGATGATGCAAGCATTACCAACAGTCTCATAGCTTCTATCCATGTAG 2494
Ath_NRPD2E2_cDNA        GTCTGGCTCCTCT-TCTTTGCGCTAGGTGTTTCGTCAGACAAAGAAGCCATGGAT-CTAATTGCTTTTGATGGTGATGATGCAAGCATTACCAACAGTCTCATAGCTTCTATCCATGTAG 1148
Aly_NRPD2E2_gDNA        TTCAGGCTTCGTTGTTTTTAAGATTGATGATTTTGTCTGCTCCCAATCTTTAGATGTTTCTTGCTTTTTTT----CCGGGCCAAAATTTGAATTGTGATTACTTTTTCTTGTAGTAGTGG 1288
                         **:**** *  * * ***..*.*:*.**:**    . .*:.. ** * :*.***  *:.******** :*    . *.  ***..:**...:: :*: *..*:  *  *.*. ::**.*

Ath_NRPD2E2_gDNA        CTGATGCAGTTTGTGAAGCTTTTCGCTGTGGGAACAATGCTTTAACATATGTTG-AACAGCAGATCAAAAGCACCAAATTCCCTCCTGCTGAAAGTGTGGATGAGTGCCTCCATCTGTAT 2613
Ath_NRPD2E2_cDNA        CTGATGCAGTTTGTGAAGCTTTTCGCTGTGGGAACAATGCTTTAACATATGTTG-AACAGCAGATCAAAAGCACCAAATTCCCTCCTGCTGAAAGTGTGGATGAGTGCCTCCATCTGTAT 1267
Aly_NRPD2E2_gDNA        GTGCTCAAACGAAATAAGCTTTAGTTTGT---TTCATT--TTAAAGATTGGATGCAATAAAAGAAAAACATCTTCAGCTTTTTATTTATT-TAGTTCTTCCCCATTCCCTCACTGTGCTT 1402
                         **.* .*.  :.: *******:   ***   ::**:*  **:** **: *:** ** *..***:.**.* *: **..**   :  *. * :*. * *   .  * * ****..* ** :*

Ath_NRPD2E2_gDNA        TTGTTTCCAGGCCTCCAAAGTTTGAAGAAGAAAGCTCGATTCCTGGGCTATATGGTGAAGTGCCT-----TCTGAACTCGTATGCGGGAAAAAGAAAATGCGAAAACAG------GGACA 2722
Ath_NRPD2E2_cDNA        TTGTTTCCAGGCCTCCAAAGTTTGAAGAAGAAAGCTCGATTCCTGGGCTATATGGTGAAGTGCCT-----TCTGAACTCGTATGCGGGAAAAAGAAAATGCGAAAACAG------GGACA 1376
Aly_NRPD2E2_gDNA        TAATTTGAGTGTTTCATGCTTGTGTGCAATGACTCTTGTACTATCAAACTTTTGATGCTGTTTCTGTTTTGCTGTCCATGTATCTTATTCTTATAAATGTAGTTTATTGTCTAACTGCCT 1522
                        *:.*** .. *  **.:.. * **:. ** .*. ** *::  .* ... :*:**.**.:**  **      ***:.*: ****   . :.::* ***:  .*:::* :*      *.*:
```

```
Ath_NRPD2E2_gDNA      GTTTCCGGAATAAGCGAATTGAGCTCGCTGGAGAACTATTGGAGAGGGAGATAAGGGTGCATCTGGCACATGCTAGAAGAAAGATGACCAGGGCCATGCAGAAACACCTCTCAGGCGATG 2842
Ath_NRPD2E2_cDNA      GTTTCCGGAATAAGCGAATTGAGCTCGCTGGAGAACTATTGGAGAGGGAGATAAGGGTGCATCTGGCACATGCTAGAAGAAAGATGACCAGGGCCATGCAGAAACACCTCTCAGGCGATG 1496
Aly_NRPD2E2_gDNA      CTTCACTTTATAAATTCACTAGGTGTTCATAAAAACTGTTGTTAAAAGCGACAAATTCAAGACGGGACAAGACGAATATGTCGAGAAGAAGATACTTGAG--GTCAAAAAGCAGGACATT 1640
                       ** .*  :****.  .* *..*    *: .*.****.*** :.*..*.** **.    ...:* **...* .* *.:* .:.** .* .**. .*:**..  .:**..:. ****. **

Ath_NRPD2E2_gDNA      GTGATTTGAAGCCTATTGAGCATTATTTGGATGCTTCTGT-TATCACAAATGGGCTTAGTAGAGCCTTCTCTACTGGAGCATGGTCTCATCCTTTCAGGAAGATGGAAAGGGTTTCAGGT 2961
Ath_NRPD2E2_cDNA      GTGATTTGAAGCCTATTGAGCATTATTTGGATGCTTCTGT-TATCACAAATGGGCTTAGTAGAGCCTTCTCTACTGGAGCATGGTCTCATCCTTTCAGGAAGATGGAAAGGGTTTCAGGT 1615
Aly_NRPD2E2_gDNA      CTAATTGGTAGCATTCCTGTCATGGTGAAATCTGTCCTTTGCAAAACAAGCGAGAAAGGAAAAGAAAACTG--CAGAAAGGGGGATTGTGCCTTTGATCAGGGTGGCTAT--TTTGTGAT 1756
                       *.*** *:***.*:    . *** .* :..:    * ** *   *:.****. *.*.::.*:*.**..::**   *:*.*. . **: * : ***** *  *.*.***.:*   *** :*.*

Ath_NRPD2E2_gDNA      GTTGTGG-CTAATTTGGGTCGTGCAAATCCATTGCAGACTCTGATTGATCTG----AGGAGAACGCGACAGCAAGTCTTATATACCGGCAAGGTTGGAG-ATGCTAGATATCCGTAAGTG 3075
Ath_NRPD2E2_cDNA      GTTGTGG-CTAATTTGGGTCGTGCAAATCCATTGCAGACTCTGATTGATCTG----AGGAGAACGCGACAGCAAGTCTTATATACCGGCAAGGTTGGAG-ATGCTAGATATCCG------ 1723
Aly_NRPD2E2_gDNA      AAAGGGGGCTGAGAAGGTTAGT--TAAACTAATACATACATATATGCATATTGCCATTCAATACTTAAAATAAACTTTATTTTCTAAGCCAAAACGGATTTTGTTTGTCAGCAATTT--- 1871
                      .::* ** **.* ::** *.**  :**:* *:*.** **: : **  **.*     :  *.:**  .*.* .** * *::*:*. ..**.*..: ***  :** *:*: * *..

