Whole-Genome Microarray in *Arabidopsis* Facilitates Global Analysis of Retained Introns

Hadas Ner-Gaon and Robert Fluhr*

Department of Plant Sciences, Weizmann Institute of Science, Rehovot 76100, Israel

(Received 2 May 2006; revised 19 July 2006; published online 16 September 2006)

Abstract

Alternative splicing (AS) is an important post-transcriptional regulatory mechanism that can increase protein diversity and affect mRNA stability. Different types of AS have been observed; these include exon skipping, alternative donor or acceptor site and intron retention. In humans, exon skipping is the most common type while intron retention is rare. In contrast, in *Arabidopsis*, intron retention is the most prevalent AS type (~40%). Here we show that direct transcript expression analysis using high-density oligonucleotide-based whole-genome microarrays (WGAs) is particularly amenable for assessing global intron retention in *Arabidopsis*. By applying a novel algorithm retained introns are detected in 8% of the transcripts examined. A sampling of 14 transcripts showed that 86% can be confirmed by RT–PCR. This rate of detection predicts an overall total AS rate of 20% for *Arabidopsis* compared with 10–22% based on EST/cDNA-based analysis. These findings will facilitate monitoring constitutive and dynamic whole-genome splicing on the next generation WGA slides.

Key words: microarray; TILING; alternative splicing; *Arabidopsis*; intron retention; NPR-1; GIGANTEA

1. Introduction

Alternative RNA processing pathways are a result of combining different splice junctions that are present in pre-mRNA transcripts. In this way, a variety of mRNA and proteins can be created from the same gene. Alternative splicing (AS) is thought to play a major role in expanding the potential informational content of eukaryotic genomes. Recent evidence indicates a high incidence (32–60%) of AS in the human genome, predominately in the form of exon skip while a minor form was intron retention (5–16%).1–4 Although rare, intron retention can play an important biological role. It has been linked to tumor growth5 and is part of developmental regulation of a proinsulin expression.6

Plants are thought to exhibit less AS (10–22%).3,7,8 Computational analysis of AS in *Arabidopsis* by EST-pair alignment identified 436 alternatively spliced genes.9 Unexpectedly, intron retention was found to be the most common type (45%). Retained introns were shown to be present in RNA derived from polyribosomes, demonstrating that these intron retention events are not the byproduct of incomplete splicing but are found in a translatable context in the cytoplasm.9 Iida et al.7 aligned 248 514 RAFL (RIKEN *Arabidopsis* Full-Length) cDNA/EST sequences to the *Arabidopsis* genome using a BLAST-based method.10 They identified 15 214 transcription units (TUs) containing at least two sequences each and observed AS for 11.6% of these TUs of which 44% were retained introns.7 In a recent study using a large collection of EST/cDNA, 4707 (21.8%) of the transcripts showed 8264 AS events. Approximately 56% of these events were of the intron retention type.8 These studies confirmed that a low percentage of genes (10–22%) are alternatively spliced and that intron retention is the most prevalent AS type in *Arabidopsis*.

Calculating the rate of AS using expressed sequence data (ESTs and cDNAs) can underestimate the amount of AS for the following reasons. Collectively, the EST and
cDNA database in plants is relatively small and, thus, may lack low-abundance transcripts. In addition, AS predictions based on cDNA data are in many cases biased towards the termini of transcripts owing to the preponderance of end-sequence reads among ESTs and oligo(dT)-based priming for reverse transcription. Recently, high-density oligonucleotide-based whole-genome microarrays (WGAs) have emerged as a novel platform for genomic analysis beyond gene expression profiling. Unbiased WGAs can be used for ‘epigenomic’ mapping and have been used to monitor exon skipping in humans.\(^\text{12,13}\) By comparing the expression levels of all the exons within genes, it was shown that a large fraction (~80\%) of expressed genes on chromosomes 21 and 22 exhibit exon skipping. This estimate is significantly higher than previous predictions, which was based on computational analysis of ESTs and cDNAs.\(^\text{9}\) The array-based approach to analyzing AS also has limitations. For example, the detection of relatively rare splicing events is difficult. Furthermore, in a manner analogous to obtaining multiple EST libraries from different tissue, a large number of distinct tissues or cell populations must be surveyed to obtain sufficient confidence to call a splice variant at a specific splice junction.

In this work by using published WGA information, we monitor intron retention in \textit{Arabidopsis} and show that it is readily detectable, generally leading to transcripts containing new reading frames that would result in truncated gene products.

