# Uncontrolled Enhancement: How Introns Differ in Widespread and Niche Genes

*Zhe Zheng, Ian Korf*

## Abstract

Intron mediated enhancement (IME) is the phenomenon in which certain introns increase expression in the associated gene. To measure how introns affect gene expression, researchers have developed the IMEter, which uses a language-like model to predict which introns are likely to cause IME in genes. This opens up possibilities of using introns to increase expression in specific genes but poses additional problems. In particular, transgenes fitted with powerful introns have been shown to lose tissue specific gene expression in addition to increased expression. To determine if this is a general phenomenon, we examined genes in various genomes and compared their IMEter scores to their gene expression and tissue specificity. In line with our prediction, we find that genes with high IMEter scores are also broadly expressed across multiple tissues. This supports previous interpretations that IME signals provide a non-specific boost in gene expression.

## Introduction

Although seemingly useless, introns provide multiple roles in gene expression, allowing for the production of new proteins through both exon shuffling (Long et al., 1995) and alternative

splicing (Maniatis & Tasic, 2002). In addition to their role in allowing for alternative mRNA transcripts, however, introns also directly affect gene expression of the associated gene transcripts in a phenomenon known as intron-mediated enhancement (IME). Introns with IME content typically boost expression anywhere between 2- to 10-fold and have been observed to do so significantly more in certain organisms, such as monocots and plants (Maas et al., 1991). The ability of introns with IME content to increase expression of their associated genes comes with immense practical applications in biotechnology. By incorporating introns that enhance gene expression, researchers have been able to achieve higher levels of protein production in various organisms. In several studies, the inclusion of specific introns in transgenes resulted in substantial increases in gene expression and protein yield, with the potential for industrial and agricultural applications (Rose, 2002). In particular, introns significantly increased luciferase activity and mRNA accumulation compared to intronless controls in barley, demonstrating IME as a valuable tool for achieving high and stable transgene expression in crops (Bartlett et al., 2009).

Despite its potential applications and upsides, IME's mechanisms are not completely understood and remain a current object of research, and not all introns show an ability to induce IME (Gallegos & Rose, 2015). However, researchers have identified several factors that make introns more likely to affect gene expression. In particular, introns must be positioned close to the start of transcription to have an effect on expression, and those placed upstream of the promoter or far downstream from the start site do not exhibit the same enhancement effects (Gallegos & Rose, 2018). In addition, researchers have found that certain motifs within introns are critical to their enhancing effect; such sequences are also over-represented in promoter-proximal introns and can

convert a non-stimulating intron into one that significantly increases mRNA accumulation (Gallegos & Rose, 2019). Using these features, researchers developed the IMEter, a tool that uses existing introns and their positions on the genome to predict whether new introns have IME content (Rose et al., 2008).

In many cases, inserting introns into genes have caused a wide array of unpredictable and sometimes disruptive effects. An experiment that incorporated an intron from the UBQ10 gene with known IME content did cause gene expression to increase in all cases compared to intronless controls, but effects were varied. Interestingly, the insertion of the UBQ10 intron altered the expression patterns of some genes; for example, it caused strong root activity in genes that did not normally express in roots, indicating a loss of tissue specificity (Emami et al., 2013).

The loss of tissue specificity observed in transgenes with introns containing IME content brings to mind the idea that IME content may differ among genes based on how specific their gene expression is among tissues in the organism. In line with the experiment in Emami et al. (2013), genes with high specificity should have introns with low IME content, as introns with high IME content would cause such genes to lose their specificity.

# Materials and Methods

## Materials

To conduct our experiment, we employed the use of several databases and computational tools. To analyze the intron content in the genes and predict their IME content, we reimplemented the

IMEter (Rose et al., 2008) as described in its initial release, with slight modifications to its code, thresholds, and methods of calculations. For our databases, we analyzed the genomes of *A. thaliana*, obtained from The Arabidopsis Information Resource (TAIR) at https://www.arabidopsis.org/, and *O. sativa*, with gene expression data obtained from the Rice Genome Annotation Project (Trapnell et al., 2010) at http://rice.uga.edu/expression.shtml. The expression information was downloaded separately from the intron data for each gene and then subsequently spliced together into a single .txt document for easier access and use. Code was written in Python and graphs were generated with the help of various modules, while data processing was performed in a Linux environment with command-line scripts. All code and data can be accessed from the project GitHub repository https://github.com/k-zhng/imeval.