Ath_NRPD2E2_gDNA      AATTCCACCTCCTCTGGTATATTTAAATATATCTCACGTATTTTAACTTACTGGTCTGGTCTGCATTTACTCCATTTTACATGTCTTCAGACTGTTTTAAAGATATTCGTAATAACTTTA 3195
Ath_NRPD2E2_cDNA      ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA      ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA      ATTCAATGTAGTATATGATGCCGATCACTGTTTCTGCAGTCTCTTGTCTGTGTATAATACTTATTTTGTATAGATGTTACTGCTATTAAAAAACTCTGATACTGTCTTTCTTGTTTCTTT 3315
Ath_NRPD2E2_cDNA      ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA      ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA      CTCCCTTGGTCAATCTATCTGTTGAACTCTTGAGATTATCCATTTTGGTTCCTTTTCAATGTGAGCGGTTAGACAATTAAATCGTGTTGGGAAACTGAACTATAGCTGCATTGTTTGTAA 3435
Ath_NRPD2E2_cDNA      ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA      ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA      TGTTGGCCACAGGCACCCCTCTCACTGGGGCAGAGTATGCTTTTTGTCAACTCCAG------------ACGGTGAAAATTGTGGTCTTGTGAAGAACATGTCTCTTCTGGGACTTGTGAG 3543
Ath_NRPD2E2_cDNA      -------------CACCCCTCTCACTGGGGCAGAGTATGCTTTTTGTCAACTCCAG------------ACGGTGAAAATTGTGGTCTTGTGAAGAACATGTCTCTTCTGGGACTTGTGAG 1818
Aly_NRPD2E2_gDNA      ------------ATACAACACAAACAAGAGTATAATTACAATTTTCTATCATCAGATAATAGTAGTTATCAGCAAAAAAGATGTTACAAATTAGACAATAACCTATTTGGTTCATTTTAT 1979
                                   **..*:*:.**:.*.* * *.*::  .:**** *.:..**...            :*.* .****: .** *. :.: :***..**.:*  :* *** :*:* * *

Ath_NRPD2E2_gDNA      CACCCAAAGTTTGGAGTCTGTGGTGGAAAAGCTCTTCGCTTGTGGAATGGAAGAGCTGATGGATGATACATGCACACCATTGTTTGGCAAACATAAAGTTCTTCTCAATGGAGACTGGGT 3663
Ath_NRPD2E2_cDNA      CACCCAAAGTTTGGAGTCTGTGGTGGAAAAGCTCTTCGCTTGTGGAATGGAAGAGCTGATGGATGATACATGCACACCATTGTTTGGCAAACATAAAGTTCTTCTCAATGGAGACTGGGT 1938
Aly_NRPD2E2_gDNA      TTTCCTAATGGAGATGGTTAAGAAAGTAAGAAACTTAACTTATTTATGACTTTATATGCTTAAACATACATACAACAAAACCTTT-ATCAACACAAACACATACTGAAGGAAAGGAAAAA 2098
                       : **:**   :*.:*  *.:*.:.*:**...:***..***.*   *: . :: * .**.*.*: ******.**....*:  ***  . .**** *** : .*:** ** *.*.. :...:

Ath_NRPD2E2_gDNA      TGGATTATGTGCAG-ATTCTGAATCCTTTGTCGCGGAGTTAAAAAGCAGGCGGCGCCAAAGTGAATTACCTCGTGAGGTATCTTCTGTTTCAGCAAATCTCTTGCTATATTTTGATATTC 3782
Ath_NRPD2E2_cDNA      TGGATTATGTGCAG-ATTCTGAATCCTTTGTCGCGGAGTTAAAAAGCAGGCGGCGCCAAAGTGAATTACCTCGTGAG------------------------------------------- 2014
Aly_NRPD2E2_gDNA      CAATTCAAGTACTTCAAACTTCATTAAATACTAAACAAATTAATCGTGTTTGGGATCTTTTTTCATAGCCACTGTCA------------------------------------------- 2175
                       ..:* *:**.*:  *::** .**  .::*.   ... *.:*:**:.* .   ** . *::: * .**:.**:*   ..


Ath_NRPD2E2_gDNA      TTGTGTTACTTGGTATTTGCTTTGGATTTTTCTGCTTCAGATGTGTCTATGTCGAATATTGTTTATATATATGAAACGTTCTCTGCAGATGGAAATCAAGCGAGATAAAGATGACAATGA 3902
Ath_NRPD2E2_cDNA      ----------------------------------------------------------------------------------------ATGGAAATCAAGCGAGATAAAGATGACAATGA 2046
Aly_NRPD2E2_gDNA      ----------------------------------------------------------------------------------------TGGGATTTCTATTTTGAACTATTTTAGTG-GA 2206
                                                                                                              : ***::**:*   :**:.:* :* * :. **

Ath_NRPD2E2_gDNA      GGTAAGAATTTTCACTGATG----CTGGTAGACT-ACTCCGACCTCTCTTG---GTTGTGGAAAATCTCCAAAAGTTGAAGCAAGAAAAACCTTCACAGTATCCT--TTTGACCATCT-- 4010
Ath_NRPD2E2_cDNA      GGTAAGAATTTTCACTGATG----CTGGTAGACT-ACTCCGACCTCTCTTG---GTTGTGGAAAATCTCCAAAAGTTGAAGCAAGAAAAACCTTCACAGTATCCT--TTTGACCATCT-- 2154
Aly_NRPD2E2_gDNA      AATTAGTTTTCTCGCCGTTTTCCTTTGCTCAGCAGGCTCCGTCTTTGCTTGTCCTTTATGCATACTTTATTAGCATCAAAGATATAGACTTTTTCTTTCCTTTCTGGCTTGACCATGAGG 2326
                      ..*:**::** **.* *:*      ** *...*: .*****:* *  ****    **.** *:*.* *. :*...* .***.:* *.*.:  ***: :  :* **   ******** :