2. Materials and methods

2.1. Data sources


2.2. Normalization of the whole-genome array-WGA

The hybridization value from a set of 12 oligonucleotide arrays representing 94\% of the \textit{Arabidopsis} genome sequence (110 Mb) was used.\(^\text{14}\) Each array contains 835,000 oligos of 25 mer size. Four RNA populations were hybridized to these arrays.\(^\text{14}\) The probe sequences and the hybridization values for the 12 slides were taken from GEO Accession viewer http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi? using the GEO accession GPL slide numbers (GPL432-GPL443). Hybridization data were normalized by taking the log2 of the data of all probes and then by dividing each probe intensity value in a given experiment by the median intensity value of all probes in that experiment. As reported, the median intensity value of all probes were determined to be a good estimate of the background noise level in a given experiment.\(^\text{14}\)

2.3. Retained-intron analysis

An intron was defined as candidate constitutive or as a retained intron based on its combined probes hybridization values relative to the mean hybridization value of the other introns and exons in the transcript. In order to define retained introns in an expressed transcript the probes of its exons were treated as one group, while each intron, which contained at least three probes, was treated as different group. A balanced one-way ANOVA for comparing the means of all the groups in a transcript was performed. The ANOVA test evaluates the hypothesis that the samples all have the same mean against the alternative hypothesis that the means are not all the same. To insure a False Discovery Rate (FDR) of at most 0.05, the Benjamini-Hochberg (BH) method was used.\(^\text{15}\) Transcripts in which the exons mean was at least 1.09 and the introns mean was less than 1.09 were ranked according to their \(P\) ANOVA value. The \(P\) ANOVA value was then compared with the BH critical value to select significant values. The BH critical values and the \(P\)-value for each ANOVA test are shown in Supplementary Table S2 available at www.dnaresearch.oxfordjournals.org. Using these results a test was then performed that determines which pairs of means are significantly different and which are not. This test employs the Tukey’s least significant difference procedure. PERL programs were written in order to differentiate between groups that have the same mean to groups that are different from each other. Programs are available upon request from H.N-G.

2.4. Gene ontology

Gene ontology classification of transcripts was carried out using the first level categories of the Gene Ontology classification of the biological processes of their protein products (ftp://ftp.arabidopsis.org/home/tair/Onotologies/Gene\_Ontology/ATH\_GO\_GOSLIM.20060121.txt). Gene ontology categories that deviate from cumulative binomial distribution with the probability for retained-intron genes \(P = 0.08\) and \(1 - P\) for constitutive introns genes, with significant \(P\)-value set at 0.05, are summarized.
2.5. Growth of plants, cell culture, RNA extraction and RT–PCR

*Arabidopsis thaliana* (ecotype Columbia) seeds were surface sterilized and placed onto sterilized MS containing Petri dishes. The seeds were cold-treated at 4°C for 2 days and then incubated 7 days at 16 h light/8 h dark cycle. Flower buds were collected from mature *Arabidopsis* plants (5–7 weeks old) grown on soil. *Arabidopsis* suspension-culture cells were a generous gift from Dr Gideon Grafi. The culture was maintained at 16 h light/8 h dark cycle with constant shaking and was subcultured every week and harvested then. Root tissue was obtained by incubating surface-sterilized seeds in flasks containing 100 ml Gamborg’s liquid media. The seedlings were incubated for 2 weeks with constant agitation under 16 h light/8 h dark cycle, and root tissue was collected and frozen.

RNA was extracted from the four different tissues, using RNeasy (Qiagen, Hilden, Germany). RT–PCR was conducted as previously published.16 Oligonucleotide primers were designed for simultaneous amplification of both the spliced and unspliced variants or by using an internal intron probe in which the opposing primer was from the next non-adjacent exon. A complete list of the PCR primers used in this study is shown as Supplementary Table S6 available at www.dnaresearch.oxfordjournals.org.

3. Results

3.1. Division of probes in WGA to accommodate gene structure-criteria

A unique set of whole-genome arrays representing ~94% of the *Arabidopsis* genome sequence are composed of a set of 12 microarrays (6 for the forward strand and 6 for the reverse complement), each containing ~835,000 oligos of 25 mer size. Each array contains an end-to-end ‘tile’ of oligonucleotides (from head to tail) with no overlap in their DNA sequences. In this array, only ‘perfect match’ probes were included in the whole-genome array design. The standard Affymetrix *Arabidopsis* gene expression controls are present on each of these chips. Recently, RNA from four sources were used to probe this array to determine by empirical analysis the transcriptional activity of the *Arabidopsis* genome.14 Tissues sampled included seedlings, flowers, suspension-culture cells and roots.

We wished to ascertain the possibility of refining from this data knowledge about alternative transcripts. In order to parse the data into established annotation of genes and their exons and introns structure the genes structures predicted by TAIR (The *Arabidopsis* Information Resource; http://www.arabidopsis.org/) were used. The gene structure includes the genomic location and sequence of each exon and intron. In cases in which one gene includes more then 1 transcript (e.g. known AS), the transcript structure that included the most exons was used so that a gene was divided into as many distinct groups as possible. The oligos were aligned to the exons and introns sequences in order to allocate oligos to exons and introns. The rest of the oligos that include control probes, probes with more than 1 hit in the genome, intergenic probes as well as probes that fall on intron/exon borders were removed. Thus, in total, 1,682,698 unique probes of the introns or exons were found to be amenable to analysis (Supplementary Table S1 is available at www.dnaresearch.oxfordjournals.org).

