Identification of cis-regulatory motifs in first introns and the prediction of intron-mediated enhancement of gene expression in the plant Arabidopsis thaliana

Georg Back
Max Planck Institute of Molecular Plant Physiology
Dirk Walther (walther@mpimp-golm.mpg.de)
Max Planck Institute of Molecular Plant Physiology

Research Article

Keywords: Gene expression, Introns, Intron-mediated enhancement, Sequence motifs, Random Forests, Arabidopsis thaliana

Posted Date: February 24th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-234201/v1

Identification of cis-regulatory motifs in first introns and the prediction of intron-mediated enhancement of gene expression in the plant Arabidopsis thaliana.

Georg Back and Dirk Walther*

Max Planck Institute of Molecular Plant Physiology

Potsdam, Germany

14476

back|walther@mpimp-golm.mpg.de

+49-0-331-5678108

*Corresponding author

Abstract

Background. Intron mediated enhancement (IME) is the potential of introns to enhance expression of its respective gene. This essential function of introns has been observed in a wide range of species, including fungi, plants, and animals. Studies in the plant *Arabidopsis thaliana* have shown that enhancing introns exhibit a distinct base composition and are generally the first intron located close to the transcription start site. However, the mechanisms underlying the enhancement are as of yet poorly understood. The goal of the study was to identify potential IME-related sequence motifs and genomic features found in first introns of genes in the plant Arabidopsis thaliana.

Results. Based on the rationale that functionale sequence motifs are evolutionarily conserved, we exploited the deep sequencing information available for Arabidopsis thaliana, covering more than one thousand Arabidopsis accessions, and identified 81 candidate hexamer motifs with increased conservation across all accessions, and which also exhibited positional occurrence preferences. Of those, 71 were found associated with increased correlation of gene expression of genes harboring them, suggesting a cis-regulatory role. Filtering further for effect on gene expression correlation yielded a set of 16 hexamer motifs, corresponding to five consensus motifs. While all five motifs represent new motif definitions, two are similar to the two previously reported IME-motifs, whereas three are altogether novel. To identify additional IME-related genomic features, Random Forest models were trained for classification of gene expression level based on an array of different sequence-related features. The results indicate that introns harbor information with regard to gene expression level and suggest sequence-compositional features as most informative, while position-related features, that were thought to be of central importance before, were found with lower than expected relevance.

Conclusions. Exploiting deep sequencing and broad gene expression information and on a genome-wide scale, this study confirmed the regulatory role on first-introns, characterized their intra-species conservation, and identified a set of novel sequence motifs located in first introns of genes in the genome of the plant *Arabidopsis thaliana* that may play a role in inducing high and correlated gene expression of the genes harboring them.

Keywords: Gene expression, Introns, Intron-mediated enhancement, Sequence motifs, Random Forests, Arabidopsis thaliana

Introduction

The presence of introns is one of the defining characteristics of eukaryotes (1). Introns are nucleotide sequences within a gene that are transcribed, but are removed during the maturation of the functional mRNA construct. While self-splicing introns, which catalyze their own removal from the maturing RNA, can be found in all domains of life, spliceosomal introns are unique to eukaryotes. As the name suggests, these introns require the spliceosome, a large protein/RNA complex, for removal from the nascent RNA product.

The origin of introns has been discussed intensively. Some studies suggest that introns, at their core, are selfish deoxyribonucleic acid (DNA) elements, emerging during population bottlenecks (2). Their function, if having any at all, would therefore be a secondary development. The evolutionary age of introns proved to be a controversial topic, which divided the field into two major theories. The "intron early" theory posits that introns are of very ancient origin, even predating cellular life, and all modern organisms without introns lost them secondarily. By contrast, the "intron late" theory advocated a relatively recent or even several independent origins (3). However, as more whole genome sequences for a variety of prokaryotes became available, revealing an absence of introns in the majority of this group, the intron early theory was regarded increasingly unlikely. The occurrence of spliceosomal introns in all known eukaryotes and the precursor self-splicing introns in prokaryotes indicate a very ancient origin nonetheless. This led to the proposition of a synthesis theory (4), suggesting a very early emergence of self-splicing introns in simple prokaryotes, which further developed into spliceosomal introns in a common ancestor of all eukaryotes.

The question as to which functions introns and intron splicing have, has been discussed since their discovery. The fact that they can be found in the nuclear genomes of all eukaryotic life forms suggests a central importance (1,5). This assumption also seems to reconcile the absence of an immediate functional role with the substantial burden introns impose on the harboring cells and organisms as a whole. The spliceosome complex consists of several small nuclear ribonucleic acids (snRNAs) and proteins, which all have to be available at high quantities to guarantee efficient splicing and to avoid delays in mRNA maturation, straining the cellular resources. Furthermore, transcription and intron splicing itself are energy-intensive processes. Lastly, due to the crucial role of splice junctions in the correct splicing of genes, the organisms containing introns are prone to gene loss by single point mutations. A single mutation in the splice site or a mutation that generates a new splice site can lead to incorrect splicing and unwanted retention of intronic sequence or deletion of parts of exons. This may lead to loss of function of the resulting protein or to complete gene-loss due to introduction of a premature stop codon, either

contained in the intronic sequence or by a frameshift, resulting in nonsense-mediated decay. Additionally, intron mutations can lead to an activation of non-canonical splice sites, resulting in similar effects (6). Correspondingly, a high number of human genetic diseases have been linked to mutations in introns (7). Similarly, several plant mutants related to splicing errors affecting the phenotype are known (8).

As introns have been observed across all eukaryotic cells, it seems reasonable to assume that introns play a vital role. Allowing alternative splicing is one of the leading explanations for the prevalence of introns. Alternative splicing is a process of selective removal or retention of exons or introns. This leads to multiple different mRNA transcripts that can be derived from a single gene, resulting in several different proteins, thereby efficiently expanding the protein repertoire. Alternative splicing has been observed for most multi-exon genes, and has been suggested as the main contributor to eukaryotic complexity (9). Conversely, other authors suggest that only a very small portion of alternative transcripts are actually translated into proteins at relevant levels (10). Besides alternative splicing, other functions for introns have been discovered over the years. mRNA-stability has been linked to introns, indicating that splicing stabilizes mRNA, leading to an increase of mRNA half-life (11). Furthermore, splicing can assist in the 3′-end formation of the pre-miRNA by recruiting capping factors (12). Surprisingly, the “fate” of an intron does not necessarily end after its excision from the mRNA. Introns can contain RNA genes, which are expressed upon removal from the host mRNA. Several non-coding RNA families have been found in introns, such as snoRNAs, long non-coding RNAs (lnRNAs), miRNAs and small-interfering RNAs (siRNAs) (1). The resulting RNAs can then have gene-regulatory roles. miRNAs and siRNAs inhibit translation of their target genes by binding to the mRNA and either blocking translation or initializing degradation of the mRNA. Thus, by inhibiting its host or its host antagonist, siRNAs can both promote or inhibit the expression of its host gene (13).

As perhaps one of the most essential functions of introns, the enhancement of gene expression has been reported. Studies have shown that certain introns are able to enhance the expression of their respective genes by a significant amount. Interestingly, in contrast to normal enhancers, these introns have to be transcribed to trigger this effect (14). This enhancement, known as Intron Mediated Enhancement (IME), is strong enough to be used as a tool in the repertoire of molecular biology techniques to boost the expression of specific target genes, and has been suggested to contribute to the high expression levels of housekeeping genes (15). IME was one of the earliest known functions of introns, when it was discovered in 1987 in maize (16). Since then, IME has been found in a variety of species, from plants to vertebrates and nematodes (17,18). It has been reported that IME can act via increased transcription rate, increased nuclear export of the transcript, transcript stability, and even enhanced translation

efficiency (19). The mechanisms responsible for these diverse modes of action of introns on the gene expression are as of yet not fully understood. However, a strong correlation between proximity of an intron to the transcription start site (TSS) and its capability to enhance expression has been identified, with the vast majority of reported IME found associated with the first (5'-most) intron of a gene (20). As for the actual mechanism that results in gene enhancement, both splicing-dependent and splicing-independent effects have been reported.

The splicing snRNA U1 is known to interact with and recruit polymerase II via the transcription factor TFIIH, increasing the probability of transcription initiation in the presence of a functional 5′ splice site, possibly even independent of the actual splicing process (21). This effect is the stronger the closer the site is to the TSS, as closer proximity is likely to result in an increased probability of the recruited polymerase to initiate transcription. Potentially further contributing to IME, the previously mentioned effect of splicing on the mRNA stability and capping efficiency increase the amount of functional mRNA. The exon junction complex, which forms at the junction of two exons after splicing, can increase the rate of nuclear export of the processed mRNA (22). Finally, the exon junction complex has also been linked to increase in translation efficiency, by interacting with translation factors and ribosome components and recruiting or activating them (23).

All previously described mechanisms depend largely on the splicing process and the spliceosome. This would mean that merely the presence of an intron in close proximity to the TSS is sufficient to induce IME, independently of the actual intron sequence. However, several studies have shown that different introns at the same position can have different effects on gene expression, ranging from no effect to strong stimulation (24). Furthermore, certain introns can confer tissue specificity, leading to a spatially differentiated expression enhancement (25). These findings strongly indicate a sequence-dependent mode of action of IME, which does not primarily rely on splicing alone. Concordantly, studies have shown that intron sequences can enhance expression without being spliced at all (19,26). This implies a regulation similar to expression regulation by transcription factors. However, severe reduction of IME when splicing was inhibited has been reported as well (27,28). These conflicting observations either suggest two independent mechanisms of IME, splicing-dependent and splicing-independent, or a combination thereof, which is supported by some studies reporting a reduction, but not a depletion of IME in absence of splicing (15). To understand how all these factors play together, identifying introns, which lead to IME is crucial.

Primarily, IME-introns have been identified by experimental evidence (16,25,29). While this is essential for gaining further insight into IME, it only covers a small portion of potential IME

candidate introns. To identify IME introns on a larger scale, bioinformatic methods are required. Currently, the only available computational method with this goal is IMEter, which works under the assumption that TSS-proximal introns are enriched in IME sequence motifs assumed as words (k-mers of length 5) (30,31). IMEter computes a log-odds score for an intron sequence to correspond to TSS-proximal, and hence IME-signal-bearing, introns by scoring the actually present pentamers relative to observed relative frequencies of pentamers in TSS-proximal vs. TSS-distal introns. This straightforward approach has shown promising results. Many of the previously established IME-introns were assigned high scores high by IMEter, and the authors even suggested a correlation between IME-induced mRNA-fold increase and IMEter score (32). Furthermore, in top scoring introns, two sequence motifs were detected, which, when present at high densities, are able to induce IME (25,32). These motifs even led to an increase of mRNA levels when located within an exon (15). However, not all introns that have been reported to be able to induce IME, score accordingly with IMEter or are enriched for the two reported motifs (15,32). Therefore, alternative computational approaches may identify additional regulatory motifs in introns.

One commonly used strategy to bioinformatically identify functional sequence elements, called phylogenetic footprinting, assumes that they are likely conserved across related species. With available sequence and associated single nucleotide polymorphism (SNP) information, this approach can also be applied to intra-species evolution, as applied, for example, in *Arabidopsis thaliana* (33). This method requires a large set of sequences to achieve a high motif resolution. However, when performed across species, increasing the number of species in a set also increases the divergence between sequences, leading to ambiguity in the orthologous relationships. Conversely, if the sequences are too similar, for example being from the same species, the density of accumulated mutations might be too low to determine conserved regions. Therefore, in this case, a large number of genomes is required to ensure a minimum variety among the compared sequences. The 1001-genome-project offers this opportunity, with a Single Nucleotide Polymorphism (SNP)-set for 1,135 fully sequenced *Arabidopsis thaliana* accessions (34). Moreover, a large compendium of gene expression data (microarray- and RNA-seq-based) is available, allowing to test whether introns sharing a particular motif also share a similar expression pattern as well as available methylome data, permitting to include epigenetic information in the analysis (35). A previous study was able to identify novel motifs in promoter regions using the 1001 Genome project SNP set (36). The authors compared conservation not only across a single mapping location, but compared all mapping locations of a motif. This design attempts to avoid the previously described problem of the relatively low SNP density between the Arabidopsis accessions by determining the degree of conservation of a motif over all its occurrences in the genome. As a further relevant parameter, it was shown that most motifs had a

positional preference, indicating a correlation between occurrence of a motif and SNP density.

As an alternative and broader, more general approach, machine learning can be applied to the problem of identifying factors relevant for IME. Machine learning has been applied successfully to a wide range of biological problems, such as predicting RNA and protein folding, identifying descriptors of enzyme efficiency, and prediction of regulatory motifs (37–39). While the classical machine learning algorithms, such as decision trees, lost popularity in favor of deep learning, their models allow better interpretation. By identifying intron features that can influence gene expression levels, alternative modes of action underlying IME could be discovered.

This study builds on the rationale that IME-motifs are conserved more than expected by chance and uses a SNP-based approach to identify cis-regulatory elements, initially defined as sequence hexamers, applied to intronic sequences. Introns were split into two groups, first (5'-most) introns, which are known to be able to induce IME, and all other downstream introns. By adding conservation and location distribution as characteristic features associated with IME candidate motifs, our approach attempts to extend the concepts established by IMEter, which relies on candidate motif occurrence differences in the first vs. other introns alone. Differential methylation as a potential regulator of IME was also investigated here. For validation of functional relevance, correlation of gene expression of all genes containing candidate IME motifs in their first intron was used.

To assess the information contents of intronic sequences on gene expression and to extract associated informative features, his study also includes a Random Forest (RF) classifier model for prediction of mRNA expression levels based on intron sequence information. Strongly expressed genes have been associated with IME, indicating that intron features that coincide with high gene expression of the respective gene might be related to IME (20). A number of sequence characteristics of the respective first intron, such as intron length, nucleotide composition, distance to TSS, distance to the translation start codon, and the IMEter score served as features for the Random Forest classifier. In addition, folding energetics of intronic RNA, cross-species conservation, and presence of transpons was considered as well. The goal was not only to create an accurate model, but also to extract features that contribute the most to the prediction accuracy in addition to the more targeted k-mer motif approach.

We report the identification of 16 motifs, collapsing to five consensus motifs. While all five motifs constitute new motif definitions, two resemble previously reported IMEter motifs, and three appear altogether novel. The RF-models confirm the predictive potential of introns with regard to the expression level of their host genes and suggest features associated with base composition as particularly informative. Our results shed new light on the possible mode of action responsible

for IME and may serve as a starting point for further approaches examining IME in the future.

Materials and Methods

Extraction of intron positions and sequences

The Arabidopsis Information Resource (TAIR)10 (40) General Feature Format version 3 (GFF3) file was used to extract the sequence coordinates of all mRNA introns within the *Arabidopsis thaliana* genome via exon positions to infer intron positions. All introns shorter than ten base pairs (bp) were excluded. A FASTA file containing all introns was created by using bedtools (41) and the complete TAIR10 genome as a reference. The intron set was then split into first, i.e. the promoter-proximal intron set, and the set of other introns. Introns located in the 5'UTR of a gene were detected by an overlap between an artificially length-extended (5bp at either end) intron and 5'UTR coordinates.