Ath_NRPD2E2_gDNA      TCTTGACCATGGGATTCTCGAGCTGATCGGGATTGAGGAAGAAGAAGACTGTAATACAGCATGGG-GAATCAAACAGCTTCTGAAGGAACCAAAGATATACACACATTGCGAATTGGACC 4129
Ath_NRPD2E2_cDNA      TCTTGACCATGGGATTCTCGAGCTGATCGGGATTGAGGAAGAAGAAGACTGTAATACAGCATGGG-GAATCAAACAGCTTCTGAAGGAACCAAAGATATACACACATTGCGAATTGGACC 2273
Aly_NRPD2E2_gDNA      CCATGACTATTCAAATCTT--ACAGGAAGCGTTCTTTGCAGTCATAGGCTCTGGGACAGATGACTTGACTCTGATATATACTG------CAAAAAATATTTTCAAGTT--GTTATACAAC 2436
                       *:**** **  .*:***   .*:*.:.* *:*  : *.**:..:**.** *.. ****.: .   **.**:.* *  .*:***      *.***.****: :**..**  *:::*. *.*

Ath_NRPD2E2_gDNA      TGTCATTCTTGTTGGGTGTGAGCTGTGCAGTTGTCCCATTTGCAAATCACGACCATGGGAGAAGAGT-TCTCTACCAGTCCCAGAAGCACTGCCAACAAGCCATTGGATTCTCATCAACG 4248
Ath_NRPD2E2_cDNA      TGTCATTCTTGTTGGGTGTGAGCTGTGCAGTTGTCCCATTTGCAAATCACGACCATGGGAGAAGAGT-TCTCTACCAGTCCCAGAAGCACTGCCAACAAGCCATTGGATTCTCATCAACG 2392
Aly_NRPD2E2_gDNA      TTCCTAACGTGATTATATTGTGTTTTGCAGGTG-----TTTATAGCTCAAGAACAGATGTGCACAAAGAGACTGTGGATTTCTAACTCACCATGGACAGTCTCTT----TCAGGTCCGAA 2547
                      *  *:::* **:* . : **:* * ***** **     ***. *..***.**.** . *:*.* *.: : :**.  ..*  *:.*. *** .  .***. * .**    **: .**....

Ath_NRPD2E2_gDNA      AACCCTAACATCCGCTGCGATACGCTGTCCCAGCAGCTGTTCTATCCTCAGAAGCCACTGTTCAAGACATTGGCGTCGGAGTGTCTTAAAAAAGAAGTGCTGTTCAATGGCCAGAACGCA 4368
Ath_NRPD2E2_cDNA      AACCCTAACATCCGCTGCGATACGCTGTCCCAGCAGCTGTTCTATCCTCAGAAGCCACTGTTCAAGACATTGGCGTCGGAGTGTCTTAAAAAAGAAGTGCTGTTCAATGGCCAGAACGCA 2512
Aly_NRPD2E2_gDNA      AATAAAAGAAATAGGTTCATTGTGCG--CCTCTCGGAGAATGAGAAATCAGAAGACTAT----AAGAAAAGGGAG----AAAGTACTGACAGTGTACTTCTTGTCGACTGAGATTCCAGT 2657
                      ** ..:*..*: .* * *.:*. **   ** . *.*. .:* :.:..*******.*:.*    ****.*: **.*    *.:**. *.*.*.:*:* * **  **.*  *. * :.*. :

Ath_NRPD2E2_gDNA      ATTGTTGCTGTGAATGTTCATCTCGGGTACAACCAAGAGGATTCCATTGTGATGAACAAGGCTTCATTGGAACGTGGTATG--TTCCGTTCAGAGCAGATTAGAAGCTACAAAGCAGAGG 4486
Ath_NRPD2E2_cDNA      ATTGTTGCTGTGAATGTTCATCTCGGGTACAACCAAGAGGATTCCATTGTGATGAACAAGGCTTCATTGGAACGTGGTATG--TTCCGTTCAGAGCAGATTAGAAGCTACAAAGCAGAGG 2630
Aly_NRPD2E2_gDNA      CTGGCTCCTGTTCTT----TGCGCTGGGTGTTTCGTCAG-ACAAAGAAGCCATGAATCTGATTGCTTTTGATGGTGATGATGCAAGCATTACCAACAGTCTCATAGCTTCTATCCA---- 2768
                      .* * * ****  .:*    : * * ** : :: *.: ** * :...::*  *****  .:*. * *:** **: ***.*.:   :: *.**.. *.***: *..:****:*:*: **

Ath_NRPD2E2_gDNA      TTGATGCTAAAGACTCAGAGAAGAGGAAGAAGATGGATGAGCTTGTTCAGTTTGGAAAGACACACAGCAAAATCGGCAAAGTAGACAGCCTTGAAGATG-ACGGGTTTCCTTTCATTGGT 4605
Ath_NRPD2E2_cDNA      TTGATGCTAAAGACTCAGAGAAGAGGAAGAAGATGGATGAGCTTGTTCAGTTTGGAAAGACACACAGCAAAATCGGCAAAGTAGACAGCCTTGAAGATG-ACGGGTTTCCTTTCATTGGT 2749
Aly_NRPD2E2_gDNA      -TGAAGCTGATG-CAGTTTGTGAAGCTTTTCGCTGTGGGAACAA-TGCTTTAAGTTATGTTGAACAGCAGATCAAACCTTGGAG---GCCTGGATGACAGGCAAGTATCTCTGACAAGCA 2882
                       ***:***.*:* *: : :*:..** :: :.*.** . **.*:: * *: *::* :*:*: ..******.*: ...*.::* **   **** **:** . .*..**:**  *  ..::* :