3.2. Criteria for differentiating introns and exons in the whole-genome screen

Our aim was to devise statistical tests for each transcriptional unit so that aberrant behavior of introns and exons could be registered. Hybridization values were normalized by dividing each probe intensity value in a given experiment by the median intensity value of all probes in that experiment as done previously.14 As hybridization values are distributed normally in a log scale, we transformed all the values to log2 and then normalized by dividing each probe intensity value in a given experiment by the median intensity value of all probes in that experiment.

We first examined the possibility of assessing retained introns by comparing the expression data of individual introns compared with a transcripts exon hybridization value. An intron was defined as ‘candidate constitutive’ or as ‘retained’ intron based on its mean hybridization value relative to the mean hybridization value of all introns and exons in a TU. To establish global statistical parameters that can be used to distinguish between intron and exon scores in a transcriptional unit, we first explored the hybridization values of introns and exons in the full genome expression database. Genes that do not contain introns or that have multiple non-unique probes in their introns are not included in this analysis. The hybridization values of 18,275 genes were used (Table 1). The mean hybridization values of all the probes that belong to the exons of a specific gene were calculated, as well as the mean hybridization values of all the probes that belong to the gene introns. This assumes that the vast majority of plant introns are constitutively spliced. The mean hybridization values distribution of all the transcripts exons and introns is shown in Fig. 1. The two samples are statistically different based on the goodness-of-fit hypothesis test (Two-sample Kolmogorov–Smirnov, \(P < 10^{-100}\)). A cut-off minimum for the mean hybridization value of exons of a transcript of 1.09 was chosen for analysis as at that level 58% (10,574) of the transcripts could be analyzed but only 1.5% of the introns mean hybridization values were above this level.
3.3. Examining intron retention in whole-genome arrays

In order to identify retained introns in the 10 574 transcripts, the probes of the exons in each transcript were treated as one group, while each intron, which contained at least three probes, was treated as a different group. The three probes limitation was used to increase the significance of the statistical tests. A balanced one-way ANOVA for comparing the means of all the groups in a transcript was performed. To insure a FDR of at most 0.05, the BH method was used. In the example shown, the AT2G36530 transcript is defined as containing only constitutive introns as the exons group is significantly different from all the introns. In Class 2 transcripts, one intron is statistically similar to the exons group. Thus, in the case of transcript AT2G47470, intron 5 is defined as a retained intron (Fig. 2B). Intermediate cases are exemplified in Class 3 (intermediate retained). In this case the exceptional intron is expressed at a level that is between the exons group and all the other introns. In the example shown in Fig. 2C, the AT5G11200 transcript shows that intron 6 hybridization values are between the exon and intron groups.

Examination of the rest of the transcripts shows that 8206 out of 10 270 (80%) fall into one of the intron classes. In the remaining 2064 cases, there is no clear statistical differentiation of the intron and exon groups. In these cases, the transcript exon and intron groups are compared again using a lower confidence level \( a = 0.1 \) instead of 0.05. This causes the interval around the mean of each group to contract. At this probability level conclusions could be drawn for an additional 620 cases. In the remaining 14% of transcripts the statistical variance was too great and no clear definition of the intron/exon expression boundaries could be made. In these cases, the transcripts were termed ‘undifferentiated’. The algorithms above as applied to the whole-genome arrays are summarized in the Table 1. Thus, \(~8%\) rate of the transcripts that could be measured contained retained introns. This is slightly higher than previous EST and cDNA-derived estimates for retained introns of between 3 to 6% of total transcripts. The detailed analysis of all the transcripts are tabulated in Table S2 available at www.dnaresarch.oxfordjournals.org.

### Table 1. Intron and exon classes in the WGA survey

<table>
<thead>
<tr>
<th>Type</th>
<th>Genes</th>
<th>Introns/exons</th>
<th>Analyzed transcripts</th>
<th>Class 1</th>
<th>Class 2</th>
<th>Class 3</th>
<th>Undifferentiated</th>
<th>%IR/%ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron</td>
<td>18 275</td>
<td>59 510</td>
<td>10 270</td>
<td>8123</td>
<td>644</td>
<td>59</td>
<td>1444</td>
<td>7.97</td>
</tr>
<tr>
<td>Exon</td>
<td>19 782</td>
<td>92 456</td>
<td>10 725</td>
<td>6149</td>
<td>452</td>
<td>125</td>
<td>–</td>
<td>8.6</td>
</tr>
</tbody>
</table>

\( ^a \)All exons and introns probes in Table S1 available at www.dnaresarch.oxfordjournals.org were parsed to the specific transcript that they belong to. Of these 59 510 introns and 92 456 exons contained at least three probes/intron or exon. Those introns or exons are part of 18 275 or 19 782 genes, respectively.

