Gametophytic Selection in *Arabidopsis thaliana* Supports the Selective Model of Intron Length Reduction

Cathal Seoighe[1*], **Chris Gehring**[2], **Laurence D. Hurst**[3]

1 Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Rondebosch, South Africa, **2** Department of Biotechnology, University of the Western Cape, Bellville, South Africa, **3** Department of Biology and Biochemistry, University of Bath, Somerset, United Kingdom

Why do highly expressed genes have small introns? This is an important issue, not least because it provides a testing ground to compare selectionist and neutralist models of genome evolution. Some argue that small introns are selectively favoured to reduce the costs of transcription. Alternatively, large introns might permit complex regulation, not needed for highly expressed genes. This "genome design" hypothesis evokes a regionalized model of control of expression and hence can explain why intron size covaries with intergene distance, a feature also consistent with the hypothesis that highly expressed genes cluster in genomic regions with high deletion rates. As some genes are expressed in the haploid stage and hence subject to especially strong purifying selection, the evolution of genes in *Arabidopsis* provides a novel testing ground to discriminate between these possibilities. Importantly, controlling for expression level, genes that are expressed in pollen have shorter introns than genes that are expressed in the sporophyte. That genes flanking pollen-expressed genes have average-sized introns and intergene distances argues against regional mutational biases and genomic design. These observations thus support the view that selection for efficiency contributes to the reduction in intron length and provide the first report of a molecular signature of strong gametophytic selection.

Citation: Seoighe C, Gehring C, Hurst LD (2005) Gametophytic selection in *Arabidopsis thaliana* supports the selective model of intron length reduction. PLoS Genet 1(2): e13.

Introduction

Selection for efficiency has been proposed to explain the reduced intron lengths of broadly or highly expressed genes in several animal systems [1–5]. Because of the energetic cost of transcription [4–6], which is proportional to the length of the transcript and the amount of the transcript that is produced, highly expressed genes are likely to experience greater selective pressure for a reduction in transcript length. This model sees long introns in weakly expressed genes as the result of weakened negative selection. This interpretation of the negative correlation between intron size and gene expression level [1–5] has recently been challenged. The genomic design hypothesis suggests that the shorter introns of highly expressed genes may not be the result of purifying selection, but instead reflect a reduced level of epigenetic regulation in housekeeping genes, which are often expressed at high levels [7]. Under this hypothesis, selection actively favours the accumulation of longer introns in less highly expressed genes because many of these genes are tissue specific and require greater levels of epigenetic regulation. This is supported by the fact that intergenic distances also tend to be reduced in the vicinity of highly expressed genes [2,7,8], an observation that is not explained by the transcriptional efficiency model. Moreover, if one controls for intergene distance, it is as yet unclear whether, in humans, highly expressed genes have small introns as reports are contradictory [2,7]. Hence, the relevance of the transcriptional efficiency model is currently uncertain. The correlation between intergene distance and intronic size has also been interpreted as evidence for a regional mutational bias, coupled with neutral evolution [2]. Indeed, regions of high compaction tend to be GC rich [2,8] and hence regions of high recombination rates [9]. If recombination induces deletions, then a simple mutational bias/neutral model can be considered.

Owing to the fact that it has abundant genes that are haploid expressed, *Arabidopsis thaliana* provides a novel testing ground to examine these conflicting viewpoints. Strong selection at the gametophytic stage, owing to haploid exposure of recessive mutations and/or to strong pollen competition [10–12], has been proposed as a key aspect of plant evolutionary biology resulting in the purging of deleterious mutations in genes that are transcribed in the growing pollen tube [13,14]. A transcriptional cost view of intron length variation predicts that this strong purifying selection should cause a reduction in intron lengths in genes that are expressed in pollen compared to genes that are expressed elsewhere and that this reduction should be most pronounced in the most highly expressed genes.