Extraction of relevant single nucleotide polymorphisms (SNPs)

SNPs were extracted from the 1001 genome project variance calling file (VCF) (34). All variants that were positioned in one of the introns were extracted. A threshold of 50 was set as the minor allele frequency for SNP positions to be considered and 500 valid (i.e. non-"N") alleles called, with alleles counted as haploid counts (i.e. counts per chromosome). With VCFtools (42), the resulting VCF file was used to extract all SNP positions. In total, 2,426,458 SNPs were used, of which 382,016 were located in introns.

Selection of candidate hexamers

Selection of k-mer size

As a compromise between specificity of motifs (favoring longer motifs) and the combinatorial increase associated with increasing motif-length, a k-mer size of k=6 was chosen, from here on termed hexamers. For each hexamer, their respective positions in each intron were determined using the extracted intron sequences. To avoid bias towards hexamers containing part of the highly conserved splice sites, the first and last three sequence positions of each intron were excluded from the analysis. From the obtained hexamer positions, the frequency and distribution of hexamers within the introns were determined. For analyzing conservation, frequency, and location distribution, results for reverse complementary hexamers were combined with their forward definitions and treated as one hexamer.

Relative frequency of hexamers

Similar to IMEter (32), the frequency of hexamers in first introns compared to other introns was taken as the initial criterion for the identification of potential regulatory hexamers. For both intron

sets, first and other introns, the total occurrence of each hexamer, H, over all introns in the Col-0 reference genome was determined, and then normalized by the total occurrence of all hexamers for each intron group, respectively. Afterwards, the relative frequency, F, was calculated by dividing the normalized frequency of hexamers in the first by the normalized frequency of hexamers in the other introns, with

$$F_{H_i} = \frac{C_{f,H_i} / \sum_{j=1}^{N} C_{f,H_j}}{C_{o,H_i} / \sum_{j=1}^{N} C_{o,H_j}}, \qquad \text{Eq. 1)}$$

where C stands for counts, f and o for first and others, respectively. N is the total number of observed hexamers (N=2080).

Degree of conservation of hexamers, conservation rate

To assess the degree of conservation of each hexamer, the total number of occurrences of each hexamer introns was compared to the occurrence of the same hexamer with SNP positions masked, performed separately for first and other introns. The masking was done by replacing each position containing a SNP with a symbol not in the nucleic acid notation, in this case "*". The degree of conservation was calculated as the ratio of hexamer counts, C_H, with SNPs masked and the counts without masking. This provides a position and alignment independent measure of conservation with ratio-values near 1 suggesting high conservation and smaller ratios suggesting increasing variability. For comparison, the randomly expected conservation was computed as:

$$C_r = (1 - \frac{N_{SNP}}{N_{bp}})^k, \qquad \text{Eq. 2)}$$

where N_{SNP} is the number of SNP-positions found in introns and N_{bp} is the total number of positions in respective introns, computed separately for first and other introns. C_r corresponds to the probability of a k-mer not containing any SNP position given the background SNP-density.

Positional distribution of hexamers in introns

Two factors were considered for the location distribution of hexamers within introns. First, since most cis-regulatory elements show preferences for specific localization, we hypothesized that relevant hexamers should show a characteristic distribution, which significantly differs from a uniform distribution. To examine this, the relative positioning of each occurrence of a hexamer in an intron was determined by dividing the first position of each hexamer occurrence by the length of the respective intron. These relative start positions were then binned into ten bins covering an interval of (0, 1). Based on the binned occurrence counts, positional preferences were expressed as position entropies, S_m , withfigure

$$S_H = -\sum_{b=1}^{10} p_{H,b} log(p_{H,b}), \qquad \text{Eq.3)}$$

where $p_{H,b}$ is the relative frequency of hexamer motif (k-mer) H occurring in bin b.

For each hexamer, 10,000 random uniform distributions with the same number of occurrences were simulated and the entropy for each distribution was calculated. Since uniform distributions have the largest possible entropy (over a finite interval), non-uniform distributions should be significantly smaller. By comparing the entropy of the actual hexamer entropy relative to the random entropy, an empirical p-value was calculated.

As a second criterion, to be considered a candidate hexamer motif, the distribution of hexamers was required to be significantly different in first introns compared to the distribution in other introns. A Fisher's exact test on the binned data was used to determine whether there was a significant difference between the two distributions.

For both metrics, the Benjamini–Hochberg method of False Discovery Rate (FDR) adjustment was applied (43).

Multiple sequence alignments/ Consensus motif generation

For the identification of a consensus motif from candidate hexamers, a Multiple Sequence Alignment (MSA) on a subset of hexamers considered candidate motifs was performed. The multiple alignment using fast Fourier transform (MAFFT) tool (44) was used to perform the alignment. JalView (45) was utilized for tree visualization. For comparison of consensus motifs, the STAMP tool (46) was used. Collapse of hexamer motifs into consensus motifs is, by its nature, to some degree, arbitrary and was performed requiring a minimum support per consensus position of two individual motifs and guided by the dendrogram of sequence-distance-clustered motifs (Figure 3a) with the objective to group similar motifs together, while unique motifs separate..

Calculation of IMEter score

IMEter (31) is a tool scoring the similarity of a sequence to introns close to the TSS. IMEter version 2.2 was downloaded from the KorfLab/IME github repository. IMEter was trained with the Phytozome dataset as described in the IMEter use manual (47). The IMEter score for each first intron was then calculated. The introns were subsequently ranked by their IMEter score.

Comparison of gene expression

As a gene expression dataset, microarray expression data from Craigon *et al.* (2004) was used,

covering 20,922 genes with unique probe-geneID mappings, profiled in 5,295 hybridizations/ conditions (48). The data was normalized as described in Korkuc *et al.* (2014) (36). For comparing the gene expression of sets of genes, Pearson correlation of normalized, log-transformed expression levels across all samples was used. For each gene subset, the correlations between all possible combinations of two genes was calculated based on the determined expression levels in the samples contained in the expression dataset. To compare two subsets, a Cohen's d analysis of effect size on the two sets of correlations was performed. This yielded both an evaluation of the direction as well as the magnitude of the effect. Confining the analysis to genes with introns, annotated 5'UTR>0bp, and requiring a log(median_expression)>0.1 left 13,504 genes for expression analysis.

In general, gene subsets can be compared to a set of random genes of equal set size, or other gene subsets. To avoid correlation related to homology present within a gene subset containing a certain hexamer, comparison to subsets of genes containing other, but specific hexamers was performed. For this, hexamers with occurrences similar to the hexamers of interest (+/-10%) were chosen, and correlations for their respective gene subsets were calculated. Then, Cohen's d values for the gene set containing the hexamer of interest and each of the new subsets were calculated. Finally, the mean effect size was determined.

Potential motifs were compared to high IME-scoring introns as judged by the IMEter tool. The correlation of the hexamer gene set was compared to the set of genes with the highest IMEter score with equal set size by calculating Cohen's d effect size.

Analysis of differentially methylated regions

For the analysis of differential methylation, information on differentially methylated regions (DMRs) from Kawakatsu *et al.* (2016) (35) was used. These cover three different types of methylation, CG-DMRs, representing differential methylation only in the CG context; CH-DMRs, which cover only regions that are differentially methylated in the CHG/CHH context; and C-DMRs, which are regions with differential methylation in both contexts. For all sets, all differentially methylated positions (positions that are part of DMRs) within first introns were extracted and summarized for each intron, respectively.

Identification of new motifs and motif binding comparison

The tool Tomtom was used to compare candidate motifs to a set of 872 sequence motifs reported as part of the published DAP-seq motif dataset for *Arabidopsis thaliana* (49). DAP-seq motifs correspond to transcription factor binding sites motifs derived from binding assays of transcription factors to "naked" genome DNA segments.

GO-term enrichment

Gene Ontology (GO)-term enrichment analysis was performed based on a Fisher's exact test with FDR correction. The terms were extracted from the GO-slim-term subset available from TAIR10 (40).

Prediction of expression level with Random Forest

Selected features

All features chosen to characterize introns were directly or indirectly linked to information contained in first introns. Table 1 lists all features along with a short description. The length of the first intron, the distance of the first intron to the coding sequence, the distance of the first intron to the transcription start site and intron retainment of the proximal intron were derived from the extracted intron GFF3 file. The relative base-type frequencies were derived from the extracted fasta file of the first introns, with the flanking three bp bordering the splice sites masked. The relative dimer counts were calculated in a similar fashion as the hexamers described above, but with k=2. All possible dimers were determined, their occurrence in each first intron, excluding the splice sites, were assessed, and the count of reverse complementary dimers were combined. Finally, the counts were normalized by dividing by the respective intron length.

Information about differentially methylated regions (DMRs) was derived as described above. Similarly, the IMEter score for the first introns was calculated as described above. The SNP-frequency per bp was calculated using the VCF file.

The minimum folding energy was calculated using mfold (50). The mfold script was modified, removing unnecessary plotting and non-relevant calculations. For each first intron, an overhang of 20 bp into the flanking exons on both sides were included in the calculation. The minimum energy was then normalized by dividing by intron length with 40 bp for the overhang added.

For considering the presence of conserved non-coding sequences (CNS), a dataset from Haudry *et al.* (2013) was used (51). A position was considered conserved if an associated CNS sequence was found present in at least four of the nine Brassicaceae species examined in (51). The relevant positions, i.e. positions that overlapped with first introns, were extracted. For every intron, the total number of CNS positions was determined, and normalized by intron length.

Transposable elements were extracted from the TAIR10 transposable element dataset (40). The total number of transposable elements per intron was normalized by intron length.

As an indication of functional relevance, we probed introns for evidence of retention in annotated splice variants as reported in the GFF-file. If an intron sequence was found to overlap with an exon of an alternative transcript, it was considered retainable (retention=1), otherwise not (retention=0).

Classification

As a target variable for prediction, gene expression level was chosen. For the expression set, the microarray data (48) was utilized. The median expression for each gene across all samples was determined. A binary classification into high/low expression was chosen using the median as a set division threshold. To potentially increase prediction performance, models were also created for a modified dataset, which contained only genes found in the upper and lower quartile of RNA expression levels. The goal was to create two more distinct groups to allow better classification (increased contrast).

Model selection

For creating the actual prediction model, the Random Forest (RF) classifier as implemented in the sklearn (52) module was used. Hyperparameter tuning via random grid search with cross-validation to increase performance and reduce overfitting of the model was performed. The final RF-models contained 6000 trees. Each tree had a maximum depth of 10 with a minimum number of samples per split of 5, and a minimum of two samples at the leaf nodes. Number of features to choose from at every split was sqrt(total_number_of_features).

Dataset selection

For training the Random Forest model, the dataset for the introns was randomly split into training and test dataset with a ratio of 80% and 20%. For the creation of learning and ROC curve analysis, ten-fold cross-validation on the whole set was performed.

Feature importance

For determining the feature importance, permutation feature importance was selected. It has been suggested that this method provides better results than the "Mean Decrease in Gini" method, which is used by the sklearn classifier (53). After training the classifier, one feature of the test set was permuted randomly and the accuracy was scored. This was repeated five times for each feature, and the mean decrease in accuracy (MDA) was calculated, respectively. This process was repeated for all features.

SHAP importance

The Shapley Additive explanation (SHAP) method explains individual predictions of a model (54). It is based on Shapley Values, which have their origin in game theory. A Shapley value of a feature is the average contribution to all possible feature combinations. Calculation of Shapley values is computational expensive due to combinatorial explosion. SHAP therefore uses sampling to approximate Shapley values to reduce the computational burden. The Python package SHAP (55) was used to calculate SHAP values for the trained models, and to visualize the results.

Statistical analysis and visualization

All statistical analyses were done in Python 3.7 (56). The modules scipy (57), numpy (58), and pandas (59) were used. Visualization and plotting was performed with the modules matplotlib (60) and seaborn (61). In cases of single test statistics, reported p-values less than p=0.001 are not specified further and indicated as p<0.001.

Results

Comparison of SNP-frequencies in first versus other introns

Since it has been shown that specifically the first intron bears the capacity to influence expression of the gene it is part of, the set of Arabidopsis introns was split into two sets, one with only the first introns, i.e. the 5'-most, of each gene, and another for all remaining introns, termed "other introns". The average intron length of first introns was determined as 259.7bp, with a median of 161bp, and a mean of 160.8bp for the other introns, with a median of 100bp, respectively. For both intron sets, the respective SNP-density was calculated by using the variants data of the 1001 genome project (34). Only positions with at least 50 alleles containing a different variant (minor allele) were considered as SNP positions, and the first and last five positions of each intron were excluded to avoid over-representation of splice sites. Surprisingly, first introns were observed to have a slightly higher total SNP-density of 0.0164 SNPs (i.e. polymorphic positions) per bp compared to the other introns with 0.016 SNPs per base position (Mann–Whitney U test, p<0.001). A visualization of the relative SNP-frequency for the first (5' end of intron) 20 bp positions, including a 20 bp overlap into the preceding exon clearly shows this difference (Figure 1a). This effect is not only observable in the introns itself, but also in the preceding exons, likely explained by the embedding of other introns in coding regions with associated conservation pressure, whereas first introns often are found in a non-coding UTR context. The position-resolved conservation profiles, shown in Figures 1a, b, also confirm the expected lower SNP-frequency on and near the exon/intron splice site as well as the expected three-bp periodicity within the exon/coding region. To test whether the difference in conservation effect is related to the positioning of introns in the 5′ untranslated region (UTR), which could potentially explain reduced conservation, first introns were separated into introns positioned in the 5'-UTR and introns positioned in the CDS. Surprisingly, first introns in 5'-UTRs were found to have a lower SNP-density than first introns in the CDS, with an average SNP-density of 0.0141 for the 5'-UTR introns and 0.0169 for the CDS introns (Mann–Whitney U test, p<0.001) (Figures 1 b, c, d). By contrast, upstream intron-flanking regions showed the expected behavior with UTR-exons being less conserved than CDS-exons (Figure 1b).

High sequence conservation, as reflected by a low SNP-density, can be an indicator of

functionality (62). This agrees well with IME-function predominantly being found in introns close to the TSS, and therefore close to, or even within, the 5'-UTR, indicating a possible correlation between conservation and IME function, but within CDS regions, first and other introns do not follow the expected conservation pattern.