## Methods

To investigate whether genes with low specificity had higher IME content compared to genes with more specific expression across tissues, we analyzed the genome of *Arabidopsis thaliana* (Berardini et al., 2015) and compared the IMEter scores of introns in various genes with the expression specificity of the associated gene across different tissues. To do so, we first re-implemented the IMEter as described in its initial release (Rose et al., 2008).

### IMEter

Based on the idea that promoter-proximal introns had higher effects on gene expression while promoter-distal introns were less likely to contain IME content, we first created a list of all possible 5-mers (substrings in the genome of length 5, for a total of $4^5 = 1024$ 5-mers) and iterated through the introns of the entire *A. thaliana* genome. An arbitrary threshold of 400 was

set to be the cutoff between the promoter-proximal and promoter-distal regions of introns. Each occurrence of a 5-mer was counted as an occurrence in the promoter-proximal or promoter-distal regions of the introns based on the set threshold. As an improvement upon the original IMEter, each k-mer was evaluated separately as a promoter-proximal or promoter-distal intron, as compared to the original model, which categorized each intron as proximal or distal based on the beginning of its associated gene. At the end, the proximal and distal counts of each 5-mer were normalized to frequencies of the total promoter-proximal and promoter-distal counts and an IMEter score was calculated for each 5-mer based on $\log_2\left(\frac{P_{wi}}{Q_{wi}}\right)$, with $P_{wi}$ being the frequency of the 5-mer in promoter-proximal introns and $Q_{wi}$ being the frequency of the 5-mer in promoter-distal introns.

Following the initialization of k-mers IMEter scores, we then re-iterated through introns in the *A. thaliana* genome to generate an IMEter score for each intron. To do so, we iterated through each intron and summed the IMEter scores for each 5-mer in that intron to output an IMEter score that represented how likely it was that the intron in question affected expression and contained IME content. The process could be expressed as follows:

$$S = \sum_{i=1+D}^{i \leq L-K-A} \log\left(\frac{P_{wi}}{Q_{wi}}\right)$$

in which $S$ is the IMEter score, $i$ the position of the sequence, $L$ the length of the intron, $K$ the k-mer size, $w_i$ a k-mer of length $K$ starting from position $i$, $D$ the length of the splice donor site

consensus (usually 5), $A$ the length of the splice acceptor site consensus (usually 10), $P$ the frequency distribution of k-mers of size $K$ in promoter-proximal introns, and $Q$ the frequency distribution of k-mers of size $K$ in promoter-distal introns.

## Entropy

Following re-implementation of the IMEter, we sought to quantify the specificity of genes. To do so, we employed Shannon entropy, which measures uncertainty and randomness in a set of probabilities, to determine how specific or broad gene expression levels were across twelve tissues in *A. thaliana*.

To calculate Shannon entropy, given the expression levels $(e_1, e_2, \ldots, e_n)$ of a gene in $n$ different tissues, we first normalize the expression levels of the gene in each tissue as a frequency of the total expression levels in that tissue across all genes by summing the expression levels in that tissue across the genome and then dividing the expression level in the tissue in the particular gene by the total sum. Following that, Shannon entropy $H$ can be calculated as follows:

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

for which $n$ is the number of tissues, and $p_i$ is the probability distribution of normalized expression levels across the tissues in gene $i$, as described by:

$$p_i = \frac{x_i}{\sum_{j=1}^{n} x_j}$$

where $x_i$ is the normalized expression level in tissue $j$. Substituting $p_i$ into the entropy formula, we get:

$$H = -\sum_{i=1}^{n} \left( \frac{x_i}{\sum_{j=1}^{n} x_j} \right) \log_2 \left( \frac{x_i}{\sum_{j=1}^{n} x_j} \right)$$

Low entropy values indicated that the expression of the gene was concentrated in a few tissues, making it highly specific, while high entropy values meant that gene expression was spread out more evenly across different tissues and suggested that the gene was broadly expressed and not specific.