Ath_NRPD2E2_gDNA      GCTAACATGAGTACTGGCGATATTGTCATTGGCAGATGCACCGAGTCTGGGGCTGATCACAGTATAAAGCTCAAGCACACTGAGAGAGGAATTGTGCAAAAAGTGGTATTATCATCTAAT 4725
Ath_NRPD2E2_cDNA      GCTAACATGAGTACTGGCGATATTGTCATTGGCAGATGCACCGAGTCTGGGGCTGATCACAGTATAAAGCTCAAGCACACTGAGAGAGGAATTGTGCAAAAAGTGGTATTATCATCTAAT 2869
Aly_NRPD2E2_gDNA      AGTATCTCTG--ACAGGCAAAATAGAAGTG----AAAGCCCTGGTACAGAGATACTTGCCTGTCATAT-------ATCTCTGT---AAGACTAAAAAACTAAGAAGTTTCCAGGCCTCCA 2986
                      . **:*:  .   **:***.*:**:*:..*     .*:**.* *. :*:*.*. : :* .*:**.::*:       .:*:***:   *.**.*:.:..*.:***:.**:* .: . **..:

Ath_NRPD2E2_gDNA      GATGAAGGGAAGAATTTTGCTGCGGTTTCTCTGAGACAGGTAAGTTCCAGATCATACTAAATCGAGCTGTTTTTTCAGAGAATGCATTCCTATGTATGAATCGAATGTTCCATTGATTGG 4845
Ath_NRPD2E2_cDNA      GATGAAGGGAAGAATTTTGCTGCGGTTTCTCTGAGACAG---------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA      AGTACAGAGATATATGACAGGCAAGTATCTCT--GACAG---------------------------------------------------------------------------------
                      ..*..**.**:.:** : .   ..**:*****  *****

Ath_NRPD2E2_gDNA      CTTTTACATCTTACAGGTTCGTTCTCCATGCCTTGGAGATAAGTTTTCCAGTATGCATGGCCAGAAGGGTGTTTTAGGCTACCTAGAGGAACAGCAGAATTTTCCTTTCACGATCCAAGG 4965
Ath_NRPD2E2_cDNA      ----------------GTTCGTTCTCCATGCCTTGGAGATAAGTTTTCCAGTATGCATGGCCAGAAGGGTGTTTTAGGCTACCTAGAGGAACAGCAGAATTTTCCTTTCACGATCCAAGG 3012
Aly_NRPD2E2_gDNA      ------------------GCAAAATAGAAG---TGAGCATGAGTT-----ATATGACAGGCAAGTAT-CTCTGTAAGACTAAAAAACTAAGAAGTTCAATGTTCTCTGGTTGATTAATAC 3116
                                        *.::.*. *:*   **.. **.****     .****..:***.**:*   * * *:**.***..:*.. .*..** : *** ***  *  : *** .*:.

Ath_NRPD2E2_gDNA      ---CATAGTTCCTGATATTG---TGATAAACCCGCACGCTTTCCCTTCTAGGCAAA-CACCAGGTCAACTCTTGGAGGCTGCTCTCTCCAAAGGAATCGCTTGTCCTATACAAAAGGAGG 5078
Ath_NRPD2E2_cDNA      ---CATAGTTCCTGATATTG---TGATAAACCCGCACGCTTTCCCTTCTAGGCAAA-CACCAGGTCAACTCTTGGAGGCTGCTCTCTCCAAAGGAATCGCTTGTCCTATACAAAAGGAGG 3125
Aly_NRPD2E2_gDNA      TTCTATTGTTCCTGAAAAACGTCTAGAGAATACACAAAAATAGGCTCAAAAGCAATGTACCAGTATATAAATTAGTTAGAGGATTGATGCTGTGAGCCTTGTGATTTATGTCTGATTCAT 3236
                          **:********:*::    *..:.** .*.**...:*:  ** .:*.****:  ***** : *:.:.**.*: . :* : * :  .:. **. *   **:  ***. .:.*  ..


Ath_NRPD2E2_gDNA      GTAGCTCTGCTGCATACACCAAATTGACACGTCATGCCACTCCTTTCTCCACTCCGGGTGTCACTGAAATCACCGAGCAGCTTCACAGGTACATTCTTCACATTGTCTCTTGGTTTTAGC 5198
Ath_NRPD2E2_cDNA      GTAGCTCTGCTGCATACACCAAATTGACACGTCATGCCACTCCTTTCTCCACTCCGGGTGTCACTGAAATCACCGAGCAGCTTCACAG-------------------------------- 3213
Aly_NRPD2E2_gDNA      TTAACCTTTCT-----TAGATTATTGTTGATTCTTG--AGTCCTGATTCCATTAC------CAATGGTAAATATTTGTGGTTAG------------------------------------ 3307
                       **.*  * **      * .::****: .. **:**  * **** : **** *.*      **.**.:*:.:.  :* .* *:

Ath_NRPD2E2_gDNA      TCGTAAAACAGAATATAAAATTATATGCTATAACAGATTTACATTTGCTTCCTATACAAATAGATGATATCATTAAGGGCAGGAACATATTATTGATAATATTTCCTCGTTGAAGATGTT 5318
Ath_NRPD2E2_cDNA      ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA      ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA      TAAACTTGGAGACTTTGGCTACAGAATTTCCAAAAGTTGATTGAGCTAATATACTGCACAAGGCACTAAGCTAGATTTGAGCACCTTACTTGAAACATGTCATAGTGGATCTTTATTTTG 5438
Ath_NRPD2E2_cDNA      ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA      ------------------------------------------------------------------------------------------------------------------------
```

```
Ath_NRPD2E2_gDNA          TTTTCTAGGAGTACTAGAAGTGAGCATGAGTTATCTGTCTCTGTAAGACTAAAAAACTAAGAAGTTCAATGTTCTATGGTTGATTAATTTCTTGTATTGTGCCTGAAAAACGTCTAGAGA 5558
Ath_NRPD2E2_cDNA          ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA          ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA          ATACAGAAAAATAGGCTCAAGAGTCATGTACCAGTATATAATTAGTTAGAGGATTGATGCTGTGAGCCTTGTGATCTATGTATAATTCATTTAACCTTTCTTAGATTATTGTTGATTCTT 5678
Ath_NRPD2E2_cDNA          ------------------------------------------------------------------------------------------------------------------------
Aly_NRPD2E2_gDNA          ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA          GAGTCCTGATTCATTACCAATGGTAAATATTTGTGATTAGGGCCGGCTTTTCAAGATGGGGAAACGAAAGGGTCTACAACGGTAGATCAGGTGAGATGATGCGTTCTATGATATTCATGG 5798
Ath_NRPD2E2_cDNA          ----------------------------------------GGCCGGCTTTTCAAGATGGGGAAACGAAAGGGTCTACAACGGTAGATCAGGTGAGATGATGCGTTCTATGATATTCATGG 3293
Aly_NRPD2E2_gDNA          ----------------------------------------GGCCGGCTTTTCAAGATGGGGAAACGAAAGGGTCTACAATGGTAGATCGGGTGAGATGATGCGTTCTCTGATATTCATGG 3387
                                                                  *************************************** ********.******************.************