Selection criteria for potential cis-regulatory intron motifs

For identifying candidate intron motifs associated with IME, a k-mer-based strategy similar to IMEter was applied, with additionally utilizing conservation and relative position in introns as informative criteria, similarly as described by Korkuc *et al.* (2014) (36). As a compromise between specificity of a sequence motif and combinatorial explosion, a k-mer length of k=6 was chosen. All counts of reverse complement hexamers were combined, leading to a total of 2080 unique potential 6-mer (hexamer) motifs. Four properties were examined for determining whether a hexamer was considered a candidate: 1) higher sequence conservation in first introns than in other introns, 2) higher relative occurrence in first introns than in other introns, 3) non-uniform distribution of the motif within the first intron, and 4) dissimilar positional distribution of the motif between first and other introns. Criteria 3 and 4, which impose positional preferences, were introduced to follow the rationale that similarly to transcription factor binding sites (36,63), intronic motifs may exhibit such positional preferences as well. Of those criteria, criterion 2 follows the approach of IMEter, while criteria 1, 3, and 4 are introduced in addition in this study.

Evolutionary conservation of hexamers

Our approach builds on the rationale that functional motifs show increased conservation. Therefore, and if indeed IME is associated specifically with first introns, we expect potential motifs to be more evolutionarily conserved in first introns than in other introns. The mean conservation rate (see Methods for definition) over all hexamers was determined as 0.9131, higher than the randomly expected rate, C_r, Eq. 2, of 0.905 (Figure 2a). Similarly, other introns had an average hexamer conservation of 0.915 compared to the expected value of 0.907 (Figure 2b). At first, it may be surprising that the average observed hexamer conservation is higher than that based on the expected background conservation (Eq. 3). This apparent contradiction can be explained as an indication that SNPs are not completely randomly distributed within introns, but tend to positionally cluster. Similar observations have previously been reported (64). This could be due to either a bias in the sequencing technology or some biological process. Also, hexamers with very low occurrences tend to have higher SNP-rates (Figures 2a, b). This may point to a sequencing artifact as well (homo-oligomeric stretches). A total of 929 hexamers were determined to have a higher conservation in first introns relative to other introns, while 1151 hexamers were more conserved in other introns, which reflects the observed higher SNP frequency, and hence, lower conservation, in first vs. other introns (Fig 2a).

Relative occurrence of hexamers in first vs. other introns

Under the assumption that functional sequence motifs induce IME, it appears plausible to expect that these motifs show a higher relative occurrence in first introns compared to other introns, since the vast majority of reported IME-introns are first introns of a gene (31). Inspecting relative hexamer counts (count of a particular hexamer divided by the total number of detected hexamers), 843 hexamers were detected with higher relative occurrence in first compared to other introns, while for 1237 hexamers, the inverse was true. A closer examination of the relative count distribution of hexamers revealed a significant difference between the distribution of hexames with lower relative frequency versus those with higher relative frequency in first introns (Figure 2c, Kolmogorov-Smirnov test $p<0.001$). While there are fewer hexamers with higher relative occurrence in first vs. other introns than what is observed in reverse, those that are overrepresented in first introns show a pronounced tail (at around a twofold enrichment factor) that may point to the ones that are functionally significant and, thus, enriched.

Non-uniform positional distribution of hexamers in introns

Studies have shown that functional sequence motifs often exhibit a positional preference (36,63), including signals associated with IME (31). Assuming that potential functional motifs in introns exhibit this preference as well, hexamer positional distributions were tested for deviation from uniformity (see Methods), yielding 1448 hexamers detected with significantly non-uniform positional distributions in first introns.

Dissimilar positional distributions of hexamers in first vs. other introns

Finally, to exclude positional preferences unrelated to hexamer IME function, only hexamers with significantly different positional preferences in first and other introns were considered further. A Fisher's Exact test comparing positionally binned distribution of hexamers (ten bins, see Methods) within first introns to other introns respectively yielded a subset of 459 introns, which were significantly differently distributed in first vs. other introns.

In total, 81 hexamers met all four requirements laid out above, and were investigated further.

Analysis of identified candidate hexamers

Expression correlation of genes containing candidate intronic hexamer motifs

To test for any regulatory effects of the identified 81 candidate first-intron motifs, correlation of gene expression level was taken as an indicator. Under the assumption that an intron motif regulates gene expression, those genes that harbor a particular motif should exhibit a higher correlation of gene expression amongst them than a comparable set of random genes. However, increased correlation among genes with a specific intron motif could not only indicate regulatory

effects, but also originate from the genes being homologous. Closely related genes might exhibit a similar expression profile and will also be more sequence-similar to one another with a correspondingly increased probability to find the same hexamer in their introns. Therefore, candidate motifs were compared to hexamers with similar occurrences as the one under consideration (within a 10% interval of higher/lower occurrence) to account for this effect. Gene expression correlation of the gene subset containing the hexamer of interest was computed, and then compared to the correlation of genes observed to each contain a comparable hexamer in their first intron. Of note, as a control, we compared the matching k-mer approach to the naive approach to simply use all other genes and found concordant results (Supplementary Figure S1).

The median Cohen's d effect size, i.e. the magnitude of the difference of correlation values for the two gene sets across all 81 motifs was 0.018 (std.dev.=0.029), with only 10 hexamers having a negative mean effect size (Table 2; for the complete set of 81 candidate motifs, see Supplementary Table 1). Thus, a significant majority (71 in total) of the 81 selected hexamers exhibited higher correlation than hexamers of similar occurrence (p=1.8E-12, binomial test, with p_{prior}=0.5). Sixteen candidate motifs with a mean effect size of greater than an arbitrarily chosen threshold of +0.05 (5%) were selected and investigated further.

Analysis of candidate hexamer subset with evidence of expression regulation

For each of those 16 hexamers, the average gene expression level of genes harboring them in their first introns was significantly higher than the average of the whole set (p<0.001), with Cohen's d effect size ranging from 0.18 for ACCCTA to 0.45 for AGATCG. This result is in line with motif-mediated IME being associated with highly expressed genes.

Candidate consensus motifs

The set of hexamers contained sequence-related hexamer sequences, which may be equivalent in function or part of a larger consensus motif, e.g. AGATCG and TCGATC (with its reverse complement GATCGA). Using the program MAFFT, the set of 16 motifs was collapsed into five motifs, GATTCG, TTTCGA, KCGAGAR, ACYCYR, and ARATCGA. Three of these are consensus motifs from several individual hexamers, and two remain as their original hexamer definition (Figure 3a).

Next we tested, whether the identified motifs correspond to known binding sites of known DNA-binding proteins, such as transcription factors. A motif comparison analysis performed with the motif comparison tool Tomtom against the DAP-seq database for Arabidopsis transcription factors and their associated target motifs revealed no significant overlap (all E-values>1) with any of the 872 DAPseq-reported motifs.

To elucidate the biological role of the genes harboring the candidate motifs, a GO-term enrichment analysis for genes, whose first introns contain the five motifs were performed, with gene sets considered separately for every motif. With regard to GO-cellular components, gene sets for all motifs were significantly enriched for cytosol (p_{FDR}<0.0036) and cytoplasmic components, with gene sets associated with three motifs being significantly enriched for Golgi apparatus (ARATCGA, GATTCG, TTTCGA) (p_{FDR}<0.037), and two for endoplasmatic reticulum (ARATCGA, GATTCG) (p_{FDR}<0.024) and nucleus (p_{FDR}<0.025), respectively. All motif sets were significantly depleted for mitochondrial (p_{FDR}<0.04) and extracellular (p_{FDR}<5.9e-11) genes, with four sets being also depleted for chloroplast genes (ACYCYRA, ARATCGA, GATTCG, KCGAGAR) (p_{FDR}<0.042). Testing GO-function terms, all motifs were found enriched for protein binding (p_{FDR}<1.2e-4). Furthermore, structural molecule activity (ARATCGA, GATTCG, TTTCGA) (p_{FDR}<0.016) and DNA/RNA-binding (ACYCYRA, ARATCGA, GATTCG) (p_{FDR}<0.027) were overrepresented in gene sets of three motifs, respectively. The GO-term “unknown molecular function” (p_{FDR}<4.3e-4) was significantly underrepresented for all motifs. Additionally, gene sets of three motifs were depleted for transcription factors (ARATCGA, GATTCG, TTTCGA) (p_{FDR}<0.027). Lastly, significantly enriched process terms were and DNA/RNA metabolism (ACYCYRA, ARATCGA, KCGAGAR, TTTCGA) (p_{FDR}<6.3e-4), cell organization (ACYCYRA, ARATCGA, KCGAGAR, TTTCGA) (p_{FDR}<0.036), and transport (ARATCGA, GATTCG, KCGAGAR, TTTCGA) (p_{FDR}<0.049) with four motif sets each, while signal transduction (ARATCGA, GATTCG, KCGAGAR, TTTCGA) (p_{FDR}<0.014) and unknown processes (all) (p_{FDR}<0.0021) were underrepresented. Thus, generic housekeeping functions appear overrepresented, while signalling and transcription factor activities appear to be less present in the gene sets associated with the five identified IME-candidate-motifs.

Comparison of potential regulatory motifs to IMEter

To further evaluate the newly identified motifs, they were compared to the most commonly used tool for identifying potentially IME introns. IMEter scores whole introns (32), or, in a new version, a sliding window of 50bp (31). For all first introns, the IMEter score was calculated, and then sorted by score. Genes with the highest scoring introns were correlated amongst themselves at significantly higher levels than a subset of random genes of equal set size (p<0.001), with an average Cohen’s d of 0.183 (Figure 4a). The mean expression level of the top 2000 IMEter score genes was significantly higher than that of the whole gene set (p<0.001, Cohen's d effect size of 0.43). By comparison, correlation of expression amongst genes containing either one of the five consensus motifs reported in this study was either comparable to or only slightly below that of the IMEter set (Figure 4b, |Cohen’s d| <0.1), suggesting a potentially cis-regulatory role.

The overlap between the candidate motif gene sets and the corresponding IMEter sets of equal size was large, with an average overlap of 34% (p<0.001, regarding the overlap in percent,

note that sets were always of the same size). This is expected, since both approaches partly employ similar strategies for identifying IME. When the candidate motifs were compared to the IMEter set not containing overlapping genes, the effect size associated with the candidate motif gene set generally increased, resulting in comparable gene expression correlation of genes within the respective gene sets as observed for the size-matched top-IMEter gene sets, with effect size ranging from 0 to 0.1 (Figure 4c), suggesting an even stronger, albeit slightly, regulatory effect associated with the identified five consensus motifs compared to IMEter-selected gene sets.

Based on the IMEter tool, Rose *et al.* (2008) (32) and Parra *et al.*, (31) identified two motifs, CGATT and TTNGATYTG, which were overrepresented in introns with high IMEter scores, and were shown to be associated with induction of gene expression (65). Judged by their sequence, of the five identified consensus motifs two motifs (ARATCGA, GATTCG) showed some resemblance with the two IMEter motifs, albeit not identical, while three motifs (ACYCYRA, TTTCGA, KCGAGAR) can be considered more distinct, and thus, potentially novel functional IME motifs (Figure 3b).

Compared to the top scoring IMEter genes, genes containing either one of the two IMEter motifs had a comparable correlation amongst each other, with a Cohen's d of 0.004 and -0.072, respectively (Figure 4b, star-labeled motifs). With overlapping genes removed from the IMEter set, Cohen's d increased to 0.088 and 0.089 (Figure 4c, star-labeled motifs), thus, requiring the motifs to be present alone yielded a significant co-expression signal. The two IMEter motifs exhibited a significantly higher mean expression than the total set ($p<0.001$), with effect size of 0.31 for CGATT and 0.37 for TTNGATYTG. By comparison, our consensus motifs were found with corresponding effect sizes of ARATCGA: 0.33, GATTCG: 0.32, KCGAGAR: 0.23, TTTCGA: 0.21, ACYCYRA: 0.14 (mean: 0.25). Thus, the two consensus motifs detected as sequence-similar to the reported IMEter motifs (ARATCGA and GATTCG) showed the largest effect size and comparable to the two IMEter motifs, while the other three consensus motifs were found with slightly lower, but still very strong effects.

Taken together, our consensus motifs resulted in similar effects as compared to the IMEter-based intron scoring and the two IMEter-motifs, and yielded novel motif definitions and/or altogether novel motifs that may function in a cis-regulatory fashion.

Effect of differential methylation in first introns on gene expression

A study of vertebrates by Anastasiadi *et al.* (2018) has shown a strong inverse correlation between methylation in the first intron and gene expression (18). They also showed that first introns exhibit the highest density of differentially methylated regions (DMRs) of any genomic feature, and that certain DMRs could positively correlate with gene expression. These findings suggest a potential influence of DMRs on IME and therefore on gene expression, which was further investigated here using published DMR data (35). The gene sets associated with the two

methylation contexts, CG- and C-DMRs, that each were found with sufficient numbers of observations (CH-DMRs were not considered as fewer than 100 cases of overlaps with first introns were observed) had very little overlap with either the top scoring IMEter genes, or any of the potential hexamer motif gene sets. Genes containing C-DMRs in their first intron were significantly more correlated than a set of random genes, with an average effect size of 0.1. However, C-DMR genes had a significantly lower gene expression than the set average ($p<0.001$) with an effect size of -0.74. Conversely, genes with intronic CG-DMRs were expressed at significantly higher levels than the set average ($p<0.001$) with an effect size of 0.07. Yet, the CG-DMR subset showed a comparatively lower expression correlation than the C-DMR set, with only 0.054 as the average effect size compared to a subset of random genes of equal set size. With regard to overlap of DMR-set genes and the gene sets associated with any of the five candidate motifs reported here, for C-DMR, no significant overlap was detected. By contrast, the CG-DMR genes overlapped significantly with all five consensus motifs ($p<0.004$), with enrichment factors of 1.2-fold and higher.

In conclusion, no coherent picture emerges with regard to the role of DMRs in IME. While genes with CG-DMRs in their first introns are expressed at higher than average levels, the corresponding set of genes does not show correlated gene expression, a feature, which we considered evidence of regulation used to identify IME motifs.

Random Forest model for prediction of expression level based on intron features

IME has been connected to highly expressed genes such as housekeeping genes. Thus, it appears possible to cast the problem of identifying features responsible for IME as a feature extraction problem with Machine Learning methods applied to the prediction of expression level. By only including features of the first intron, the goal was to investigate the predictive value of first introns with regard to expression level of their respective genes. Random Forest Classifiers were trained for the prediction of expression level. Initially, genes were binned into two groups, sets with high and low expression level, respectively, used as classes for building the classification models, with the global median expression level taken as the threshold value. To increase contrast, binning of genes was performed based on the upper and lower quartile of expression levels. A spectrum of sequence-dependent and sequence-independent intron features, which we considered potentially predictive, were selected and tested (Table 1).

Using the median-split gene classes, the achieved model performance was modest (area under the ROC (AUC) of 0.68 and an average accuracy in a tenfold cross-validation of 63%. When increasing the gene expression difference between the two considered gene sets by using the upper and lower quartile of expressed genes to train models, a substantial increase of model performance was observed (AUC=0.78, accuracy=72%) (Figure 5a).