Results/Discussion

Introns, particularly those toward the 5′ ends of genes, may often have regulatory functions [15]. The lengths of introns in

Received March 25, 2005; Accepted May 31, 2005; Published August 5, 2005
DOI: 10.1371/journal.pgen.0010013

Abbreviations: bp, base pair; SAGE, serial analysis of gene expression

Editor: Greg Gibson, North Carolina State University, United States of America

* To whom correspondence should be addressed. E-mail: cathal@science.uct.ac.za

Synopsis

Genes are odd things. Small proteins are often encoded by big genes. In the process, much of the excess material has to be cut out and thrown away. The size of the parts that are discarded (introns) differs greatly between genes. Why should this be so? The authors test three different ideas, making use of the unusual fact that in plants genes are expressed in pollen. As pollen has only one copy of every gene, natural selection is expected to work somewhat better. The authors find that the non-coding parts of genes that are especially active in pollen are particularly small. They also find that being active in pollen tends to make introns small. This provides strong support for the idea that small introns are the result of selection to reduce costs of making too much material that is only going to be thrown away.

animals have been shown to decrease as a function of the position of the intron, counting from the 5′ end, and to depend to some extent on the breadth of expression (i.e., the number of tissues in which the gene is expressed [5]). We find a similar reduction in intron length as a function of intron position in *Arabidopsis* (Figure 1). In order to reduce the impact of positional effects and regulatory elements associated with proximal introns, we restricted our analysis of intron lengths of genes that are expressed in pollen to distal introns (from intron 5 to intron 10) since they are less likely to have a role in regulation (our analyses were not sensitive to the cut-off used to classify distal introns).

Using publicly available serial analysis of gene expression (SAGE) data, we compared intron lengths between genes that are expressed in pollen and the sporophyte. A summary of the dataset is provided in Table 1. The average intron length for the pollen genes was 107.7 base pairs (bp), compared to 123.4 bp for introns from genes expressed in at least one of four sporophyte conditions ($p = 0.0002$). In spite of significant differences in means, the mode of the distribution of intron lengths remained approximately the same in both groups and for all intron positions. Comparison of the distributions of intron lengths shows that there were fewer longer introns among the genes that were expressed in the gametophyte compared to the sporophyte, as indicated by curvature away from the diagonal in a quantile–quantile plot (Figure 2). We also compared intron lengths between genes expressed in the sporophyte and gametophyte with expression level as a covariate, using expression levels from the pollen SAGE dataset and the largest [16] of the four sporophyte SAGE datasets in the study. We found significant evidence for both a negative correlation between intron length and gene expression level ($p = 0.01$) and a reduced intron length in genes expressed at a given level in the gametophyte compared to the sporophyte ($p = 0.001$). This latter result suggests that introns from genes expressed in pollen remain shorter than introns from genes in the sporophyte when we control for expression level.

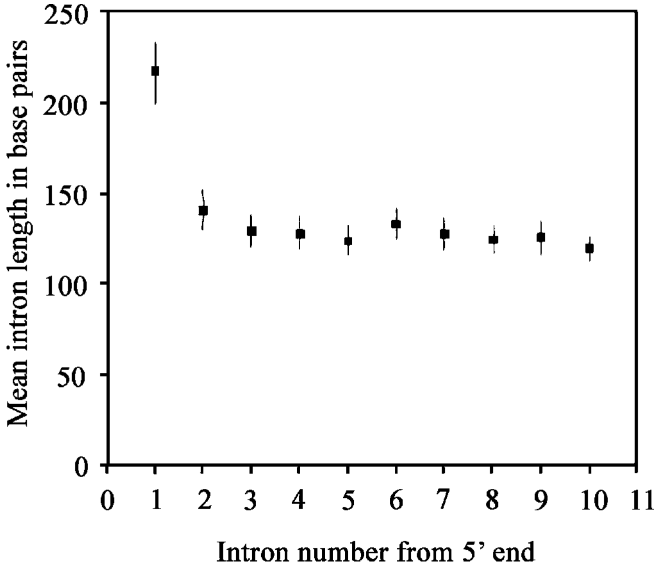

Figure 1. Mean Intron Length as a Function of Intron Position, Counting from the 5′ End of the Gene

Intron length was nearly constant for introns 5 to 10. Proximity to the 3′ end of the gene was not correlated with intron length. Error bars show ± twice the standard error. The data shown are for genes with exactly ten introns so that positional effects from the 3′ ends can also be assessed.