\( ^b \)Values from the four different RNA samples were combined. Transcripts were used for further analysis if the mean hybridization value of its exons was \( \geq 1.09 \) and the mean hybridization value of its introns was \( <1.09 \). Transcripts in which the \( P \)-value of the ANOVA test was above the BH critical value were also excluded. The BH critical values and the \( P \)-value for each ANOVA test are shown in Supplementary Table S2 available at www.dnaresarch.oxfordjournals.org. In the exon analysis 727 and 449 exons that were the first or last exon skipped in class 2 and 3, respectively, were not included as their veracity could not be determined.

\( ^c \)Transcripts classes are as described in the text; Class 1, constitutive; Class 2, retained or skipped; Class 3, intermediate.

\( ^d \)Intron retention and Exon skipped rate are computed by dividing the total number of Class 2 and 3 transcripts by the number of statistically significant cases (analyzed transcripts less undifferentiated transcripts).

![Figure 1. Distribution of the mean hybridization values of the probes in exons and introns transcripts. Probes were divided into exons and introns. The mean hybridization value of all the exons in a transcript as well as the mean hybridization value of all its introns were calculated.](image-url)
5 constitutive-type introns was examined by RT–PCR. The genes were selected such that they represent low- and high-expression value transcripts, retained introns in different parts of the transcripts and contain intermediate and retained-type introns. Each PCR was carried out using RNA that was extracted from four different samples including; seedlings, flowers, suspension-culture cells and roots that were combined together. Retained introns were traced either by primers that flank the intron or by using an internal intron probe in which the opposing primer was from the next non-adjacent exon. In this way, retained (I) and the smaller spliced fragment (S) could be recovered. However, if a fragment larger than that predicted by

\[ \text{'I'} \] was detected, i.e. a fragment that included the non-adjacent but normally constitutively spliced intron was detected, this indicated that the RT–PCR amplification used pre-mRNA as its template and was rejected. All fragments containing putative introns were confirmed by sequencing (data not shown). The results shown in Fig. 3 and summarized in Table 2 indicate that 12 out of 14 introns (86%) that were predicted to be retained by the expression analysis were detected by RT–PCR. In a similar manner, when constitutive introns were examined 4 out of 5 could be verified as being constitutive.

3.5. Feasibility of detecting ‘exon skip’ in whole-genome arrays

In theory, a modification of the same algorithm should be applicable to estimating exon skip. However inspection of Fig. 1 shows that the distribution of the introns mean hybridization value is much narrower than the exons mean hybridization value distribution. Thus, the lower hybridization values expected from potentially skipped exons would fall well within the normal exons distribution values, making the direct application of WGA difficult. The limitation of this application is exemplified when the same statistical methods used for detection of retained introns are applied to detect exon skip. Thus, to detect exon skip in a transcript, the probes of its introns were treated as one group, while each exon, which contained at least three probes, was treated as a different group. Exon skip of the first or last exons were excluded as these are probably due to 5' and 3' transcription start/finish differences and do not necessarily arise from splitting. Furthermore in that case, any data retrieved would be impossible to verify owing to the lack of flanking sequence for RT–PCR priming. The results of the search for exons skips were divided as for the introns. Examination of internal exons using the statistical tests and parameters as those applied for examination of the introns showed an exon skip level of more than 8% (Table 1). A sample of previously un-annotated AS events in Arabidopsis was experimentally tested by RT–PCR. Each PCR was done using RNA that was extracted from four different samples as carried out for intron retention analysis. However, in this case only a low rate verification was achieved, 3 out of 8 (37.5%; Supplementary Table S3 available at www.dnaresearch.oxfordjournals.org). Thus, owing to the difficulty of separating the exons distribution from the introns distribution in the available database, the detection of exon skip is unreliable.

3.6. Comparison of the results for intron retention to EST-derived estimates

Identification of alternative transcripts have relied on comparison of sequences from EST and cDNA databases, a method that is inherently different from the type of
analysis carried out here, and, therefore, it is worthy to compare the results. The TAIR6 release contains 31,407 putative genes of which 3159 (10%) have annotated splice variants. Using TAIR annotation (ftp://ftp.arabidopsis.org/home/tair/Maps/seqviewer_data/SV_gene_feature.data) the type of each splice variant was determined. Out of the 3159 genes, 2251 can be analyzed in the WGA experiments and were composed of two transcripts that could be compared to one another. Of these 468 were shown to be of the retained-intron type, a rate that is considerably lower than the more than 40% that was previously observed.3,7–9 In order to compare the efficiency of both methods the limitations applied to the WGA analysis were applied to the retained introns in the TAIR database. Of the 468 retained introns, 257 met the criteria employed to analyze the existing WGA data. Of these 45 or 18% were detected in common by the WGA method (Supplementary Table S4 is available at www.dnaresearch.oxfordjournals.org). These results indicate the overlapping but complementary nature of both approaches as a tool to determine AS.