Feature importance

For the trained Random Forest models, feature importance, judged as MDA, the mean decrease in prediction accuracy, was determined. For the best performing model, sequence composition features seemed most important, with the percentage of guanine (G) and adenine (A) having the highest impact on model performance (Figure 5b). IMEter score, which is derived from the distribution of pentamer-motifs in introns, the relative occurrence of TC-dimers and percentage of Cytosine (C) were also found to have high feature importance, further suggesting sequence-dependent effects. Finally, intron-length and distance to the TSS had only a small positive effect on prediction performance. This is surprising, since IME has been closely associated with distance to the TSS. However, MDA, while powerful, is susceptible to correlated features, as influences can be masked.

SHAP values are an alternative and very informative way to assess feature importance and decision making of a model. They are calculated for all predictions individually, making them ideal for analyzing the effect of feature values on the prediction. For the model at hand, positive SHAP values indicate that this feature value increases the chance of the model classifying the sample as highly expressed, while negative SHAP values increase the chance of low expression classification. The importance of features determined by SHAP was assessed similarly as by MDA (Figure 5c). Sequence features, such as A-, G-, TC-content, and IMEter-score again had the biggest impact on model prediction. Low values for A-content resulted in positive SHAP values, while high values resulted in negative SHAP value. A similar pattern was observed for the dinucleotide TA. By contrast, for IMEter, length, G and TC, high feature values generally resulted in a higher SHAP value, with lower values having a negative impact. Notable differences between MDA importance and SHAP importance were observed for the features distance-to-the-TSS and CDS-start, which were considered less important by SHAP, and number of differentially methylated regions (methylation C), which had a stronger effect on model prediction according to SHAP. In the case of C-DMRs, an interesting pattern was observed. While a low number of differentially methylated sites had no effect on the model prediction, high numbers resulted in a negative SHAP value, indicating that the model associated them with lower gene expression. This is consistent with the significantly lower mean gene expression level of C-DMR genes reported above. The low impact of both distance features (distance_TSS, distance_CDS) was yet again surprising, since IME has been associated with both, a short distance to the TSS, as well as being positioned in the UTR. Even more surprising is that very small feature values (distances <200b) were associated with negative SHAP values, i.e. the model was more likely to classify the respective gene as expressed at low levels (Figure 5d).

As considered above for the relevance of motifs, it needs to be considered whether there

is indeed a specific signal in introns that causes increased gene expression of the associated genes, or whether our classifier simply picks up on features associated with genes that are expressed at high levels, such as housekeeping genes. To test for that, we extracted the same set of features, as considered for introns, for first exons of the same genes and built RF-models using exon-only, intron-only, and exon-intron-combined features. As shown in Figure 6a, while the performance of exon-only and intron-only features is comparable (AUC=0.78), considering both combined leads to a significant increase of predictability (AUC=0.81). We interpret this as evidence that introns hold information over and above that, which is associated with recognition of highly-expressed gene families alone, for which exon-only and intron-only serve as a suitable point of reference. Furthermore, both exon and intron features were considered equally important (Figures 6b, c)

Taken together, these classification results imply that there is indeed relevant intronic information for determining the expression level class (high vs. low) of genes and suggest a number of informative features.

Discussion

Intron-mediated enhancement (IME) has been discussed as an important regulator of gene expression, found in nearly all eukaryotic systems tested so far [23]. It has been associated with highly expressed genes [20,67,68]. However, the exact mode of action for inducing expression enhancement is not yet understood. Several different mechanisms have been proposed, with the seemingly biggest open question being the importance of splicing [23]. Several studies suggest that the recruitment of splicing factors, even in the absence of splicing, are the major determinants for the induction of enhancement [23,27]. Conversely, studies have shown that some introns and intronic motifs were able to induce enhancement irrespective of splicing and splicing factors, and suggested a mechanism acting at the level of genomic DNA [20,65,67].

Here, we reported the identification of 16 hexamer and five resulting consensus DNA sequence motifs that may be related to IME in the plant *Arabidopsis thaliana*. Building on previous studies on sequence signals associated with IME (31,32), our study exploited the available deep sequencing information of more than one thousand *Arabidopsis thaliana* accessions allowing us to probe for conservation as a hallmark of functional relevance. Furthermore, we imposed more explicitly that motifs should show positional preferences within introns, an assumption that appears supported by prior findings (36,63), and tested as evidence of a functional effect that motif harboring genes exhibit correlated gene expression in addition to elevated expression level, making use of the plethora of available expression information. Thus, we postulated that IME not only leads to increased expression, but also includes a regulatory

component, leading to concerted gene expression of subsets of genes. In line with this assumption, out of 81 motifs, identified based on conservation and evidence of preferential intron locations alone, 71, i.e. almost all, were associated with increased correlation (positive, rather than negative effect size, $p_{bionial,\ prior=0.5}$=1.8E-12, Supplementary Table S1) and 16 (19.7%) were found associated with significantly elevated co-expression (effect size, Cohen's d>5%) (Table 2).

Contributed by the seminal studies on IME (31,32), IME has been associated with whole introns (IMEter tool and score) and two motifs (CGATT and TTNGATYTG, with the first being a sub-motif of the other) have been implied as functional. Our study enlarges this set to 16 hexamer and five consensus motifs that now can be explored further and experimentally characterized. The IMEter tool has been shown to be a good indicator of IME, with experimentally identified IME introns having consistently high IMEter scores, and the level of enhancement of known IME introns correlating with IMEter score (25,31,65,66). The motifs identified here and based on conservation, relative occurrence, and positional distribution were comparable with regard to their effect on correlation of gene expression and expression level to genes with the highest IMEter score (Figures 4b,c). Therefore, it seems likely that the discovered hexamer and consensus motifs are truly related to IME. Building on conservation using intra-species sequence variation, as done here, also is supported by previous observations indicating that regions with high IMEter score were conserved among different species (24,31).

As the mode of action of IME still is a big unknown, the fact that we did succeed in identifying motifs that are, based on our filter and test criteria, associated with IME, suggests that either a molecular recognition event - such as binding by proteins - may be at work, or that the motifs play a RNA-structural role relevant for splicing. At this point, using the approaches presented here, we cannot interpret the data in favor of either of the two alternatives. However, our study provides novel candidates for targeted follow-up studies.

In addition to a search for sequence motifs, we performed a Random-Forest-based classification of genes with regard to gene expression level. Here, the goal was to a) prove predictability of expression level using intron-based information, and b) to identify additional features relevant for IME. Indeed, we were able to show that introns hold information on expression level over and above the information provided by the gene context (exon-related information), (Figure 6a). Overall, base compositional features (most significantly, G-contents) were found most informative, and more important than other parameters such as distance to the TSS or other parameters that would allow to arrive at more interpretable conclusions with regard to mode of action of IME (such as DMRs, folding energy and others) (Figures 5b, 6b). Low A- and high G-content of the first intron were pivotal for classification as a highly expressed gene. In

contrast to the k-mer-based IMEter score, A- and G-content are more general features, describing the composition of the intron and the pre-mRNA. This could indicate a motif-independent influence of first introns on gene expression. Studies have shown that intron composition can regulate splicing by influencing pre-mRNA folding around the splice sites (67), which could explain the observed effects. Compositional effects have also been reported to influence mechanical properties of genomic DNA, such as bending flexibility (68). However, initial attempts by us to use machine learning to associate reported sequence-dependent DNA flexibility measures to IME proved unsuccessful (not shown). However, considering mechanical properties of pre-mRNA may offer a fruitful avenue for further research.

Confirming the validity of prior approaches, the previously published IMEter score was among the most important features. Correlation between IMEter score and enhanced gene expression by IME has been established, also experimentally for selected gene sets (31). The results of this study show that this also applies to the whole genome.

Regarding the role of DNA-methylation, more specifically, differential methylation (DMRs), given the data and approaches used here, no consistent picture emerged. While C-DMR regions in introns were found associated with increased correlation of the corresponding genes, they were expressed at low levels. Conversely, CG-DMR intron genes showed higher expression, but low correlation. Hence, a cis-regulaotry role of DMRs in introns related to IME appears unlikely.

With regard to intron-related cis-regulatory functions, first intons (5'-most) have been considered most relevant (69). Our observation that first introns, on average, show a slightly increased SNP-density compared to the remaining introns (Figure 1) appears counter-intuitive. However, introns located in 5'-UTRs exhibit a reduced SNP-density, and hence increased conservation. Therefore, UTR-located introns may play a different functional role than introns embedded in coding regions, which is consistent with previous reports that several UTR introns have been reported to induce IME (69).

Also, when considered as a predictive feature, the distance of the intron to the transcription start site (TSS), with close distances having been discussed as more associated with IME, was not found to be particularly informative. This relatively low impact of the distance to the TSS on the expression level prediction (Figures 5b,c) is surprising. Previous studies had suggested that proximity to the TSS is an essential property of IME introns [70], and that IME effect declines with distance to the TSS [32]. Our observations were to some degree contradictory. While large distances of introns to the TSS had very little impact on the prediction accuracy of the model, distances shorter than 200 bp increased the likeliness of a gene to be classified as expressed at low, not high levels (Figure 5d). Parra *et al.* (2011) observed a similar

pattern when comparing the IMEter score of introns to their distance to the TSS. Relatively low IMEter scores were found for introns close to the TSS, with the highest IMEter scores observed at a distance of about 200 bp (31). However, in their analysis, the observed IMEter scores were still positive, suggesting enhancement, while in the case of the Random Forest model, very small distances were an indication of low expression (Figure 5d). Our findings suggest that the distance of the first intron to the TSS, as such, is perhaps less important than previously thought, and TSS-proximal introns must, in addition, exhibit a particular composition to lead to IME. The sharp drop in SHAP values, even into the negative value range, for very small distances to the TSS (Figure 5d), which suggests low expression, may perhaps also indicate gene annotation problems, which need to be inspected on a case-by-case basis.

On the technical side, with regard to the employed gene expression data, this study made use of the large microarray-based dataset covering ~20K genes and thousands of different conditions. As RNAseq has increasingly become de-facto standard, we checked whether consistent results would have been obtained had this study been performed with available RNAseq datasets. Using TravaDB (70), a large compendium of Arabidopsis thaliana RNAseq data (158 conditions), we determined a high correspondence of gene expression level (r=0.88, Supplementary Figure S2), and, as reported previously [66], also a high correspondence of pairwise correlation (r=0.49). It should be noted that the probed conditions were very different. Hence, expression level and pairwise correlation proved robust, reflecting condition-independent, coherent expression regulation. Thus, as expression level and pairwise correlation were the two criteria tested for in this study, the microarray data used here can be considered representative.

We performed our analysis within one species (Arabidopsis thaliana) with sequence variations amounting to single nucleotide polymorphisms (SNPs). While this intra-species approach eliminates the alignment challenges associated with inter-species studies, evolutionary conservation is confined to a relatively short divergence time (about 5mya (71)). The associated limitations have been discussed before in a study on promoter elements (36) and correlated mutations (72) and apply here as well.

Concerning the employed classification methodology, we employed Random Forest classifiers. While recently, deep learning architecture (Recurrent and/ or Convolutional Neural Networks (RNNs, CNNs), have proven to be powerful sequence-based classification approaches (73), RFs, in addition to being a powerful classification engine, allow for a more direct assessment of feature importance, which specifically was the goal of our study.

Conclusions

Exploiting deep sequencing and broad gene expression information and on a genome-wide scale, this study confirmed the regulatory role on first-introns, characterized their intra-species conservation, and identified a set of novel sequence motifs located in first introns of genes in the genome of the plant *Arabidopsis thaliana* that may play a role in inducing high and correlated gene expression of the genes harboring them.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

All relevant data are made public as part of this publication or are publicly available.

Competing interests

None.

Funding

Max Panck Society

Authors' contributions

D.W. conceived the study. G.B. performed all computations, except Suppl. Fig.S2, performed by D.W., G.B. and D.W. planned the analyses, interpreted the results, and wrote the manuscript.

Acknowledgements

Not applicable

References

1. Chorev M, Carmel L. The Function of Introns. Front Genet. 2012;3:55.
2. Lynch M. Intron evolution as a population-genetic process. Proc Natl Acad Sci. 2002;99(9):6118–23.
3. Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. Cold Spring Harb Perspect Biol. 2014;6(6):a016071.
4. Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct. 2006;1(1):22.
5. Lane CE, van den Heuvel K, Kozera C, Curtis BA, Parsons BJ, Bowman S, et al. Nucleomorph genome of Hemiselmis andersenii reveals complete intron loss and compaction as a driver of protein structure and function. Proc Natl Acad Sci. 2007;104(50):19908–13.
6. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017;136(9):1093–111.
7. López-Bigas N, Audit B, Ouzounis C, Parra G, Guigó R. Are splicing mutations the most frequent cause of hereditary disease? FEBS Lett. 2005;579(9):1900–3.
8. Brown J. Arabidopsis intron mutations and pre-mRNA splicing. Plant J Cell Mol Biol. 1996;10(5):771.
9. Bush SJ, Chen L, Tovar-Corona JM, Urrutia AO. Alternative splicing and the evolution of phenotypic novelty. Philos Trans R Soc B Biol Sci. 2017;372(1713):20150474.
10. Tress ML, Abascal F, Valencia A. Most alternative isoforms are not functionally important. Trends Biochem Sci. 2017;42(6):408–10.
11. Gupta SK, Carmi S, Ben-Asher HW, Tkacz ID, Naboishchikov I, Michaeli S. Basal splicing factors regulate the stability of mature mRNAs in trypanosomes. J Biol Chem. 2013;288(7):4991–5006.
12. Martinson HG. An active role for splicing in 3′-end formation. Wiley Interdiscip Rev RNA. 2011;2(4):459–70.
13. Gao X, Qiao Y, Han D, Zhang Y, Ma N. Enemy or partner: relationship between intronic micrornas and their host genes. IUBMB Life. 2012;64(10):835–40.
14. Rose AB. Requirements for intron-mediated enhancement of gene expression in Arabidopsis. RNA. 2002 Nov;8(11):1444–53.
15. Gallegos JE, Rose AB. An intron-derived motif strongly increases gene expression from transcribed sequences through a splicing independent mechanism in Arabidopsis thaliana. Sci Rep. 2019 Dec;9(1):13777.
16. Callis J, Fromm M, Walbot V. Introns increase gene expression in cultured maize cells. Genes Dev. 1987;1(10):1183–200.
17. Crane MM, Sands B, Battaglia C, Johnson B, Yun S, Kaeberlein M, et al. In vivo measurements reveal a single 5′-intron is sufficient to increase protein expression level in Caenorhabditis elegans. Sci Rep. 2019 Dec;9(1):9192.
18. Anastasiadi D. Consistent inverse correlation between DNA methylation of the first intron and gene expression across tissues and species. 2018;17.
19. Gallegos JE, Rose AB. The enduring mystery of intron-mediated enhancement. Plant Sci. 2015 Aug;237:8–15.
20. Rose AB. Introns as Gene Regulators: A Brick on the Accelerator. Front Genet. 2019;9:6.
21. Guiro J, O'reilly D. Insights into the U1 small nuclear ribonucleoprotein complex superfamily. Wiley Interdiscip Rev RNA. 2015;6(1):79–92.
22. Valencia P, Dias AP, Reed R. Splicing promotes rapid and efficient mRNA export in