DOI: 10.1371/journal.pgen.0010013.g001

Might the reduced intron lengths of genes expressed in pollen be sensitive to the method of measurement of gene expression? We compared intron lengths between genes that are expressed in pollen but not in the sporophyte and vice versa using microarray data from the Expression Atlas of *Arabidopsis* Development [17]. The mean intron lengths for pollen-specific genes was 109.4 bp compared to 134.7 bp for the genes expressed in the sporophyte but not in pollen ($p = 3 \times 10^{-9}$). The expression level of the pollen-specific genes was higher, on average, than for genes that were expressed in pollen and the sporophyte. If expression level in pollen is included as a covariate, the length of introns remained significantly lower in genes that are specific to pollen compared to genes that are specific to sporophyte ($p = 5 \times 10^{-5}$). Introns from genes that were highly expressed in pollen were also significantly shorter than introns from genes that were highly expressed in at least one sporophyte sample, regardless of whether the gene was specific to pollen or expressed in both pollen and sporophyte ($p = 0.0007$).

The reduction in intron lengths in genes expressed in the pollen SAGE dataset did not appear to be affected by whether the genes were also expressed in the sporophyte, illustrating the potential impact of strong gametophytic selection on sporophyte evolution. In the SAGE dataset, genes that were specific to pollen and genes that were expressed in pollen as well as one of the sporophyte datasets had similar average intron lengths (99.1 bp and 109.7 bp; $n = 13$ and $n = 58$, respectively; $p = 0.93$), while in both cases the introns were significantly or marginally significantly shorter than introns of genes expressed in the sporophyte but not expressed in pollen ($p = 0.06$ and $p = 0.0009$). This additionally provides evidence that the observed difference in intron lengths between genes expressed in pollen and the sporophyte is not the result of a lack of intronic regulatory elements in genes that are expressed exclusively in pollen. Contrary to the results from SAGE, the reduction in intron lengths was confined to genes that were specific to pollen in the microarray datasets, possibly due to hybridisation cross-reactivity between homologous genes. Pollen has a high proportion of genes that appear to be expressed in pollen only [18]. The reduced intron lengths observed in such genes is not in keeping with the genomic design argument that

Table 1. Summary of the Dataset

Expression	Number of Genes	Length of Distal Introns	Number of Introns	Total Exonic Length	Intron Density/kb
In pollen[a]	245	107.7	3.5	1,409.8	2.4
Pollen but not sporophyte[a]	190	109.7	3.6	1,452.2	2.3
In at least one sporophyte sample[a]	2,176	123.4	4.2	1,392.9	3.0
Pollen[b]	6,713	130.4	5.4	1,648.3	3.3
Pollen but not sporophyte[b]	641	109.4	3.6	1,470.0	2.3
In at least one sporophyte sample[b]	15,390	133.3	4.8	1,642.6	2.9

Each cell represents the mean value of the quantity in the column for the subset of genes indicated in the row. The complete dataset used is available as Dataset S1 with this article.

[a]SAGE data

[b]Microarray data

DOI: 10.1371/journal.pgen.0010013.t001

suggests that regulation of narrowly expressed genes is responsible for their longer introns compared to broadly and highly expressed genes.