Ath_NRPD2E2_gDNA          GCCCAACTTTCTACCAGCGACTTGTCCACATGTCAGAGGACAAAGTCAAGTTCAGGAACACTGGACCAGTCCACCCGCTCACACGCCAGCCAGTTGCAGACAGGAAGAGATTTGGCGGGA 5918
Ath_NRPD2E2_cDNA          GCCCAACTTTCTACCAGCGACTTGTCCACATGTCAGAGGACAAAGTCAAGTTCAGGAACACTGGACCAGTCCACCCGCTCACACGCCAGCCAGTTGCAGACAGGAAGAGATTTGGCGGGA 3413
Aly_NRPD2E2_gDNA          GCCCAACTTTCTACCAGCGACTTGTCCACATGTCAGAGGACAAAGTCAAGTTCAGGAACACCGGACCAGTCCACCCGCTCACACGCCAGCAAGTCGCAGACAGGAAGAGGTTTGGCGGGA 3507
                          ************************************************************* ****************************.*** **************.**********

Ath_NRPD2E2_gDNA          TAAAATTTGGAGAAATGGAGCGAGACTGCCTAATAGCTCACGGTGCATCAGCTAATCTGCATGAGCGTCTCTTCACTCTAAGTGACTCTTCTCAGATGCACATCTGCAGAAAATGTAAGA 6038
Ath_NRPD2E2_cDNA          TAAAATTTGGAGAAATGGAGCGAGACTGCCTAATAGCTCACGGTGCATCAGCTAATCTGCATGAGCGTCTCTTCACTCTAAGTGACTCTTCTCAGATGCACATCTGCAGAAAATGTAAGA 3533
Aly_NRPD2E2_gDNA          TAAGGTTTGGAGAAATGGAGCGAGACTGCCTAATAGCTCACGGTGCATCTGCTAATCTGCACGAGCGTCTCTTCACTCTAAGTGACTCTTCTCAGATGCACATCTGCAGAAAATGTAAGA 3627
                          ***..********************************************:*********** **********************************************************

Ath_NRPD2E2_gDNA          CCTATGCGAATGTGATCGAGAGGACTCCAAGCAGTGGAAGAAAGATTAGAGGGCCATATTGTAGAGTCTGCGTATCCTCAGACCATGTGGTTAGGGTCTATGTTCCGTATGGAGCTAAGC 6158
Ath_NRPD2E2_cDNA          CCTATGCGAATGTGATCGAGAGGACTCCAAGCAGTGGAAGAAAGATTAGAGGGCCATATTGTAGAGTCTGCGTATCCTCAGACCATGTGGTTAGGGTCTATGTTCCGTATGGAGCTAAGC 3653
Aly_NRPD2E2_gDNA          CCTATGCGAATGTGATCGAGAGGACTCCAAGCAGTGGAAGAAAGATCAGAGGGCCATATTGTAGAGTCTGCGTATCCTCAGACCATGTGGTTAGAGTCTATGTTCCGTATGGAGCTAAAC 3747
                          ********************************************** ***********************************************.***********************.*

Ath_NRPD2E2_gDNA          TTCTGTGTCAGGAGCTGTTCAGCATGGGCATCACTCTCAACTTCGACACCAAGCTATGCTGATTCCCCCTCTTTATTATGTAAATGGCTTATTGCCTTAAGACCATGTTATGTGTAGTTT 6278
Ath_NRPD2E2_cDNA          TTCTGTGTCAGGAGCTGTTCAGCATGGGCATCACTCTCAACTTCGACACCAAGCTATGCTGATTCCCCCTCTTTATTATGTAAATGGCTTATTGCCTTAAGACCATGTTATGTGTAGTTT 3773
Aly_NRPD2E2_gDNA          TTCTGTGTCAGGAGCTGTTCAGCATGGGCATCACTCTCAACTTCGACACCAAGCTCTGCTGATTACCCCTCTTTATTATGTA-------------------------------------- 3829
                          *******************************************************.********.*****************

Ath_NRPD2E2_gDNA          GCTTCAGTCCCGGTTCTGGTTAGTAGTATAGGTTTTGGTTTGGTTGATTCGGTAAGGGTTATCCGAACCGAAGAAATCGTAAAACCGAGCCACTGATGACTGAACTAACCCGTAAGTGTT 6398
Ath_NRPD2E2_cDNA          GCTTCAGTCCCGGTTCTGGTTAGTAGTATAGGTTTTGGTTTGGTTGATTCGGTAAGGGTTATCCGAACCGAAGAAATCGTAAAACCGAGCCACTGATGACTGAACTAACCCGTAAGTGTT 3893
Aly_NRPD2E2_gDNA          ------------------------------------------------------------------------------------------------------------------------