mammalian cells. Proc Natl Acad Sci. 2008;105(9):3386–91.
23. Shaul O. How introns enhance gene expression. Int J Biochem Cell Biol. 2017 Oct;91:145–55.
24. Morello L, Gianì S, Troina F, Breviario D. Testing the IMEter on rice introns and other aspects of intron-mediated enhancement of gene expression. J Exp Bot. 2011 Jan;62(2):533–44.
25. Laxa M, Müller K, Lange N, Doering L, Pruscha JT, Peterhänsel C. The 5′UTR Intron of Arabidopsis GGT1 Aminotransferase Enhances Promoter Activity by Recruiting RNA Polymerase II. Plant Physiol. 2016 Sep;172(1):313–27.
26. Rose AB, Emami S, Bradnam K, Korf I. Evidence for a DNA-based mechanism of intron-mediated enhancement. Front Plant Sci. 2011;2:98.
27. Akua T, Berezin I, Shaul O. The leader intron of AtMHX can elicit, in the absence of splicing, low-level intron-mediated enhancement that depends on the internal intron sequence. BMC Plant Biol. 2010;10:93.
28. Clancy M, Hannah LC. Splicing of the maize Sh1 first intron is essential for enhancement of gene expression, and a T-rich motif increases expression without affecting splicing. Plant Physiol. 2002;130(2):918–29.
29. Gianì S, Altana A, Campanoni P, Morello L, Breviario D. In trangenic rice, α-and β-tubulin regulatory sequences control GUS amount and distribution through intron mediated enhancement and intron dependent spatial expression. Transgenic Res. 2009;18(2):151–62.
30. Korf IF, Rose AB. Applying Word-Based Algorithms: The IMEter. In: Belostotsky DA, editor. Plant Systems Biology. Totowa, NJ: Humana Press; 2009. p. 287–301. (Methods in Molecular Biology™; vol. 553).
31. Parra G, Bradnam K, Rose AB, Korf I. Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Res. 2011 Jul;39(13):5328–37.
32. Rose AB, Elfersi T, Parra G, Korf I. Promoter-Proximal Introns in *Arabidopsis thaliana* Are Enriched in Dispersed Signals that Elevate Gene Expression. Plant Cell. 2008 Mar;20(3):543–51.
33. Hong RL, Hamaguchi L, Busch MA, Weigel D. Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. Plant Cell. 2003;15(6):1296–309.
34. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell. 2016 Jul;166(2):481–91.
35. Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell. 2016 Jul;166(2):492–505.
36. Korkuc P, Schippers JHM, Walther D. Characterization and Identification of cis-Regulatory Elements in Arabidopsis Based on Single-Nucleotide Polymorphism Information. PLANT Physiol. 2014 Jan 1;164(1):181–200.
37. Greener JG, Sternberg MJE. AlloPred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. BMC Bioinformatics. 2015;16:335.
38. Heckmann D, Lloyd CJ, Mih N, Ha Y, Zielinski DC, Haiman ZB, et al. Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. Nat Commun. 2018;9(1):1–10.
39. Wang X, Lin P, Ho JW. Discovery of cell-type specific DNA motif grammar in cis-regulatory elements using random Forest. BMC Genomics. 2018;19(1):153–60.
40. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The Arabidopsis

information resource: making and mining the “gold standard” annotated reference plant genome. genesis. 2015;53(8):474–85.
41. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinforma. 2014;47(1):11–2.
42. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
44. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
45. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91.
46. Mahony S, Benos PV. STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W253–8.
47. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):D1178–86.
48. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S. NASCArrays: a repository for microarray data generated by NASC’s transcriptomics service. Nucleic Acids Res. 2004;32(suppl_1):D575–7.
49. O’Malley RC, Huang SC, Song L, Lewsey MG, Bartlett A, Nery JR, et al. Cistrome and epicistrome features shape the regulatory DNA landscape. Cell. 2016;165(5):1280–92.
50. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.
51. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013 Aug;45(8):891–8.
52. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
53. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007 Dec;8(1):25.
54. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Adv Neural Inf Process Syst 30. 2017;4765–74.
55. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020 Jan;2(1):56–67.
56. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.
57. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020;
58. Van Der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Comput Sci Eng. 2011;13(2):22.
59. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference. Austin, TX; 2010. p. 51–6.

60. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90.
61. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. mwaskom/seaborn: v0. 8.1 (September 2017). Zenodo Doi. 2017;10.
62. Ponting CP. Biological function in the twilight zone of sequence conservation. BMC Biol. 2017;15(1):1–9.
63. Xie X, Lu J, Kulbokas E, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature. 2005;434(7031):338–45.
64. Amos W. Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence? Proc R Soc B Biol Sci. 2010 May 7;277(1686):1443–9.
65. Rose AB, Carter A, Korf I, Kojima N. Intron sequences that stimulate gene expression in Arabidopsis. Plant Mol Biol. 2016 Oct;92(3):337–46.
66. Akua T, Shaul O. The Arabidopsis thaliana MHX gene includes an intronic element that boosts translation when localized in a 5′ UTR intron. J Exp Bot. 2013;64(14):4255–70.
67. Zafrir Z, Tuller T. Nucleotide sequence composition adjacent to intronic splice sites improves splicing efficiency via its effect on pre-mRNA local folding in fungi. RNA. 2015;21(10):1704–18.
68. Basu A, Bobrovnikov DG, Qureshi Z, Kayikcioglu T, Ngo TTM, Ranjan A, et al. Measuring DNA mechanics on the genome scale. Nature. 2021 Jan;589(7842):462–7.
69. Laxa M. Intron-mediated enhancement: a tool for heterologous gene expression in plants? Front Plant Sci. 2017;7:1977.
70. Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA. A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J. 2016;88(6):1058–70.
71. Koch MA, Matschinger M. Evolution and genetic differentiation among relatives of Arabidopsis thaliana. Proc Natl Acad Sci. 2007 Apr 10;104(15):6272–7.
72. Perlaza-Jiménez L, Walther D. A genome-wide scan for correlated mutations detects macromolecular and chromatin interactions in Arabidopsis thaliana. Nucleic Acids Res. 2018 Sep 19;46(16):8114–32.
73. Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. 2020 Feb;30(2):214–26.

Tables

Table 1. Features used for the prediction of expression level based on Random Forest models.

Feature	Abbreviation	Description
intron length	length	length of the first intron
distance to CDS-start	distance_CDS	distance of the first intron to the translation start codon of its gene
distance to TSS	distance_TSS	distance of the first intron to the transcription start site
IMEter score	imeter	calculated IMEter score of the first intron
SNP ber bp	SNP_per_bp	SNP rate per base pair
DMRs C context	DMR_C	number of differentially methylated areas with CG/CHG/CHH context in the intron
DMRs CG context	DMR_CG	number of differentially methylated areas with CG context in the intron
transposable elements	n_transposon	normalized number of transposable elements in the proximal intron
intron retainment	IR	“1” if first intron is retained in some isoforms as reported in the GFF file, otherwise “0”
CNS	CNS	number of conserved non-coding sequence (CNS) sections in the intron
minimum folding energy	min_fold_energy	normalized minimum folding energy of the first intron
A/T/C/G content	A/T/C/G	base-type occurrence percentage of A/T/C/G of first introns, excluding the splice sites
dimer percentages	TA/CG...	relative frequency of all possible dimers in the first intron, with reverse complement dimers combined. Splice sites are excluded

Table 2. Hexamers with potential regulatory function as evidenced by increased conservation, positional preferences, and co-expression of genes harboring respective motifs in their first introns. 'Cohen's d correlation' is the effect size of difference in the distribution of correlation coefficients between the expression levels of genes harboring the respective motif relative to a gene set containing frequency-matched random hexamer motifs across all experimental conditions present in the expression dataset. 'Cohen's d expression level' refers to the effect size related to expression level of genes containing the respective motif in the first intron relative to all other intron-harboring genes. Listed also are the numbers of genes, in which the respective intron motif was found. Listed are all motifs with 'Cohen's d expression'>0.05. For a complete listing of all 81 candidate motifs, identified based on conservation and positional preference alone, see Supplementary Table 1.

Hexamer	Cohen's d, Correlation, comparable, random hexamer	Cohen's d, Expression level	Number of genes
AGATCG	1.45E-01	0.46	1807
ACCCTA	9.82E-02	0.18	2964
TCGATC	9.16E-02	0.34	2014
TCGGAG	8.58E-02	0.27	857
TCTCGC	8.13E-02	0.19	785
GATTCG	7.68E-02	0.32	2516
ATCGAA	7.07E-02	0.31	4188
AAATCG	7.00E-02	0.28	4086
AATCGA	6.88E-02	0.31	4406
TTAGGG	6.76E-02	0.19	2896
ATCGAG	6.20E-02	0.28	1773
TCTCGA	5.79E-02	0.22	2044
CTCTCG	5.77E-02	0.23	1124
AAACCC	5.33E-02	0.18	4970
TTCTCG	5.27E-02	0.19	2188
TTTCGA	5.20E-02	0.21	3866

Figure legends

Figure 1. **Comparison of SNP-frequencies of intron subsets. (a)** Average relative SNP-frequency of the first 20 bp of the first introns compared to the other introns including the last 20 bp of the preceding exons **(b)** Average relative SNP-frequency of the first 20 bp of first introns in 5'-UTRs compared to first introns in CDS including the last 20 bp of the preceding exons **(c)** Comparison of confidence intervals for the average SNP-frequency per bp (SNP-density) of different intron subsets **(d)** Violin plot summing up SNP-frequencies per bp (SNP-densities) of different intron subsets. For **(a)** and **(b)** positions are relative to the exons-intron junction with zero denoting first intron position.

Figure 2. Hexamer characteristics. Conservation and occurrence of hexamers in **(a)** first introns, **(b)** other introns, **(c)** Comparison of hexamers relative occurrence distributions of hexamers that occur more (blue, top x-axis)/ less (orange, bottom x-axis) often in first than in other introns. In **(a)** and **(b)**, for definition of conservation, see Methods. Every dot represents a hexamer, the red line represents a computed running average, and the dashed black line corresponds to the respective estimated random conservation based on Eq. 2.

Figure 3. Consensus motifs and comparison to IMEter motifs. (a) Consensus motifs generated by Neighbor Joining Tree of MAFFT alignments of the 16 candidate hexamers with clusters indicated as boxes. The clustering threshold for collapsing motifs into consensus motifs, shown below to the respectively clustered motifs, was set based on visual inspection. Consensus motifs were required to be supported by two or more bases, i.e. vertically aligned motif positions. **(b)** Sequence comparison of the five consensus motifs and the two previously reported IMEter motifs (indicated by *). Dendrogram was created using the tool Stamp. Branch length proportional to distance.

Figure 4. Comparison of the five identified consensus motifs to IMEter. Density plot of intra-set Pearson correlation of gene expression comparing **(a),** the 2000 genes with the highest IMEter score (top_imeter) to a subset of random genes of equal size, **(b),** genes containing of one of the five identified candidate motifs and genes containing the IMEter-derived motifs TTNGATYTG and CGATT, designated by a "*". Each gene set was compared to a subset of genes with the highest IMEter score of equal set size and their difference expressed as Cohen's d effect size, given in the graph titles, with negative values suggesting smaller effect in motif-set relative to size-matched IMEter-score set. **(c)** same as in (b), but with overlapping genes (genes found in both sets) removed from the IMEter set. Intra-set Pearson correlation refers to the determined Pearons correlation coefficients of all possible gene pairs in sets of genes harboring the same motif or belonging to the same top-IMEter set.

Figure 5. Prediction of gene expression levels based on intron features. Random Forest Performance and Feature Importance. (a) 10-fold Cross-validated ROC curves for Random Forests trained with the median-split (whole) set and quartile expression set, respectively, **(b)** MDA feature importance for Random Forest model trained with the lower and upper quartile expression dataset, For feature explanations, see Table 1. **(c)** SHAP summary plot of Random Forest model trained with the lower and upper quartile expression dataset, **(d)** SHAP value to feature value plot for distance to TSS, with the respective distance to CDS-start values color-coded. Positive SHAP values indicate an associate with the high expression class, Negative SHAP, association with the low expression class of genes.

Figure 6. Prediction of gene expression levels based on intron and exon features. Random Forest Performance and Feature Importance. (a) Tenfold cross-validated ROC curves for Random Forests trained with intron-only, exon-only features, and both sets combined for the upper/lower quartile data set. **(b)** MDA feature importance for a Random Forest model trained with combined exon and intron features for the upper/lower quartile expression data set. **(c)** SHAP summary plot containing the 20 features with highest importance of the Random Forest model trained with combined exon and intron features for lower and upper quartile dataset. Exon features were extracted from the respective first exons of genes, as were intron-features extracted from first introns.