To test whether altered rates of insertion or deletion or a higher gene density in the genomic regions containing the genes that are expressed in pollen could be responsible for the reduced intron lengths, we calculated the average intron lengths of the closest genomic neighbours of the pollen genes from the SAGE dataset. The mean intron length of the neighbouring genes was not significantly different to the mean for all genes (132.5 bp compared to 134.8 bp, $p = 0.14$). The mean intron length for genes expressed in pollen remained significantly below the mean for genes expressed in the sporophyte, considering only the closest sporophyte-expressed neighbour for each gene expressed in pollen ($p = 0.01$). Thus, regional genomic effects evoked by the genomic design hypothesis [7] and the mutational bias hypothesis are not likely to be the cause of the reduced

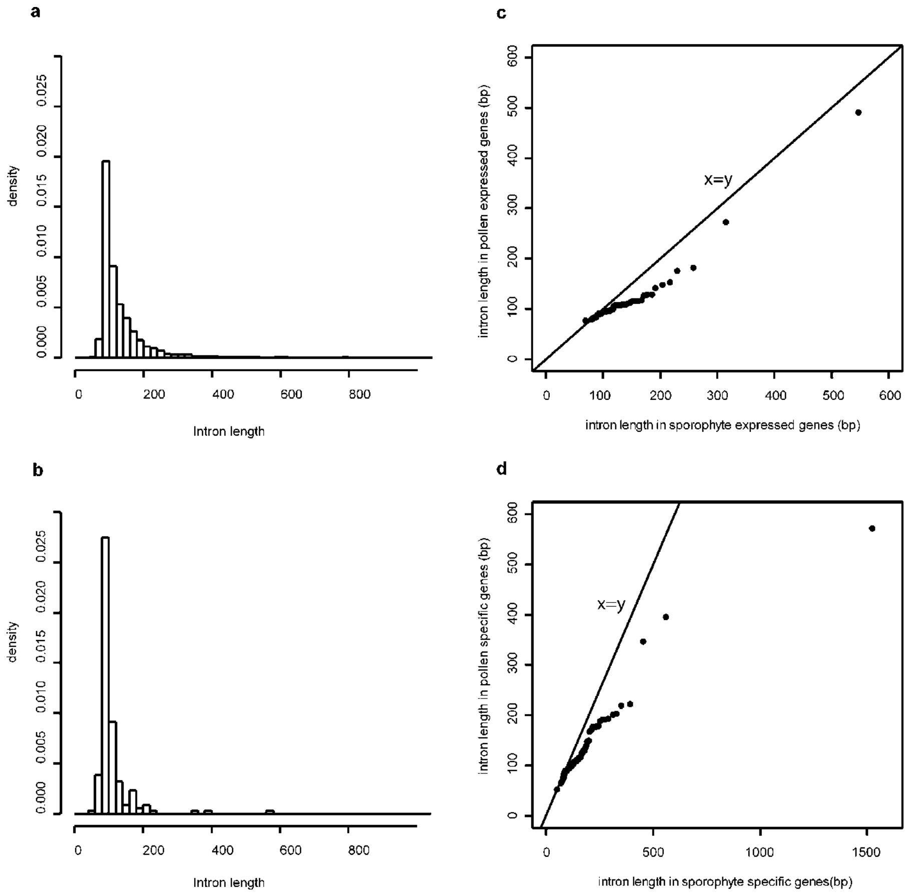

Figure 2. Histograms and Quantile–Quantile Plots of Mean Distal Intron Length

(A) Histogram for genes expressed in the sporophyte microarray datasets.

(B) Histogram for genes expressed in the gametophyte but not the sporophyte microarray datasets.

(C) Quantile–quantile plot of introns from all pollen-expressed and all sporophyte-expressed genes derived from the SAGE dataset. Quantiles of the intron length distributions for genes expressed in the gametophyte and sporophyte are on the *x*- and *y*-axes, respectively.

(D) Quantile–quantile plot of introns from pollen-specific and sporophyte-specific genes derived from the microarray dataset.

DOI: 10.1371/journal.pgen.0010013.g002

intron lengths of genes expressed in pollen. Furthermore, although the mean length of flanking regions was slightly greater for genes that were expressed in the sporophyte compared to pollen, the difference was not statistically significant (1,946 bp and 1,762 bp, for sporophyte and pollen, respectively; $p = 0.57$). Restricting to genes with at least five introns (the genes that contributed to this study), this difference is reduced, and the pollen genes, in fact, have slightly longer intergenic regions, although, again, the difference is not statistically significant (1,826 bp and 1,913 bp for sporophyte and pollen, respectively; $p = 0.15$).