Ath_NRPD2E2_gDNA          GCTTTTGTGAGATTTGACTCTTTAACCGTTAATAATTCTCGGATCTAAAGTAAAGTTTTAGG 6460
Ath_NRPD2E2_cDNA          GCTTTTGTGAGATTTGACTCTTTAACCGTTAATAATTCTCGGATCTAAAGTAAAGTTTTAGG 3955
Aly_NRPD2E2_gDNA          --------------------------------------------------------------
```

6. REFERENCES

Abdelsamad, A., and Pecinka, A. (2014). Pollen-specific activation of Arabidopsis retrogenes is associated with global transcriptional reprogramming. Plant Cell *26*, 3299-3313.

sAltschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature *408*, 796-815.

Bai, Y., Casola, C., and Betrán, E. (2008). Evolutionary origin of regulatory regions of retrogenes in Drosophila. BMC Genomics *9*, 1-9.

Baubec, T., Dinh, H.Q., Pecinka, A., Rakic, B., Rozhon, W., Wohlrab, B., von Haeseler, A., and Scheid, O.M. (2010). Cooperation of multiple chromatin modifications can generate unanticipated stability of epigenetic states in Arabidopsis. Plant Cell *22*, 34-47.

Baumbusch, L.O., Thorstensen, T., Krauss, V., Fischer, A., Naumann, K., Assalkhou, R., Schulz, I., Reuter, G., and Aalen, R.B. (2001). The *Arabidopsis thaliana* genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes. Nucleic Acids Res *29*, 4319-4333.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159-1170.

Blanc, G., and Wolfe, K.H. (2004). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell *16*, 1679-1691.

Borg, M., Brownfield, L., Khatab, H., Sidorova, A., Lingaya, M., and Twell, D. (2011). The R2R3 MYB transcription factor DUO1 activates a male germline-specific regulon essential for sperm cell differentiation in Arabidopsis. Plant Cell *23*, 534-549.

Borges, F., Calarco, J.P., and Martienssen, R.A. (2012). Reprogramming the epigenome in Arabidopsis pollen. Cold Spring Harbor Symp Quant Biol *77*, 1-5.

Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y., and Nurminsky, D.I. (2002). Large clusters of co-expressed genes in the Drosophila genome. Nature *420*, 666-669.

Bouyer, D., Roudier, F., Heese, M., Andersen, E.D., Gey, D., Nowack, M.K., Goodrich, J., Renou, J.-P., Grini, P.E., Colot, V.*, et al.* (2011). Polycomb repressive complex 2 controls the embryo-to-seedling phase transition. PLoS Genet *7*, e1002014.

Chartier-Harlin, M.C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M.*, et al.* (2004). Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. Lancet *364*, 1167-1169.

Chen, X. (2009). Small RNAs and their roles in plant development. Annual review of cell and developmental biology *25*, 21-44.

Clauss, M.J., and Koch, M.A. (2006). Poorly known relatives of Arabidopsis thaliana. Trends in plant science *11*, 449-459.

Coleman-Derr, D., and Zilberman, D. (2012). Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. PLoS Genet *8*, e1002988.

Comai, L. (2005). The advantages and disadvantages of being polyploid. Nat Rev Genet *6*, 836-846.

Cooke, S.L., Shlien, A., Marshall, J., Pipinikas, C.P., Martincorena, I., Tubio, J.M., Li, Y., Menzies, A., Mudie, L., Ramakrishna, M.*, et al.* (2014). Processed pseudogenes acquired somatically during cancer development. Nature communications *5*, 3644.

Creasey, K.M., Zhai, J., Borges, F., Van Ex, F., Regulski, M., Meyers, B.C., and Martienssen, R.A. (2014). miRNAs trigger widespread epigenetically activated siRNAs from transposons in Arabidopsis. Nature *508*, 411-415.

De Smet, R., Adams, K.L., Vandepoele, K., Van Montagu, M.C.E., Maere, S., and Van de Peer, Y. (2013). Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. Proc Natl Acad Sci USA *110*, 2898-2903.

Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol *3*, e314.

Deng, W., Maust, B.S., Nickle, D.C., Learn, G.H., Liu, Y., Heath, L., Kosakovsky Pond, S.L., and Mullins, J.I. (2010). DIVEIN: a web server to analyze phylogenies, sequence divergence, diversity, and informative sites. BioTechniques *48*, 405-408.

Ding, W., Lin, L., Chen, B., and Dai, J. (2006). L1 elements, processed pseudogenes and retrogenes in mammalian genomes. IUBMB life *58*, 677-685.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res *32*, 1792-1797.

Fablet, M., Bueno, M., Potrzebowski, L., and Kaessmann, H. (2009). Evolutionary origin and functions of retrogene introns. Mol Biol Evol *26*, 2147-2156.

Farrona, S., Thorpe, F.L., Engelhorn, J., Adrian, J., Dong, X., Sarid-Krebs, L., Goodrich, J., and Turck, F. (2011). Tissue-specific expression of FLOWERING LOCUS T in

Arabidopsis is maintained independently of polycomb group protein repression. Plant Cell *23*, 3204-3214.

Fawcett, J.A., Maere, S., and Van de Peer, Y. (2009). Plants with double genomes might have had a better chance to survive the Cretaceous–Tertiary extinction event. Proc Natl Acad Sci USA *106*, 5737-5742.

Fink, G.R. (1987). Pseudogenes in yeast? Cell *49*, 5-6.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol *183*, 557-564.

Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics *20*, 307-315.

Gregory, R.T., and Mable, B.K. (2005). Polyploidy in animals. In The evolution of the genome, R.T. Gregory, ed. (Elsevier), pp. 427-483.

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A.*, et al.* (2012). The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res *40*, D26-32.

Hirotsune, S., Yoshida, N., Chen, A., Garrett, L., Sugiyama, F., Takahashi, S., Yagami, K., Wynshaw-Boris, A., and Yoshiki, A. (2003). An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature *423*, 91-96.

Hoffmann, R.D., and Palmgren, M.G. (2013). Epigenetic repression of male gametophyte-specific genes in the Arabidopsis sporophyte. Mol Plant *6*, 1176-1186.

Honys, D., and Twell, D. (2003). Comparative analysis of the Arabidopsis pollen transcriptome. Plant Physiol *132*, 640-652.

Honys, D., and Twell, D. (2004). Transcriptome analysis of haploid male gametophyte development in Arabidopsis. Genome Biology *5*, R85.

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.*, et al.* (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet *43*, 476-481.