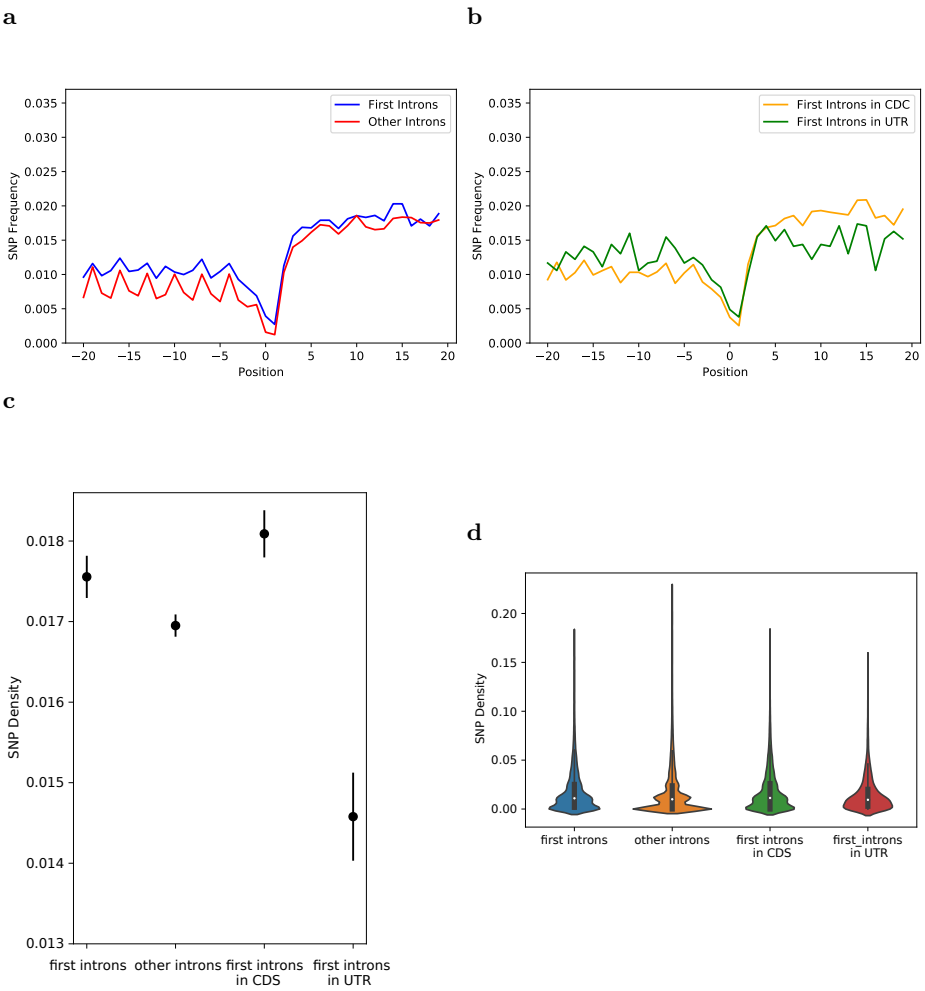

a
b
First Introns
Other Introns
First Introns in CDC
First Introns in UTR
SNP Frequency
Position
c
d
SNP Density
first introns
other introns
first introns in CDS
first introns in UTR
first_introns in UTR

Figure 1

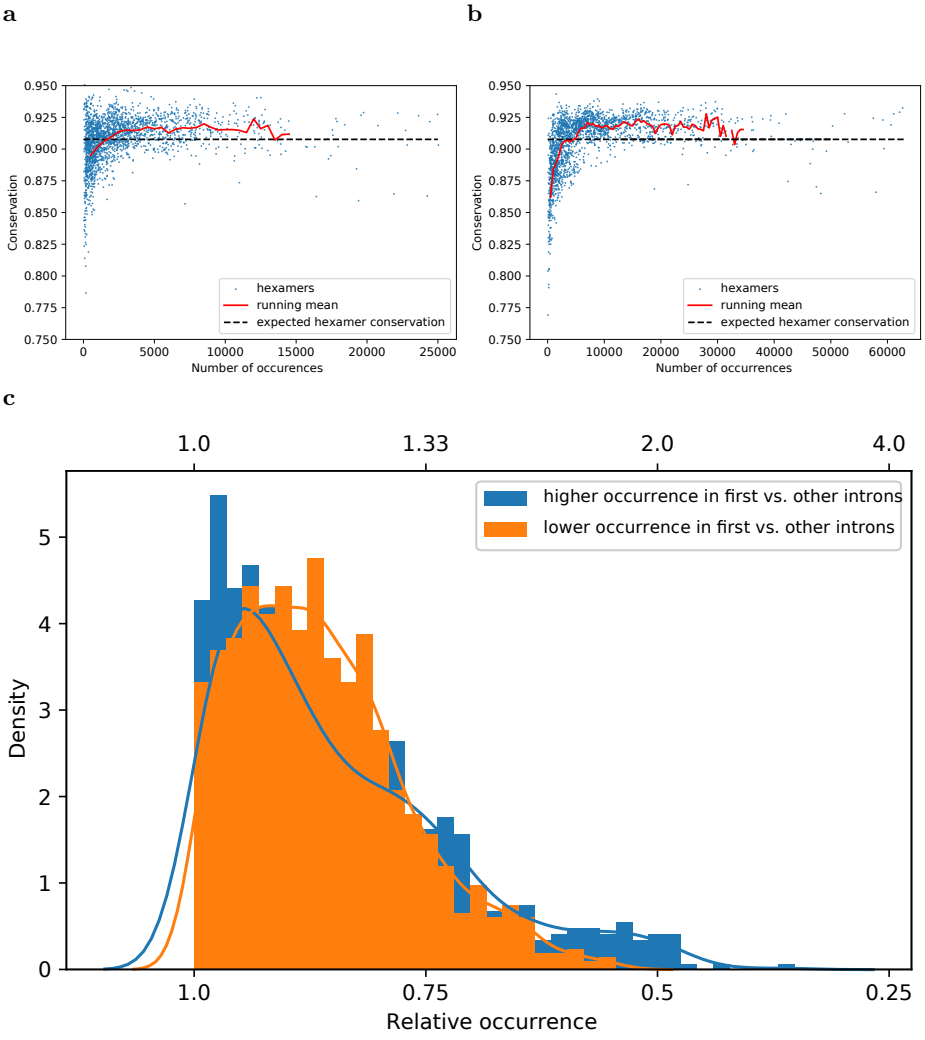
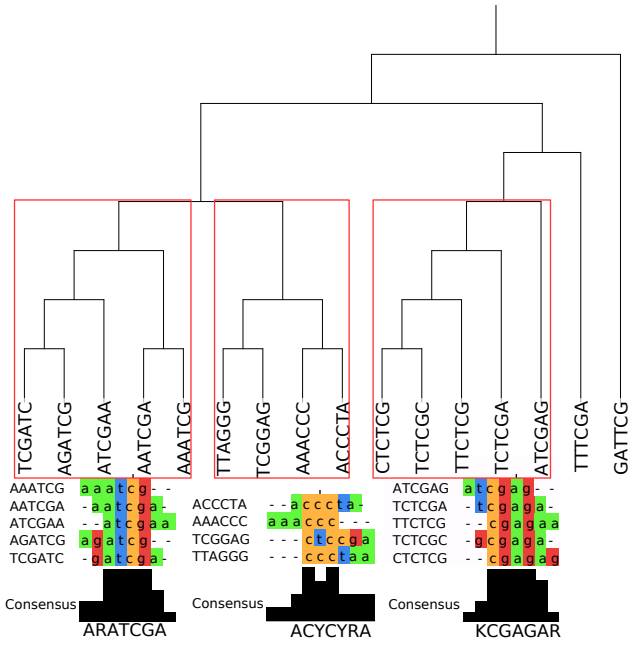

a
b
Conservation
Number of occurences
Number of occurrences
hexamers
running mean
expected hexamer conservation
c
1.0
1.33
2.0
4.0
higher occurrence in first vs. other introns
lower occurrence in first vs. other introns
Density
Relative occurrence
0.75
0.5
0.25

Figure 2

a


TCGATC
AGATCG
ATCGAA
AATCGA
AAATCG
TTAGGG
TCGGAG
AAACCC
ACCCTA
CTCTCG
TCTCGC
TTCTCG
TCTCGA
ATCGAG
TTTCGA
GATTCG
AAATCG
AATCGA
ATCGAA
AGATCG
TCGATC
ACCCTA
AAACCC
TCGGAG
TTAGGG
ATCGAG
TCTCGA
TTCTCG
TCTCGC
CTCTCG
Consensus
ARATCGA
Consensus
ACYCYRA
Consensus
KCGAGAR

b

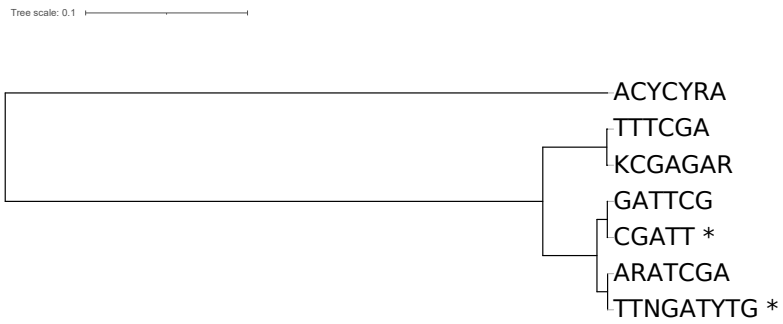
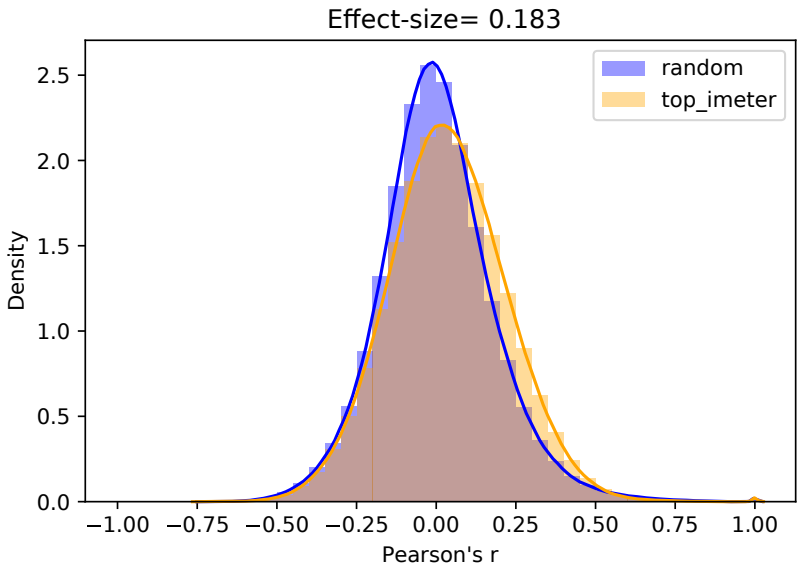
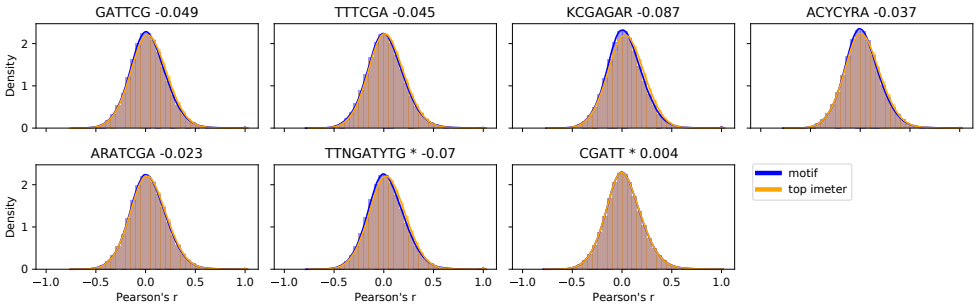

Tree scale: 0.1
ACYCYRA
TTTCGA
KCGAGAR
GATTCG
CGATT *
ARATCGA
TTNGATYTG *

Figure 3

a


Effect-size= 0.183
random
top_imeter
Density
Pearson's r

b


GATTCG -0.049
TTTCGA -0.045
KCGAGAR -0.087
ACYCYRA -0.037
ARATCGA -0.023
TTNGATYTG * -0.07
CGATT * 0.004
motif
top imeter
Density
Pearson's r

c

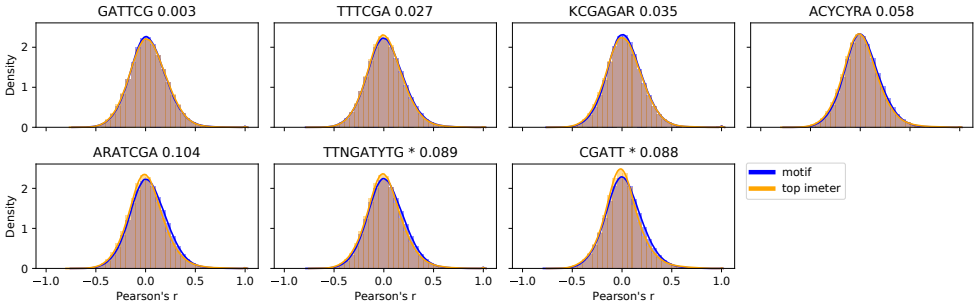
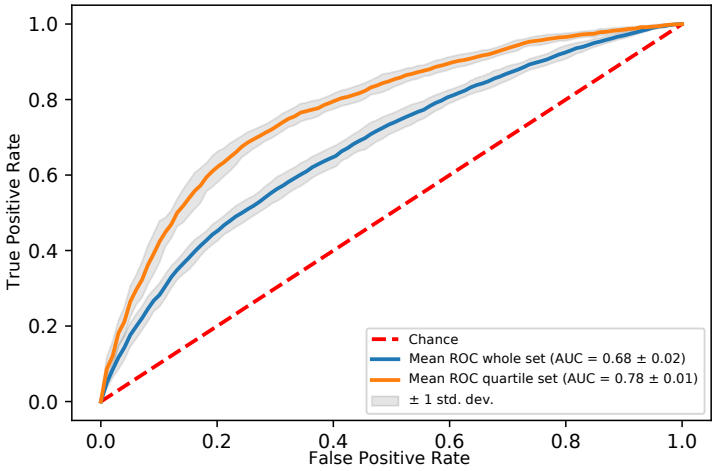
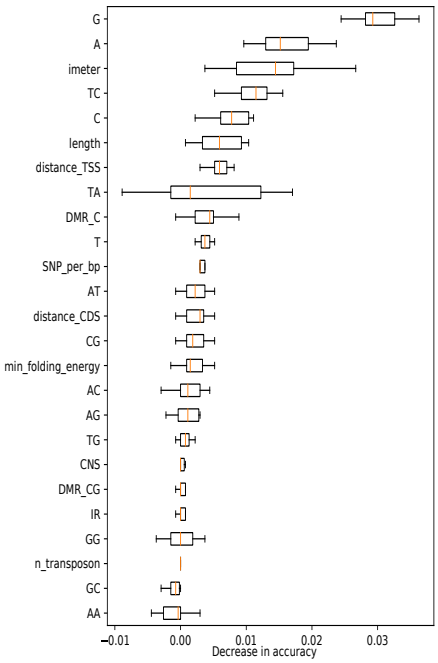

GATTCG 0.003
TTTCGA 0.027
KCGAGAR 0.035
ACYCYRA 0.058
ARATCGA 0.104
TTNGATYTG * 0.089
CGATT * 0.088
motif
top imeter
Density
Pearson's r

Figure 4

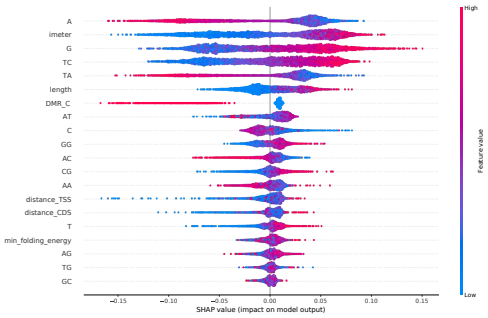
a



b



c



d

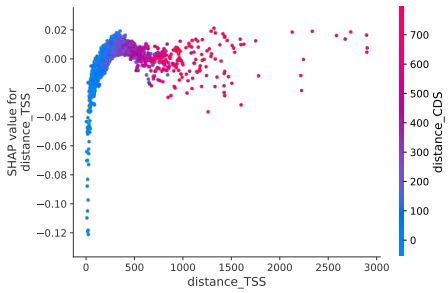
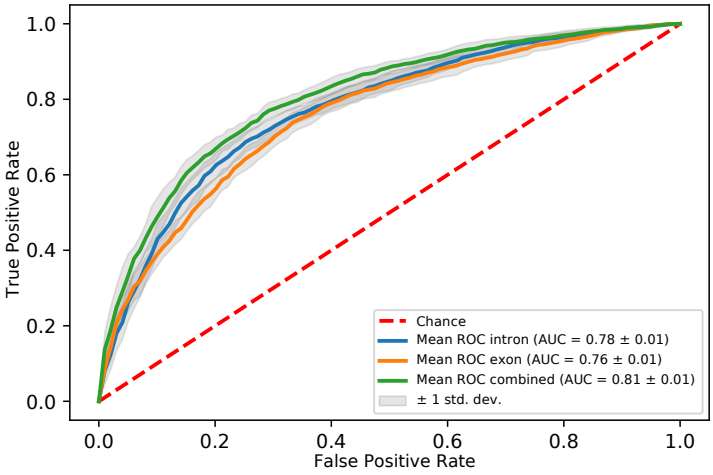
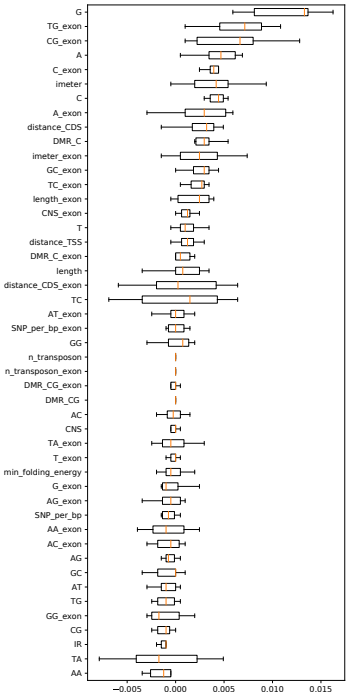