The introns of genes that were highly expressed in at least one of the sporophyte expression sites in the study were significantly reduced in length compared to all genes expressed in the sporophyte (111.1 bp compared to 123.4 bp, $p = 0.004$). Under the genomic design hypothesis, this might be explained by the fact that highly expressed genes are often ubiquitous and do not require much regulatory information in introns or flanking regions [7]. A regional mutational bias model could explain the reduced introns if highly expressed genes are associated with high rates of deletions. Both of these hypotheses are supported by a positive correlation between the lengths of introns and flanking intergenic regions in human [2,7,8]. In contrast, for most genes there is very little correlation between intron length and the mean length of 5′ and 3′ flanking regions in *Arabidopsis* (Spearman $\rho = 0.02$, $p = 0.09$). Furthermore, the length of intergenic regions was not significantly correlated with mean expression in the sporophyte, and the intergenic regions flanking genes that were highly expressed in the sporophyte were not reduced in length (2,037.8 bp, compared to the mean of 1,986.6, $p = 0.10$). Thus, we find no evidence of a contribution from gene regulation or regional mutational effects to intron length variation in *Arabidopsis,* whereas the reduced intron lengths of genes expressed in pollen strongly support the transcriptional efficiency model.

Genes that were expressed in pollen had significantly lower intron densities (number of introns per kilobase of exon) than genes that were expressed in at least one of the sporophyte conditions (2.4 introns per kb compared to 3.0 introns per kb; $p = 0.001$). The genes that were the most highly expressed in pollen had an average intron density that was lower still (1.8 introns per kb), significantly lower than for the genes that were highly expressed in at least one of the sporophyte conditions (2.6 introns per kb; $p = 0.01$). It is possible that the reduced intron densities result from a disproportionate number of partially processed retroposed genes in the pollen gene dataset rather than from selection for efficiency. Although we cannot rule this out, if retroposition is indeed responsible we might expect an increased density of introns toward the 3′ end [19]. However, we find that the relative density of introns in the 3′ and 5′ halves of genes expressed in pollen is no different to genes expressed in the sporophyte (data not shown).

The debate over whether selection to reduce the cost of transcription is indeed responsible for the shorter intron lengths observed in highly and broadly expressed animal genes has remained unresolved [2,7]. Very high levels of competition at the gametophytic stage of plants provide a useful system in which selection hypotheses can be explored. Although natural selection acting on genes that are expressed at the haploid stage is thought to be an important aspect of plant evolutionary biology [13,14], the reduction in intron lengths that we observe in the genes that are expressed in pollen represents the first well-demonstrated example of the impact of gametophytic selection on the genome of a plant. At least in the case of the genes that are expressed in pollen, there is strong evidence that selection for efficiency, rather than genomic design or regional mutational bias, plays a major role in shaping intron content.

Patterns of genome evolution can differ significantly between outcrossing organisms and self-fertilizing organisms, such as *A. thaliana* [20]. Because *Arabidopsis* has most probably become highly self-fertilizing in the relatively recent past [21] and because insertions and deletions occur on a much longer time-scale than base substitutions [22,23], we expect the evolution of intron lengths in *Arabidopsis* to be dominated by outcrossing reproduction. However, even though heterozygosity is greatly reduced in self-fertilizing organisms so that most gametophytes carry identical alleles, gametophytic competition between transient heterozygotes resulting from de novo mutations still occurs and may well be sufficient to cause the observed reduction of intron lengths in genes expressed in pollen.