3.7. Distribution of biological functionalities of transcripts with retained introns

Examining the assigned global functional activity of transcripts with alternatively spliced retained introns will contribute to understanding their function in the biology of the plant. Genes that are analyzed in this study are specific to flowers, root, cell culture or seedling and as a result their gene ontology will be different from the total genes ontology. Therefore, in our case, changes in gene ontology distribution were sought by comparing the retained introns group with the constitutive introns group analyzed in this work rather than to total genes ontology. As nearly 8% of the genes contain retained introns we expect a similar distribution to be emulated in the gene ontology. Deviations from the expected binomial distribution with the $P$-value set at 0.05 are summarized in Table 3. Strikingly, the groups that are enriched in the retained introns group include mostly RNA processing and signal transduction. In contrast, transcripts related to protein metabolism are particularly underrepresented in the intron retention group.

Figure 3. Composite of gene structure, primer choices and gel fractionation of RT–PCR products. The gene structure is from TIGR View (http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=arab). Arrows indicate the position of primers used. In cases where flanking intron probes showed two migrating fragments one gel insert is shown. In cases using intron probes, the appropriate controls are also shown. The results are summarized in Table 2. I and S indicates retained or spliced fragment, P indicate pre-mRNA, m indicates marker lane.
### Table 2. Detection of retained introns by RT-PCR

<table>
<thead>
<tr>
<th>Gene identifier</th>
<th>Exons mean</th>
<th>Introns mean</th>
<th>α*</th>
<th>Type of intron</th>
<th>Flanking primers</th>
<th>Intron primers</th>
<th>Location of intron</th>
<th>Intron results</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT4G03200</td>
<td>1.10</td>
<td>1.00</td>
<td>0.05</td>
<td>Retained</td>
<td>S</td>
<td>Pre RNA</td>
<td>Intron 9</td>
<td>-</td>
</tr>
<tr>
<td>AT1G18160</td>
<td>1.11</td>
<td>1.00</td>
<td>0.1</td>
<td>Retained</td>
<td>S</td>
<td>I</td>
<td>Intron 2</td>
<td>+</td>
</tr>
<tr>
<td>AT1G27320</td>
<td>1.12</td>
<td>1.01</td>
<td>0.05</td>
<td>Retained</td>
<td>I, S</td>
<td></td>
<td>Intron 10-3'-UTR</td>
<td>+</td>
</tr>
<tr>
<td>AT2G19620</td>
<td>1.13</td>
<td>1.02</td>
<td>0.05</td>
<td>Retained</td>
<td>S</td>
<td>I</td>
<td>Intron 3</td>
<td>+</td>
</tr>
<tr>
<td>AT1G22930</td>
<td>1.13</td>
<td>1.00</td>
<td>0.05</td>
<td>Retained</td>
<td>S</td>
<td>Pre RNA</td>
<td>Intron 7</td>
<td>-</td>
</tr>
<tr>
<td>AT1G22770</td>
<td>1.14</td>
<td>1.03</td>
<td>0.05</td>
<td>Retained</td>
<td>S</td>
<td>I</td>
<td>Intron 5</td>
<td>+</td>
</tr>
<tr>
<td>AT1G02410</td>
<td>1.14</td>
<td>1.02</td>
<td>0.05</td>
<td>Retained</td>
<td>S</td>
<td>I</td>
<td>Intron 1-5'-UTR</td>
<td>+</td>
</tr>
<tr>
<td>AT2G19940</td>
<td>1.16</td>
<td>1.00</td>
<td>0.1</td>
<td>Retained</td>
<td>I, S</td>
<td>I</td>
<td>Intron 14</td>
<td>+</td>
</tr>
<tr>
<td>AT1G76140</td>
<td>1.16</td>
<td>1.00</td>
<td>0.05</td>
<td>Retained</td>
<td>S</td>
<td>I</td>
<td>Intron 6</td>
<td>+</td>
</tr>
<tr>
<td>AT3G48440</td>
<td>1.17</td>
<td>1.03</td>
<td>0.05</td>
<td>Retained</td>
<td>I, S</td>
<td>I</td>
<td>Intron 2</td>
<td>+</td>
</tr>
<tr>
<td>AT3G13300</td>
<td>1.20</td>
<td>1.03</td>
<td>0.05</td>
<td>Retained</td>
<td>I, S</td>
<td>ND</td>
<td>Intron 1</td>
<td>+</td>
</tr>
<tr>
<td>AT1G32230</td>
<td>1.25</td>
<td>1.05</td>
<td>0.05</td>
<td>Intermediate</td>
<td>I, S</td>
<td>ND</td>
<td>Intron 4</td>
<td>+</td>
</tr>
<tr>
<td>AT1G76030</td>
<td>1.31</td>
<td>1.00</td>
<td>0.05</td>
<td>Intermediate</td>
<td>S</td>
<td>I</td>
<td>Intron 5</td>
<td>+</td>
</tr>
<tr>
<td>AT1G48920</td>
<td>1.37</td>
<td>1.01</td>
<td>0.05</td>
<td>Intermediate</td>
<td>S</td>
<td>I</td>
<td>Intron 1</td>
<td>+</td>
</tr>
<tr>
<td>AT1G53310</td>
<td>1.24</td>
<td>1.00</td>
<td>0.05</td>
<td>Constitutive</td>
<td>S</td>
<td>Pre RNA</td>
<td>Intron 2</td>
<td>+</td>
</tr>
<tr>
<td>AT4G04910</td>
<td>1.16</td>
<td>0.99</td>
<td>0.05</td>
<td>Constitutive</td>
<td>S</td>
<td>Pre RNA</td>
<td>Intron 11</td>
<td>+</td>
</tr>
<tr>
<td>AT2G20070</td>
<td>1.28</td>
<td>1.02</td>
<td>0.05</td>
<td>Constitutive</td>
<td>S</td>
<td>ND</td>
<td>Intron 3</td>
<td>+</td>
</tr>
<tr>
<td>AT3G21215</td>
<td>1.25</td>
<td>1.02</td>
<td>0.05</td>
<td>Constitutive</td>
<td>S</td>
<td>ND</td>
<td>Intron 3</td>
<td>+</td>
</tr>
<tr>
<td>AT5G20380</td>
<td>1.17</td>
<td>1.00</td>
<td>0.05</td>
<td>Constitutive</td>
<td>L, S</td>
<td>Pre RNA, I</td>
<td>Intron 11</td>
<td>-</td>
</tr>
</tbody>
</table>