Ibarra, C.A., Feng, X., Schoft, V.K., Hsieh, T.-F., Uzawa, R., Rodrigues, J.A., Zemach, A., Chumak, N., Machlicova, A., Nishimura, T.*, et al.* (2012). Active DNA demethylation in plant companion cells reinforces transposon methylation in gametes. Science *337*, 1360-1364.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat Rev Genet *11*, 97-108.

Irimia, M., Rukov, J.L., Penny, D., Vinther, J., Garcia-Fernandez, J., and Roy, S.W. (2008). Origin of introns by 'intronization' of exonic sequences. Trends in genetics : TIG *24*, 378-381.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer‐Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics *4*, 249-264.

Ishizaki, K., Johzuka-Hisatomi, Y., Ishida, S., Iida, S., and Kohchi, T. (2013). Homologous recombination-mediated gene targeting in the liverwort Marchantia polymorpha L. Scientific reports *3*, 1532.

Jeffares, D.C., Penkett, C.J., and Bahler, J. (2008). Rapidly regulated genes are intron poor. Trends in genetics : TIG *24*, 375-378.

Jelesko, J.G., Harper, R., Furuya, M., and Gruissem, W. (1999). Rare germinal unequal crossing-over leading to recombinant gene formation and gene duplication in Arabidopsis thaliana. Proc Natl Acad Sci U S A *96*, 10302-10307.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res *20*, 1313-1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat Rev Genet *10*, 19-31.

Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J. (2005). Atypical RNA polymerase subunits required for RNA-directed DNA methylation. Nat Genet *37*, 761-765.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol *14*, R36.

Konrad, A., Teufel, A.I., Grahnen, J.A., and Liberles, D.A. (2011). Toward a general model for the evolutionary dynamics of gene duplicates. Genome Biol Evol *3*, 1197-1209.

Lafos, M., Kroll, P., Hohenstatt, M.L., Thorpe, F.L., Clarenz, O., and Schubert, D. (2011). Dynamic regulation of H3K27 trimethylation during *Arabidopsis* differentiation. PLoS Genet *7*, e1002040.

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

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., *et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Liu, C., Lu, F., Cui, X., and Cao, X. (2010). Histone methylation in higher plants. Annu Rev Plant Biol *61*, 395-420.

Loraine, A.E., McCormick, S., Estrada, A., Patel, K., and Qin, P. (2013). RNA-seq of Arabidopsis pollen uncovers novel transcription and alternative splicing. Plant Physiol *162*, 1092-1109.

Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. Journal of structural and functional genomics *3*, 35-44.

Ma, Z., Coruh, C., and Axtell, M.J. (2010). Arabidopsis lyrata small RNAs: transient MIRNA and small interfering RNA loci within the Arabidopsis genus. Plant Cell *22*, 1090-1103.

Marques, A.C., Dupanloup, I., Vinckenbosch, N., Reymond, A., and Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. PLoS Biol *3*, e357.

Monk, D., Arnaud, P., Frost, J.M., Wood, A.J., Cowley, M., Martin-Trujillo, A., Guillaumet-Adkins, A., Iglesias Platas, I., Camprubi, C., Bourc'his, D., *et al.* (2011). Human imprinted retrogenes exhibit non-canonical imprint chromatin signatures and reside in non-imprinted host genes. Nucleic Acids Res *39*, 4577-4586.

Mooney, M., Bond, J., Monks, N., Eugster, E., Cherba, D., Berlinski, P., Kamerling, S., Marotti, K., Simpson, H., Rusk, T., *et al.* (2013). Comparative RNA-Seq and Microarray Analysis of Gene Expression Changes in B-Cell Lymphomas of <italic>Canis familiaris</italic>. PLoS ONE *8*, e61088.

Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J., and Baulcombe, D.C. (2009). Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. Nature *460*, 283-286.

Narsai, R., Howell, K.A., Millar, A.H., O'Toole, N., Small, I., and Whelan, J. (2007). Genome-wide analysis of mRNA decay rates and their determinants in *Arabidopsis thaliana*. Plant Cell *19*, 3418-3436.

Ohshima, K. (2013). RNA-Mediated Gene Duplication and Retroposons: Retrogenes, LINEs, SINEs, and Sequence Specificity. Int J Evol Biol *2013*, 424726.

Okamura, K., and Nakai, K. (2008). Retrotransposition as a source of new promoters. Mol Biol Evol *25*, 1231-1238.

Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. Cell *120*, 613-622.

Paradis, E. (2010). pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics *26*, 419-420.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M.*, et al.* (2012). The GENCODE pseudogene resource. Genome Biol *13*, R51.

Pennisi, E. (2012). ENCODE project writes eulogy for junk DNA. Science *337*, 1159-1161.

Pink, R.C., Wicks, K., Caley, D.P., Punch, E.K., Jacobs, L., and Carter, D.R. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? Rna *17*, 792-798.

Popova, O.V., Dinh, H.Q., Aufsatz, W., and Jonak, C. (2013). The RdDM pathway is required for basal heat tolerance in Arabidopsis. Mol Plant *6*, 396-410.

Potrzebowski, L., Vinckenbosch, N., Marques, A.C., Chalmel, F., Jégou, B., and Kaessmann, H. (2008). Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol *6*, e80.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Ream, T.S., Haag, J.R., Wierzbicki, A.T., Nicora, C.D., Norbeck, A.D., Zhu, J.K., Hagen, G., Guilfoyle, T.J., Pasa-Tolic, L., and Pikaard, C.S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. Molecular cell *33*, 192-203.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of molecular biology *314*, 1041-1052.

Roudier, F., Ahmed, I., Berard, C., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L.*, et al.* (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO J *30*, 1928-1938.

Roy, S.W., and Irimia, M. (2009). Mystery of intron gain: new data and new models. Trends in genetics : TIG *25*, 67-73.

Sakai, H., Mizuno, H., Kawahara, Y., Wakimoto, H., Ikawa, H., Kawahigashi, H., Kanamori, H., Matsumoto, T., Itoh, T., and Gaut, B.S. (2011). Retrogenes in rice (Oryza sativa L. ssp. japonica) exhibit correlated expression with their source genes. Genome Biol Evol *3*, 1357-1368.