Is it conceivable that a slightly different model might apply, one in which speed rather than cost of transcription was important, owing to the fact that only one copy of the genome is present in pollen? If the increased time required to transcribe long introns rather than the energetic cost of transcription [3,6,24] is the primary selective force acting to reduce intron lengths, then it could be argued that the reduced availability of template in the gametophyte, rather than gametophytic selection, could explain intron length reduction in pollen genes. However, because several polymerases can be attached to the same template simultaneously [25,26], gene length need not have much, if any, impact on the steady-state capacity of the template to produce messenger RNA. The additional time required to transcribe longer genes may increase the activation time of a gene but is not expected to have a disproportionately large impact on highly expressed genes. In contrast, the energetic cost of transcription is a linear function of the amount of the transcript that is produced, irrespective of whether transcription is from one or two templates. Because the energetic cost of transcribing longer genes is the same in the gametophyte and sporophyte, we consider that the increased sensitivity to slight differences in fitness caused by strong gametophytic selection is responsible for the reduced length of the introns from genes that are expressed in pollen.

Materials and Methods

Models of *A. thaliana* genes were extracted from version 5 of the annotated *Arabidopsis* genome downloaded from TIGR (ftp://ftp.tigr.org/pub/data/a__thaliana/ath1/PSEUDOCHROMOSOMES/). Genes for which more than one gene model was available (corresponding to alternative transcript isoforms) were omitted. SAGE gene expression data derived from pollen [27], seedlings [28], seedling roots [29], root [30], and seedling aerial tissue [16] were downloaded from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) or from the source data of the original manuscripts. Only tags that were mapped to a single gene and genes to which only a single tag had been mapped were retained for each dataset. In each of the SAGE datasets, the 20% of genes with the highest tag counts were defined as highly expressed. We used only expression data from tags with counts of at least five for each dataset in order to ensure robust results and that the data were comparable between all of the datasets. All statistical tests were carried out in the R statistical computing environment (http://www.R-project.org).Two-tailed Wilcoxon Rank Sum tests were performed for all of the comparisons between sample means. We used robust regression to fit a linear model to intron lengths as a function of expression level and site of expression (sporophyte or gametophyte) considering all genes expressed in pollen and genes from the SAGE dataset representing the largest number of genes (constructed from the aerial part of the plant [16]). For the linear model, only genes from the sporophyte dataset that were not also present in the gametophyte dataset were considered. Gene expression levels in pollen and a range of sporophyte conditions (root, leaf, stem, hypocotyl, and seedling), estimated using the Affymetrix (Santa Clara, California, United States) ATH1 Arabidopsis Genome Array Gene Chip as part of the Expression Atlas of *Arabidopsis* Development [17], were obtained prior to publication with the kind permission of the authors. The data are available from the NASCArrays database (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl; slide Ids ATGE__73__A/B/C, ATGE__3__A/B/C, ATGE__91__A/B/C, ATGE __28__A2/B2/C2, ATGE__2__A/B/C, ATGE__96__A/B/C). We used the mean value of the signal for each gene that was called present in the original analysis. For each condition, the top 20% of most highly expressed genes were defined as highly expressed.

Supporting Information

Dataset S1. Gene Expression and Intron Length Data

Found at DOI: 10.1371/journal.pgen.0010013.sd001 (2.2 MB TXT).

Acknowledgments

CS and CG acknowledge the support of the South African National Research Foundation and National Bioinformatics Network.

Competing interests. The authors have declared that no competing interests exist.

Author contributions: CS conceived and designed the experiments and analyzed the data. CS, CG, and LDH suggested analyses and wrote the paper. ■