aChosen significance level value.
bIntron types were selected from Table 1.
cFlanking primers (Supplementary Table S6 is available at www.dnaresearch.oxfordjournals.org) yielded either ‘I’ which indicates fragments that includes the intron or ‘S’ which indicates fragment size indicative of a spliced transcript.
dIn the cases in which ‘I’ was not detected by using flanking primers, another PCR was done, using intron primers. ‘I’ indicates transcripts that include the retained intron but in which the other intron was spliced. ‘Pre-RNA’ indicates that the slower migrating fragment of the constitutive and retained intron was detected. ND, not detected.
eIntron that was indicated in the bioinformatics analysis as retained and was detected by RT–PCR was marked as ‘+’. Intron that was supposed to be retained but was not detected was marked ‘-’.
fIntron that was indicated in the bioinformatics analysis as constitutively spliced and was not detected in the PCR was marked as ‘+’. Intron that was supposed to be spliced out but was detected was marked ‘-’.

### Table 3. Distribution of assigned gene ontology in genes with retained introns and constitutive introns

<table>
<thead>
<tr>
<th>GO Annotation</th>
<th>Constitutive</th>
<th>Retained</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0000298 Nuclear mRNA splicing, via spliceosome</td>
<td>11</td>
<td>5</td>
<td>0.006</td>
</tr>
<tr>
<td>GO:0007165 Signal transduction</td>
<td>33</td>
<td>7</td>
<td>0.024</td>
</tr>
<tr>
<td>GO:0006468 Protein amino acid phosphorylation</td>
<td>210</td>
<td>28</td>
<td>0.010</td>
</tr>
<tr>
<td>GO:0006412 Protein biosynthesis</td>
<td>272</td>
<td>13</td>
<td>0.008</td>
</tr>
<tr>
<td>GO:0005975 Carbohydrate metabolism</td>
<td>134</td>
<td>6</td>
<td>0.035</td>
</tr>
</tbody>
</table>

The transcripts with intron retention from the WGA database were classified by the biological processes of their protein products ontology in relation to the constitutive introns from this database. (ftp://ftp.arabidopsis.org/home/tair/Ontologies/Gene Ontology/ATH_GO_GOSLIM.20060121.txt) (Details in Supplementary Table S5 available at www.dnaresearch.oxfordjournals.org). Transcripts listed as ‘biological process unknown’ (30% of retained introns) were not included. Note that gene ontology may lead to the classification of one gene into multiple categories. Only classifications containing at least 5 retained introns or 64 constitutive transcripts were included (i.e. at least 0.9% of the expressed transcripts).
3.8. Physical characterization of retained introns

Differences in the physical characteristics of introns, such as size or flanking sequence information, may serve to characterize this class of AS. The characteristics that were measured include: position of the intron in the transcript, i.e. present in the UTR or CDS, whether it contains an open reading frame (ORF), the intron size, its GC content and borders of the intron. Interestingly as a group, the median size of the retained introns is significantly smaller when compared with the size of the constitutive introns irrespective of the presence of an ORF (117 and 176 bp, respectively). The size distributions are statistically different based on the goodness-of-fit hypothesis test (Two-sample Kolmogorov–Smirnov, \( P = 9.5 \times 10^{-14} \)).