Sakharkar, M.K., Chow, V.T.K., and Kangueane, P. (2004). Distributions of exons and introns in the human genome. In Silico Biol *4*, 387-393.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat Genet *37*, 501-506.

Seymour, D.K., Koenig, D., Hagmann, J., Becker, C., and Weigel, D. (2014). Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. PLoS Genet *10*, e1004785.

Simillion, C., Janssens, K., Sterck, L., and Van de Peer, Y. (2008). i-ADHoRe 2.0: an improved tool to detect degenerated genomic homology using genomic profiles. Bioinformatics *24*, 127-128.

Slotkin, R.K., Vaughn, M., Borges, F., Tanurdžić, M., Becker, J.D., Feijó, J.A., and Martienssen, R.A. (2009). Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell *136*, 461-472.

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.*, et al.* (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet *45*, 831-835.

Song, W.-Y., Choi, K.S., Kim, D.Y., Geisler, M., Park, J., Vincenzetti, V., Schellenberg, M., Kim, S.H., Lim, Y.P., Noh, E.W.*, et al.* (2010). Arabidopsis PCR2 is a zinc exporter involved in both zinc extrusion and long-distance zinc transport. Plant Cell *22*, 2237-2252.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G., Korf, I., Lapp, H.*, et al.* (2002). The Bioperl toolkit: Perl modules for the life sciences. Genome Res *12*, 1611-1618.

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics *24*, 637-644.

Stanke, M., Tzvetkova, A., and Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol *7 Suppl 1*, S11 11-18.

Stroud, H., Greenberg, Maxim V.C., Feng, S., Bernatavichute, Yana V., and Jacobsen, Steven E. (2013). Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. Cell *152*, 352-364.

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T.Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L.*, et al.* (2008). The Arabidopsis Information

Resource (TAIR): gene structure and function annotation. Nucleic Acids Res *36*, D1009-1014.

Szczesniak, M.W., Ciomborowska, J., Nowak, W., Rogozin, I.B., and Makalowska, I. (2011). Primate and rodent specific intron gains and the origin of retrogenes with splice variants. Mol Biol Evol *28*, 33-37.

Tsukahara, S., Kawabe, A., Kobayashi, A., Ito, T., Aizu, T., Shin-i, T., Toyoda, A., Fujiyama, A., Tarutani, Y., and Kakutani, T. (2012). Centromere-targeted de novo integrations of an LTR retrotransposon of Arabidopsis lyrata. Genes & development *26*, 705-713.

Tucker, S.L., Reece, J., Ream, T.S., and Pikaard, C.S. (2010). Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. Cold Spring Harb Symp Quant Biol *75*, 285-297.

Vaucheret, H., and Fagard, M. (2001). Transcriptional gene silencing in plants: targets, inducers and regulators. Trends in genetics : TIG *17*, 29-35.

Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Scie USA *103*, 3220-3225.

Wang, W., Zheng, H., Fan, C., Li, J., Shi, J., Cai, Z., Zhang, G., Liu, D., Zhang, J., Vang, S.*, et al.* (2006). High rate of chimeric gene origination by retroposition in plant genomes. Plant Cell *18*, 1791-1802.

Wang, Y., and Ma, H. (2011). Development: a pathway to plant female germ cells. Current biology : CB *21*, R476-478.

Yenerall, P., Krupa, B., and Zhou, L. (2011). Mechanisms of intron gain and loss in Drosophila. BMC Evol Biol *11*, 364.

Yogeeswaran, K., Frary, A., York, T.L., Amenta, A., Lesser, A.H., Nasrallah, J.B., Tanksley, S.D., and Nasrallah, M.E. (2005). Comparative genome analyses of Arabidopsis spp.: inferring chromosomal rearrangement events in the evolutionary history of A. thaliana. Genome Res *15*, 505-515.

Yoshida, S., Maruyama, S., Nozaki, H., and Shirasu, K. (2010). Horizontal gene transfer by the parasitic plant *Striga hermonthica*. Science *328*, 1128.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends Ecol Evol *18*, 292-298.

Zhang, Y., Wu, Y., Liu, Y., and Han, B. (2005). Computational identification of 69 retroposons in Arabidopsis. Plant Physiol *138*, 935-948.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., and Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. PLoS ONE *9*, e78644.

Zhu, X., Ma, H., and Chen, Z. (2011). Phylogenetics and evolution of Su(var)3-9 SET genes in land plants: rapid diversification in structure and function. BMC Evol Biol *11*, 1-11.

Zhu, Z., Zhang, Y., and Long, M. (2009). Extensive structural renovation of retrogenes in the evolution of the Populus genome. Plant Physiol *151*, 1943-1951.

7. CURRICULUM VITAE

PERSONAL INFORMATION

Ahmed Abdelsamad

📍 14 Kolibriweg, Cologne – 50829, Germany

☎ +49 (0) 176 358 727 96

✉ ahmed.m.abdelsamad@gmail.com

Gender Male | Date of birth 06/07/1986 | Nationality Egyptian

EDUCATION AND TRAINING

2011 – 2014

PhD in Genome and Epigenome Evolution Group

Max Planck Institute for Plant Breeding Research / University of Göttingen, Germany.

- Thesis: Evolution and epigenetic regulation of RNA-duplicated genes in *Arabidopsis*

Oct. 2009 – Mar. 2011

MSc in Molecular Biology

Max Planck Institute for Biophysical chemistry / University of Göttingen, Germany.

- Thesis: MicroRNA role in the regenerating pancreas of the mouse.

Sept. 2003 – July 2007

Honours BSc in Biotechnology

Faculty of Agriculture, Cairo University, Egypt.

- Genetics, biochemistry, microbiology and ecology.

WORK EXPERIENCE

Apr. 2012 – present

Assistant lecturer

Department of Genetics, Faculty of Agriculture, Cairo University, Egypt.

June 2008 – Mar. 2012

Teaching and research assistant

Department of Genetics, Faculty of Agriculture, Cairo University, Egypt.

June 2007 – May 2008

Research assistant

Agricultural genetic engineering research institute, Agricultural research center, Egypt.

RELATED PUBLICATIONS

V. Rawat, **A. Abdelsamad**, B. Pietzenuk, D. Seymour, D. Koenig, D. Weigel, A. Pecinka and K. Schneeberger. Improving the annotation of *Arabidopsis lyrata* using RNA-seq data 2015 (*Submitted to BMC Genomics*).

A. Abdelsamad, A. Pecinka. Pollen-specific transcription of Arabidopsis retrogenes is associat with global transcriptional reprogramming. (2014) *The Plant Cell*, vol. 26 (8), pp.3299-3313.

Göttingen, 28.05.2015

Signature ……………..………