Figure 5

a


True Positive Rate
False Positive Rate
Chance
Mean ROC intron (AUC = 0.78 ± 0.01)
Mean ROC exon (AUC = 0.76 ± 0.01)
Mean ROC combined (AUC = 0.81 ± 0.01)
± 1 std. dev.

b


G
TG_exon
CG_exon
A
C_exon
imeter
C
A_exon
distance_CDS
DMR_C
imeter_exon
GC_exon
TC_exon
length_exon
CNS_exon
T
distance_TSS
DMR_C_exon
length
distance_CDS_exon
TC
AT_exon
SNP_per_bp_exon
GG
n_transposon
n_transposon_exon
DMR_CG_exon
DMR_CG
AC
CNS
TA_exon
T_exon
min_folding_energy
G_exon
AG_exon
SNP_per_bp
AA_exon
AC_exon
AG
GC
AT
TG
GG_exon
CG
IR
TA
AA

c

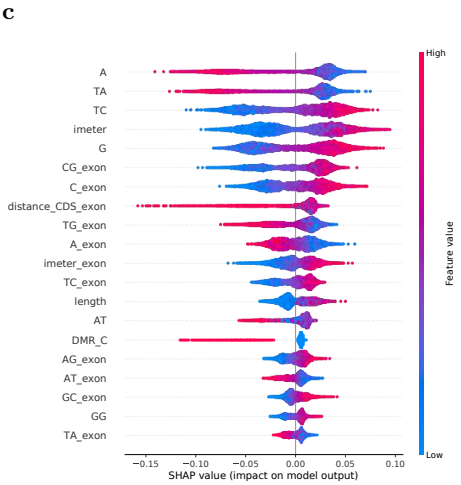

A
TA
TC
imeter
G
CG_exon
C_exon
distance_CDS_exon
TG_exon
A_exon
imeter_exon
TC_exon
length
AT
DMR_C
AG_exon
AT_exon
GC_exon
GG
TA_exon
SHAP value (impact on model output)
Feature value
High
Low

Figure 6

Figures

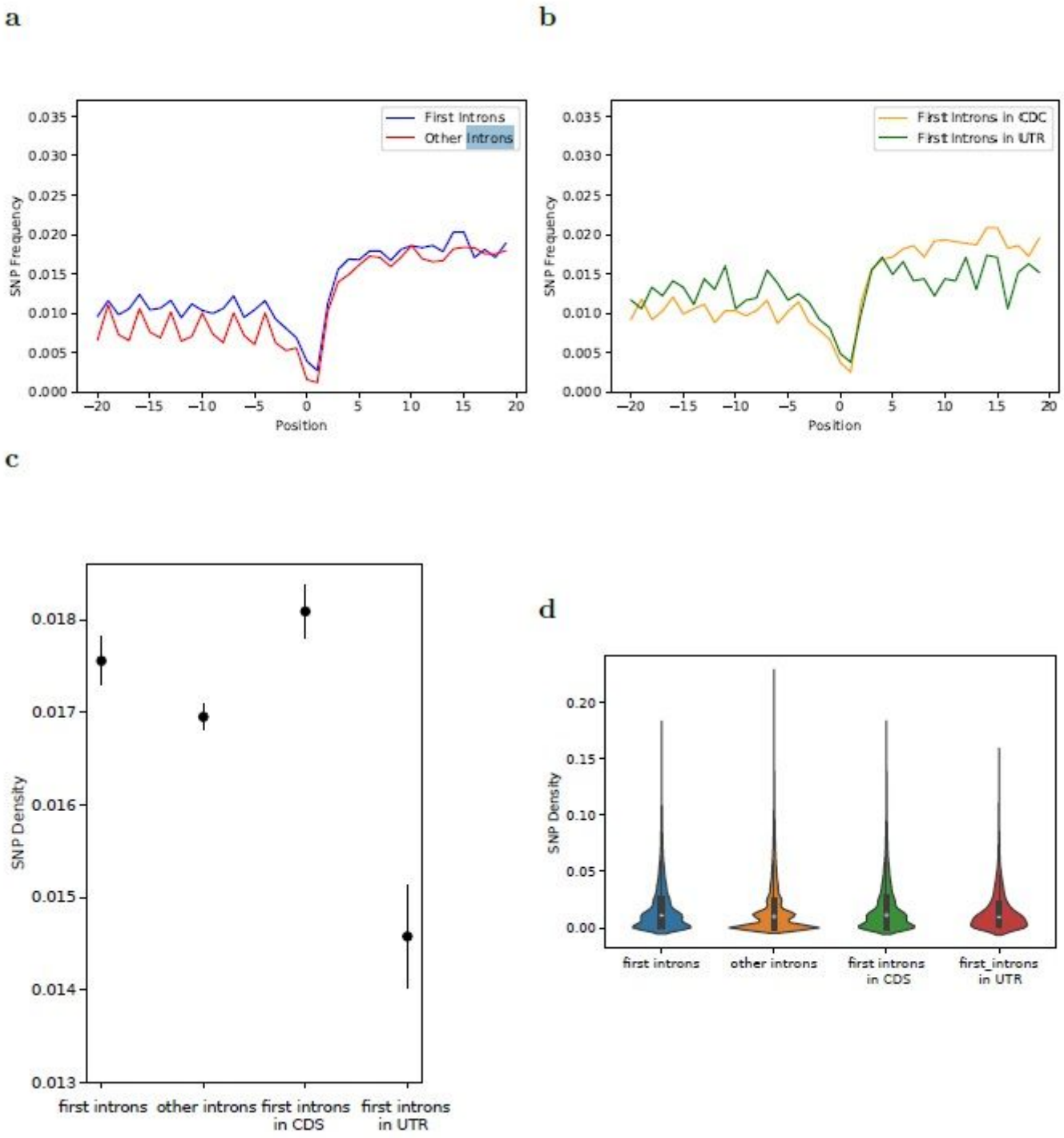


Figure 1

Comparison of SNP-frequencies of intron subsets. (a) Average relative SNP-frequency of the first 20 bp of the first introns compared to the other introns including the last 20 bp of the preceding exons (b) Average relative SNP-frequency of the first 20 bp of first introns in 5'-UTRs compared to first introns in CDS including the last 20 bp of the preceding exons (c) Comparison of confidence intervals for the average

SNP-frequency per bp (SNP-density) of different intron subsets (d) Violin plot summing up SNP-frequencies per bp (SNP-densities) of different intron subsets. For (a) and (b) positions are relative to the exons-intron junction with zero denoting first intron position.

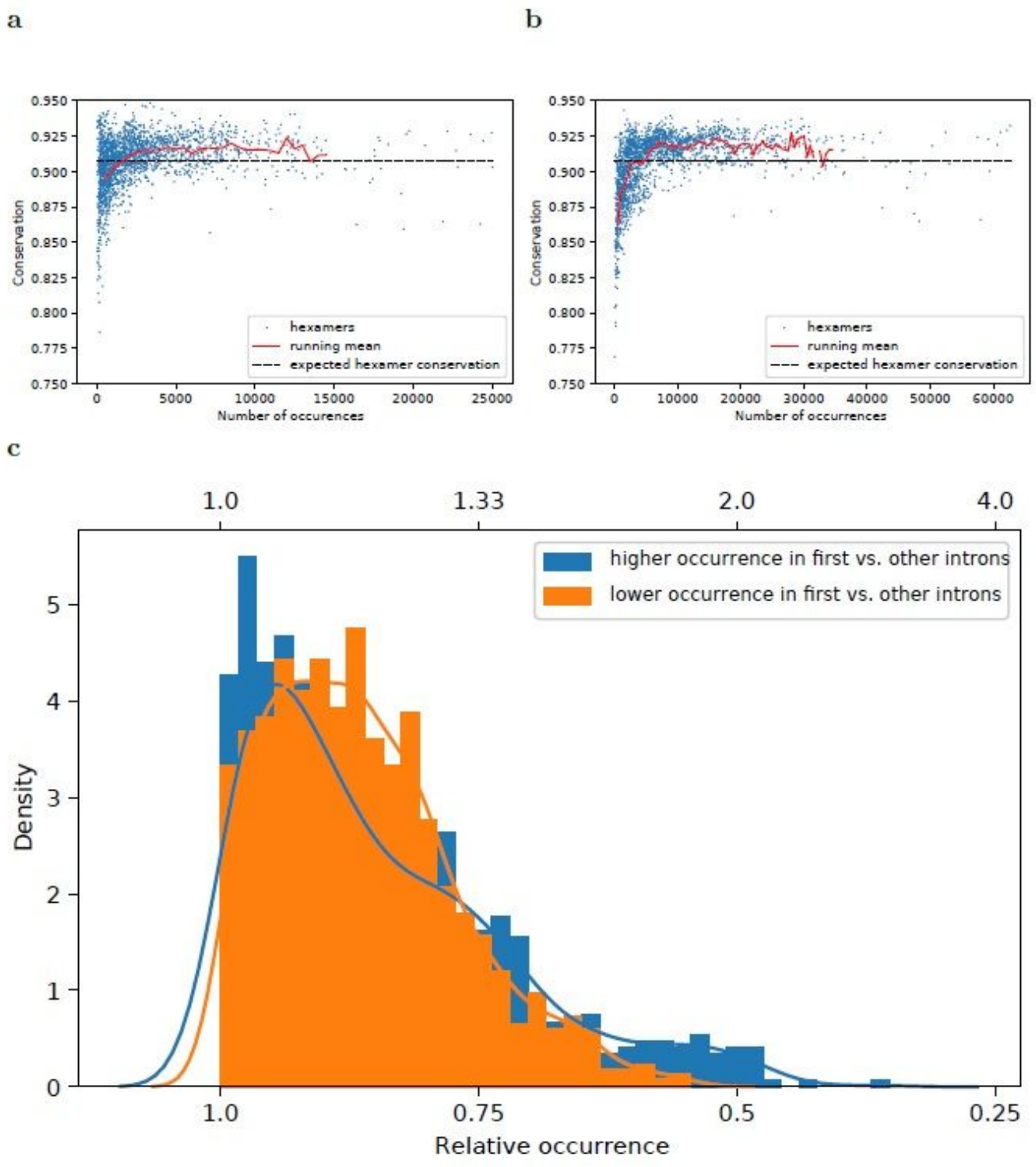


Figure 2

Hexamer characteristics. Conservation and occurrence of hexamers in (a) first introns, (b) other introns, (c) Comparison of hexamers relative occurrence distributions of hexamers that occur more (blue, top x-

axis)/ less (orange, bottom x-axis) often in first than in other introns. In (a) and (b) , for definition of conservation, see Methods. Every dot represents a hexamer, the red line represents a computed running average, and the dashed black line corresponds to the respective estimated random conservation based on Eq. 2.

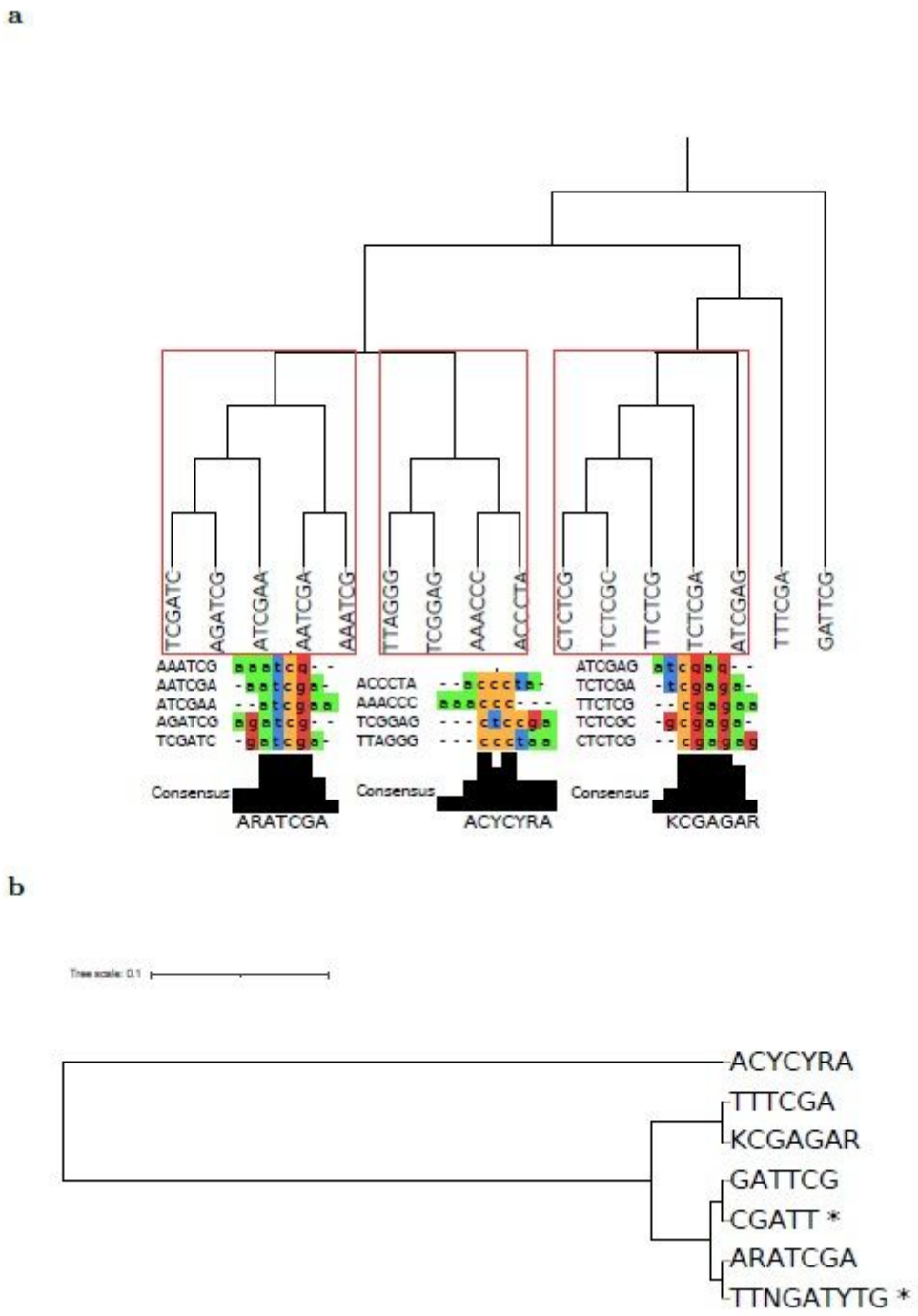


Figure 3

Consensus motifs and comparison to IMEter motifs. (a) Consensus motifs generated by Neighbor Joining Tree of MAFFT alignments of the 16 candidate hexamers with clusters indicated as boxes. The clustering threshold for collapsing motifs into consensus motifs, shown below to the respectively

clustered motifs, was set based on visual inspection. Consensus motifs were required to be supported by two or more bases, i.e. vertically aligned motif positions. (b) Sequence comparison of the five consensus motifs and the two previously reported IMEter motifs (indicated by *). Dendrogram was created using the tool Stamp. Branch length proportional to distance.