We note that the median size of introns calculated from the TAIR database (ftp://ftp.arabidopsis.org/home/tair/Sequences/blast_databases/TAIR6_intron_20060221) is 99 bp. The higher means in all types of introns measured here is a direct result of the requirement that introns be represented by at least three probes (i.e. 3 probes \( \times 25 \) bp size/probe). Interestingly, the GC content of retained introns is higher than that of constitutive introns (37% compare to 33%) but lower than the GC content of coding regions (44%).17 The two GC content distributions are statistically different based on the goodness-of-fit hypothesis test (Two-sample Kolmogorov–Smirnov \( P = 1.5 \times 10^{-139} \); Supplementary Figure S1 is available at www.dnaresearch.oxfordjournals.org). Higher GC contents have been recently detected in retained introns that were identified by genome EST/cDNA alignments.8 The possibility exists that the higher GC content reflects an artifact that is associated with enhanced non-specific hybridization of probes, i.e. they may be more sensitive to background noise. In order to assess this possibility, we examined the distribution of the GC content of the intron probes whose hybridization values were above or equal 1.09 (Supplementary Figure S2 is available at www.dnaresearch.oxfordjournals.org). In this case, the \( R^2 \) for the correlation of GC content and hybridization value was found to be 0.06. Hence for the range of hybridization values used in our survey non-specific hybridization is not correlated to GC content and the values measured are a result of true specific hybridization.

Additional qualities of note in the characterization of retained introns are that, relatively more retained introns originate from the 5′- and 3′-UTR region compared with the constitutive intron distribution (13.4% compared with 7.3%; Table 4). By comparison, the global analysis predicts that at least 9% of Arabidopsis genes have introns in their UTRs.18,19 Strikingly, inspection of Table 4 shows that 609 (86.6%) of the retained introns are inside the CDS. Yet of these, only 4.6% maintain an in-frame ORF (Table 4). Interestingly this rate, while low, is still more than 3-fold higher than that noted for constitutive introns (1.5%). Finally, Intron borders can play a crucial role in splicing efficiency.20 However a comparison between the borders of the constitutive and retained introns showed no differences (Supplementary Figure S3 is available at www.dnaresearch.oxfordjournals.org). Thus, border sequence cannot be used to predict alternative intron splicing.

### Table 4. Characterization of retained and constitutive introns

<table>
<thead>
<tr>
<th>Intron type</th>
<th>Position</th>
<th>ORF</th>
<th>Number of introns</th>
<th>% ORF</th>
<th>% position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constitutive</td>
<td>CDS</td>
<td>0</td>
<td>21310</td>
<td>98.5</td>
<td></td>
</tr>
<tr>
<td>Constitutive</td>
<td>CDS</td>
<td>1</td>
<td>325</td>
<td>1.5</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td></td>
<td></td>
<td>21635</td>
<td>92.7</td>
<td></td>
</tr>
<tr>
<td>Constitutive</td>
<td>3′-UTR</td>
<td></td>
<td>249</td>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>Constitutive</td>
<td>5′-UTR</td>
<td></td>
<td>1445</td>
<td>6.2</td>
<td></td>
</tr>
<tr>
<td>Retainedα</td>
<td>CDS</td>
<td>0</td>
<td>581</td>
<td>95.4</td>
<td></td>
</tr>
<tr>
<td>Retainedα</td>
<td>CDS</td>
<td>1</td>
<td>28</td>
<td>4.6</td>
<td></td>
</tr>
<tr>
<td>Summary</td>
<td>CDS</td>
<td></td>
<td>609</td>
<td>86.6</td>
<td></td>
</tr>
<tr>
<td>Retainedα</td>
<td>3′-UTR</td>
<td></td>
<td>31</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>Retainedα</td>
<td>5′-UTR</td>
<td>0</td>
<td>63</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

αRetained introns are Class 1 and Class 2 retained introns.

4. Discussion

We demonstrate that direct transcript expression analysis using high-density oligonucleotide-based WGA is particularly amenable for assessing global intron retention in Arabidopsis. However, the detection of exon skip is unreliable owing to the difficulty in separating the distribution of the expression values that distinguish exons from introns. We note that detection of alternative 5′ and 3′ splicing sites by analysis of WGA would be limited to those alternative sites that comprise a number of probe lengths (in the present case a minimum of 75 bp). As alternative 5′ and 3′ splice choices are usually of shorter length, they would not generally be detected.