References

1. Hurst LD, McVean G, Moore T (1996) Imprinted genes have few and small introns. Nat Genet 12: 234–237.
2. Urrutia AO, Hurst LD (2003) The signature of selection mediated by expression on human genes. Genome Res 13: 2260–2264.
3. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA (2002) Selection for short introns in highly expressed genes. Nat Genet 31: 415–418.
4. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. Trends Genet 19: 362–365.
5. Comeron JM (2004) Selective and mutational patterns associated with gene expression in humans: Influences on synonymous composition and intron presence. Genet 167: 1293–1304.
6. Ucker D, Yamamoto K (1984) Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. J Biol Chem 259: 7416–7420.
7. Vinogradov AE (2004) Compactness of human housekeeping genes: Selection for economy or genomic design? Trends Genet 20: 248–253.
8. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, et al. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. Genome Res 13: 1998–2004.
9. Fullerton SM, Bernardo Carvalho A, Clark AG (2001) Local rates of recombination are positively correlated with GC content in the human genome. Mol Biol Evol 18: 1139–1142.
10. Mascarenhas JP (1993) Molecular mechanisms of pollen tube growth and differentiation. Plant Cell 5: 1303–1314.
11. Lord EM, Russell SD (2002) The mechanisms of pollination and fertilization in plants. Annu Rev Cell Dev Biol 18: 81–105.
12. McCormick S (2004) Control of male gametophyte development. Plant Cell 16: 142–153.
13. Walbot V, Evans MM (2003) Unique features of the plant life cycle and their consequences. Nat Rev Genet 4: 369–379.
14. Bernasconi G, Ashman TL, Birkhead TR, Bishop JD, Grossniklaus U, et al. (2004) Evolutionary ecology of the prezygotic stage. Science 303: 971–975.
15. Majewski J, Ott J (2002) Distribution and characterization of regulatory elements in the human genome. Genome Res 12: 1827–1836.
16. Robinson SJ, Cram DJ, Lewis CT, Parkin IA (2004) Maximizing the efficiency of SAGE analysis identifies novel transcripts in *Arabidopsis*. Plant Physiol 136: 3223–3233.
17. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al. (2005) A gene expression map of *Arabidopsis thaliana* development. Nat Genet 37: 501–506.
18. Becker JD., Boavida LC, Carneiro J, Haury M, Feijo J (2003) A. Transcriptional profiling of *Arabidopsis* tissues reveals the unique characteristics of the pollen transcriptome. Plant Physiol 133: 713–725.
19. Mourier T, Jeffares DC (2003) Eukaryotic intron loss. Science 300: 1393.
20. Marais G, Charlesworth B, Wright SI (2004) Recombination and base composition: The case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol 5: R45.
21. Charlesworth D, Vekemans X (2005) How and when did *Arabidopsis thaliana* become highly self-fertilising. Bioessays 27: 472–476.
22. Saitou N, Ueda S (1994) Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol Biol Evol 11: 504–512.
23. Ophir R, Graur D (1997) Patterns and rates of indel evolution in processed pseudogenes from humans and murids. Gene 205: 191–202.
24. Izban MG, Luse DS (1991) Transcription on nucleosomal templates by RNA polymerase-Ii in vitro—Inhibition of elongation with enhancement of sequence-specific pausing. Genes Dev 5: 683–696.
25. Hawley DK, Roeder RG (1987) Functional steps in transcription initiation and reinitiation from the major late promoter in a HeLa nuclear extract. J Biol Chem 262: 3452–3461.
26. Femino AM, Fay FS, Fogarty K, Singer RH (1998) Visualization of single RNA transcripts in situ. Science 280: 585–590.
27. Lee JY, Lee D (2003) H. Use of serial analysis of gene expression technology to reveal changes in gene expression in *Arabidopsis* pollen undergoing cold stress. Plant Physiol 132: 517–529.
28. Du Z, Scott AD, May GD (2003) Amplification of high-quantity serial analysis of gene expression ditags and improvement of concatamer cloning efficiency. Biotech 35: 70–72.
29. Ekman DR, Lorenz WW, Przybyla AE, Wolfe NL, Dean JF (2003) SAGE analysis of transcriptome responses in *Arabidopsis* roots exposed to 2,4,6-trinitrotoluene. Plant Physiol 133: 1397–1406.
30. Munos S, Cazettes C, Fizames C, Gaymard F, Tillard P, et al. (2004) Transcript profiling in the chl1–5 mutant of *Arabidopsis* reveals a role of the nitrate transporter NRT1 in the regulation of another nitrate transporter, NRT2. Plant Cell 16: 2433–2447.