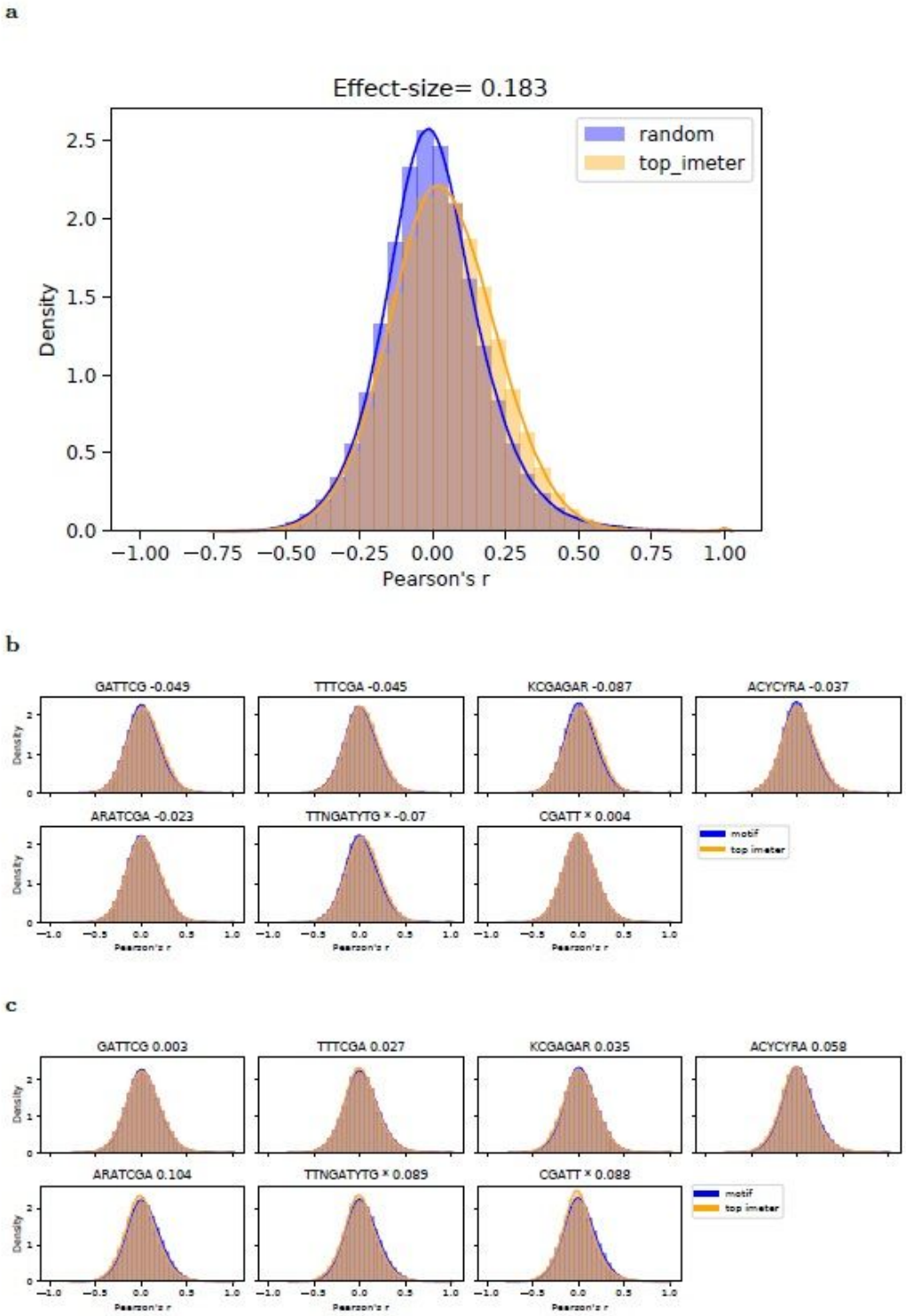


Figure 4

Comparison of the five identified consensus motifs to IMEter. Density plot of intra-set Pearson correlation of gene expression comparing (a), the 2000 genes with the highest IMEter score (top_imeter) to a subset of random genes of equal size, (b), genes containing of one of the five identified candidate motifs and genes containing the IMEter-derived motifs TTNGATYTG and CGATT, designated by a “*”. Each gene set was compared to a subset of genes with the highest IMEter score of equal set size and their difference expressed as Cohen’s d effect size, given in the graph titles, with negative values suggesting smaller effect in motif-set relative to size-matched IMEter-score set. (c) same as in (b), but with overlapping genes (genes found in both sets) removed from the IMEter set. Intra-set Pearson correlation refers to the determined Pearons correlation coefficients of all possible gene pairs in sets of genes harboring the same motif or belonging to the same top-IMEter set.

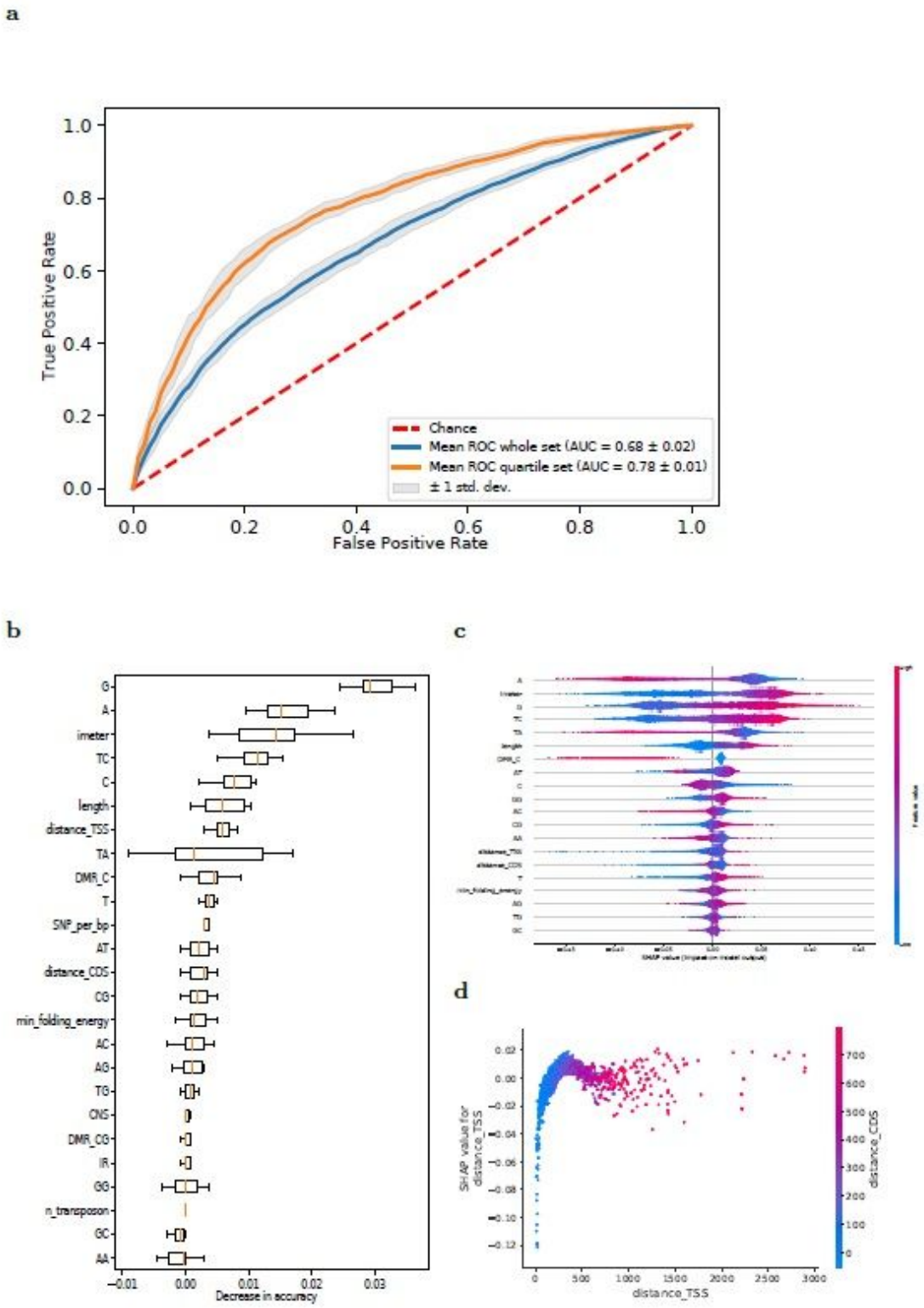


Figure 5

Prediction of gene expression levels based on intron features. Random Forest Performance and Feature Importance. (a) 10-fold Cross-validated ROC curves for Random Forests trained with the median-split (whole) set and quartile expression set, respectively, (b) MDA feature importance for Random Forest model trained with the lower and upper quartile expression dataset, For feature explanations, see Table 1. (c) SHAP summary plot of Random Forest model trained with the lower and upper quartile expression

dataset, (d) SHAP value to feature value plot for distance to TSS, with the respective distance to CDS-start values color-coded. Positive SHAP values indicate an associate with the high expression class, Negative SHAP, association with the low expression class of genes.

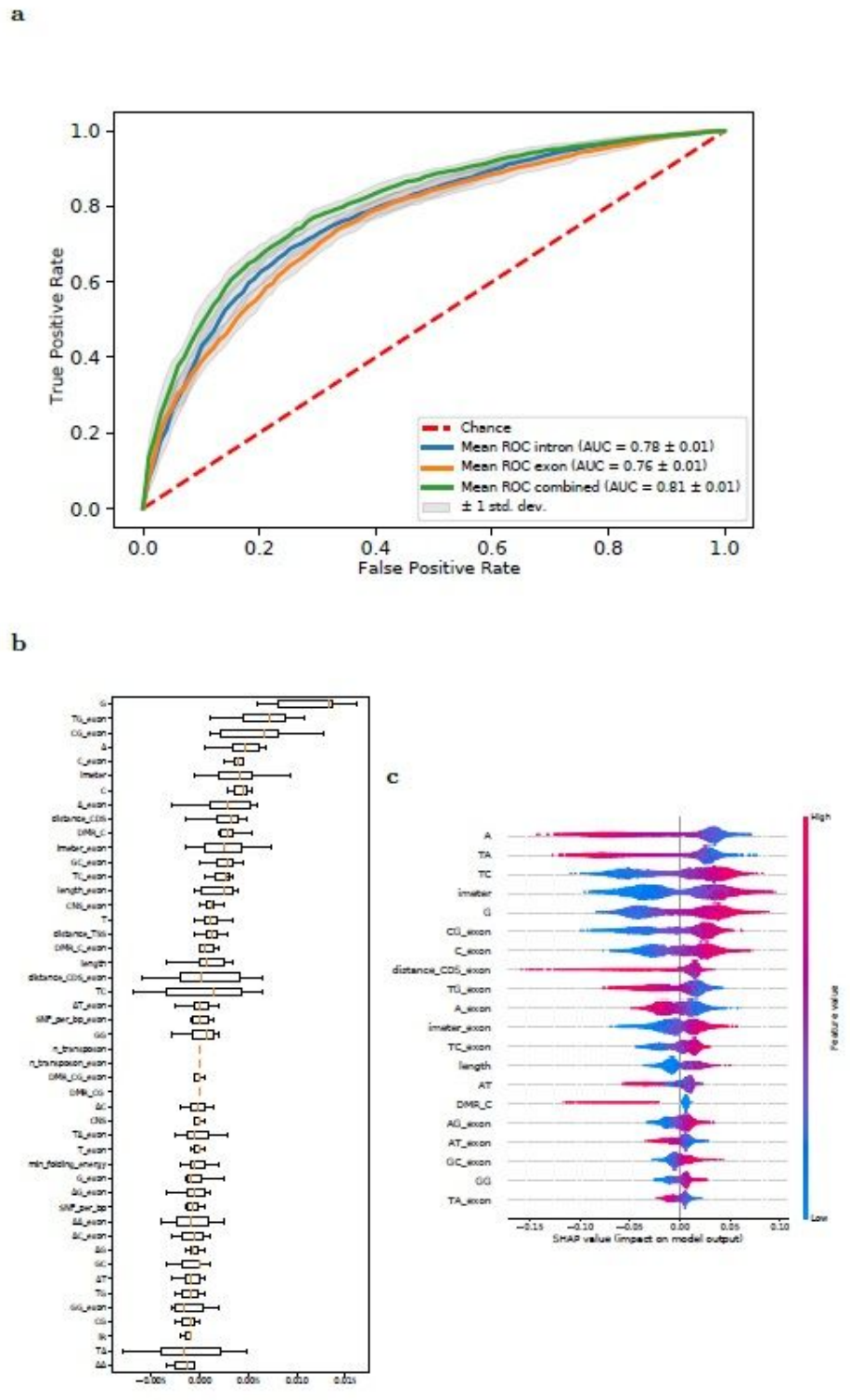


Figure 6

Prediction of gene expression levels based on intron and exon features. Random Forest Performance and Feature Importance. (a) Tenfold cross-validated ROC curves for Random Forests trained with intron-only,

exon-only features, and both sets combined for the upper/lower quartile data set. (b) MDA feature importance for a Random Forest model trained with combined exon and intron features for the upper/lower quartile expression data set. (c) SHAP summary plot containing the 20 features with highest importance of the Random Forest model trained with combined exon and intron features for lower and upper quartile dataset. Exon features were extracted from the respective first exons of genes, as were intron-features extracted from first introns.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- IntronPaperSuppl.pdf