Genome-wide detection of retained introns in Arabidopsis reveals that at least 8% of genes that can be surveyed by microarray include a transcript with a retained intron. A sample of these transcripts was tested using RT–PCR and 86% could be verified. In our analysis, results of the WGA data were combined to enable the generation of statistically reliable outcomes. Thus, tissue-specific intron retention may be lost as combining RNA from diverse biological material can obscure the AS signal. In the future, the use of sample repeats from the same tissue will undoubtedly improve the statistical veracity of the results and optimize the detection of aberrant intron behavior. Nonetheless, as intron retention comprises above 40% of all AS events3,7,9 this indicates that overall the rate of AS in
Arabidopsis is 20%. This rate is considerably higher than 10–14%, which was garnered from compilation of EST and cDNA data from different sources but similar to the one that was recently reported from a large collection of EST/cDNA. Similarly, the recent application of WGA to human chromosome 21 and 22 showed a much higher rate of AS than previously detected.

The retained-intron type is distinct from the other AS types because it represents an absence or failure of spliceosome action rather than alternative choice of splice junctions. Our analysis detects several features that are specific to retained introns. They are smaller, have higher GC content, are relatively enriched in the UTR and those in the CDS tend to have a slightly higher chance of containing an ORF when compared with a constitutive-type introns. The size of introns has been shown to play a role in AS biology. In Drosophila, the size and location of the flanking introns control the frequency and the type of AS that a pre-mRNA transcript undergoes. Furthermore, it is possible that the attribute of higher GC content could contribute to a weaker U-rich splicing signal as a result of less efficient AU protein binding. This is because intronic U-rich or AU-rich elements can influence splice site selection and splicing efficiency particularly in plants (reviewed in Simpson and Filipowicz). These sequences are likely to bind proteins (analogous to animal hnRNP proteins), which may permit intron recognition and recruit or allow access to spliceosomal components and the formation of commitment splicing complexes.

Of particular note is that 82.6% (581 out of 703) of the transcripts containing retained introns are in the CDS and would shift reading frames and have an early stop codon. Of these, in 68% (477 out of 703) of the cases the intron that is retained is not the last one. (data not shown). The position of these stop codons is important in the context of potential nonsense-mediated decay (NMD) effects. NMD serves as a surveillance mechanism that removes erroneous mRNAs containing a premature termination codon. Thus, if NMD mechanism exists in plants, its selective power of degradation is incomplete as the transcripts with AS events are readily detected. The existence of limited use of NMD in plants has recently been shown in the analysis of transcripts containing premature termination codon. However, the coupling of AS and NMD could be an important post-transcriptional regulation mechanism to adjust the level of transcript isoforms.

Inspection of the intron retention database provides biological insight. RNA processing and signal transduction transcripts appear to be differentially enriched in the retained-introns group. A serine/arginine-rich (SR) family of proteins is implicated in constitutive and AS of pre-mRNAs. They appear in alternatively spliced forms and at least one of them, SR1, contains a retained intron. Transcription factors that show AS variants were described for five genes and in one of them, At1g26260, the first intron is retained. Numerous studies have contributed to the view that SR proteins play a general role in splicing and can modulate splice site selection in a concentration-dependent manner (Ref. and reviewed by Manley and Tacke). One imaginable consequence of this is that cells may regulate the expression or activity of individual SR proteins, or their antagonists, to control the expression of one or more target genes in a tissue-specific and/or developmentally regulated fashion. Transcript that contains intron may serve as a negative regulator of the SR protein concentration.

The potential regulatory power of transcripts with retained introns can be gleaned from examination of the database (Supplementary Table S2 is available at www.dnaresearch.oxfordjournals.org). Two examples are NPR1 (At1g64280) and GIGANTEA (At1g22770). NPR1 is a key regulator of the salicylic acid (SA)-mediated systemic acquired resistance (SAR) pathway and it confers resistance to pathogens. Known mutations of this gene cause loss of SAR induction, loss of expression of PR genes and increased susceptibility to infections. WGA analysis predicts that the last intron of this gene is retained and contains an in-frame stop codon. GIGANTEA promotes flowering under long days in a circadian clock-controlled flowering pathway, together with CONSTANTS (CO) and FLOWERING LOCUS T (FT). Mutations in this gene cause late-flowering phenotype. WGA analysis predicts intron 10 (out of 13) to be retained and to contain a stop codon.

The biological function of retained introns remains an enigma. A recent survey of human transcripts shows that one-third of the AS contained premature termination codons that are the apparent targets of NMD. It was proposed that regulated unproductive splicing may be a means of modifying the level of protein expression or of introducing modified gene products. This work provides a novel database of retained introns and their genomic sites and lays a framework for future application of WGA for dynamic screening of changes in splicing during the plants’ life and as a result of environmental input.

Acknowledgements: This research was supported by the Israel Science Foundation grant No 388/02 and by the BARD United States-Israel Binational Agricultural Research and Development Fund Grant number 18-3454-03.


References