## Total Expression

We were also interested in analyzing the total expression levels of different genes in their organisms when compared with their IMEter scores and entropy values. To do so, we summed the expression levels of each gene across all tissues in the organism and then implemented a natural log scale to more clearly see the contrast in expression across genes. This could be denoted as:

$$\log \left( \sum_{t=1}^{T} E(g, t) \right)$$

where $E(g,\ t)$ is the expression level of gene $g$ in tissue $t$ and $T$ the total number of tissues.

# Results

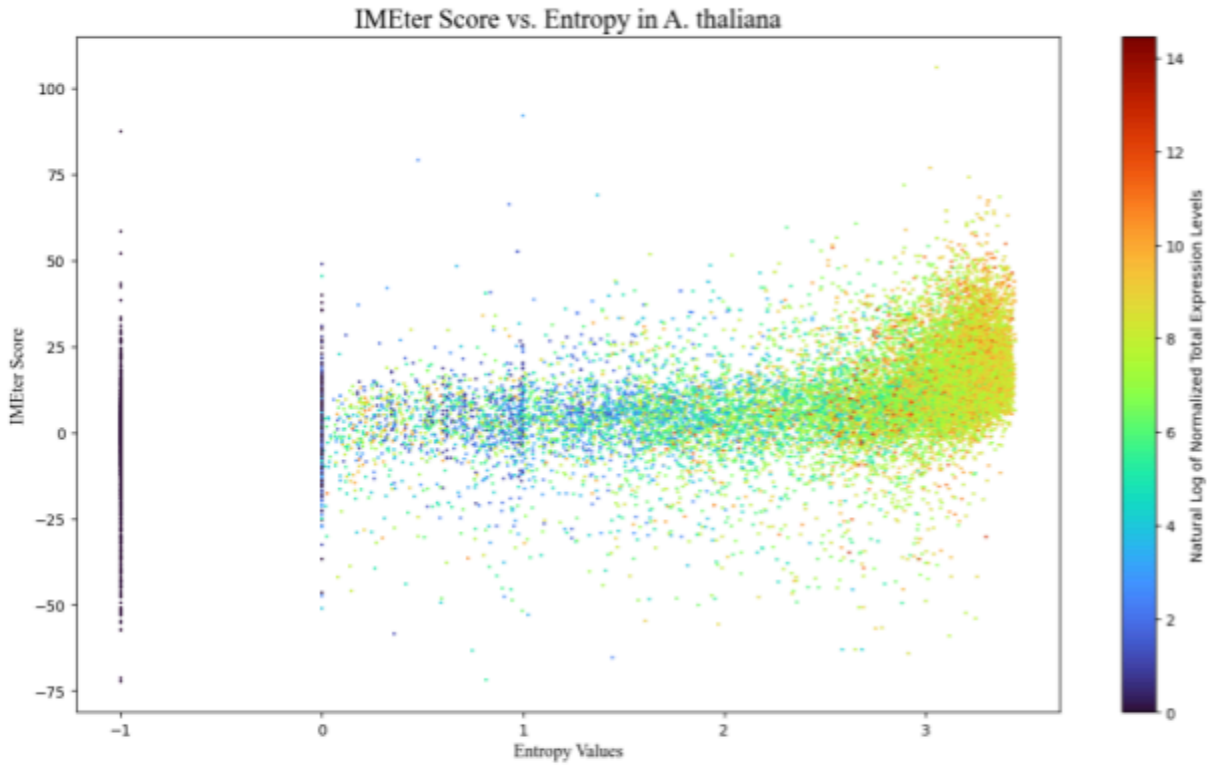


**Figure 1** Graph of IMEter score vs. entropy values in *A. thaliana*. Each data point represents a single gene. The total expression levels of each gene is expressed as a color scale for each data point. Genes with no expression were said to have an entropy of -1.

IMEter scores were graphed with respect to the calculated entropy values for each gene in *A. thaliana*. In general, as entropy values increased, the total expression of the gene also increased. In addition, we observed that various genes were highly expressed but had lower entropy values, signifying that their expression was primarily in certain tissues. There was also a noticeable uptick overall in IMEter scores as expression entropy increased. In general, although total expression levels varied among genes with lower entropy values, those with higher entropy values—and thus more broadly expressed—had relatively high total expression.

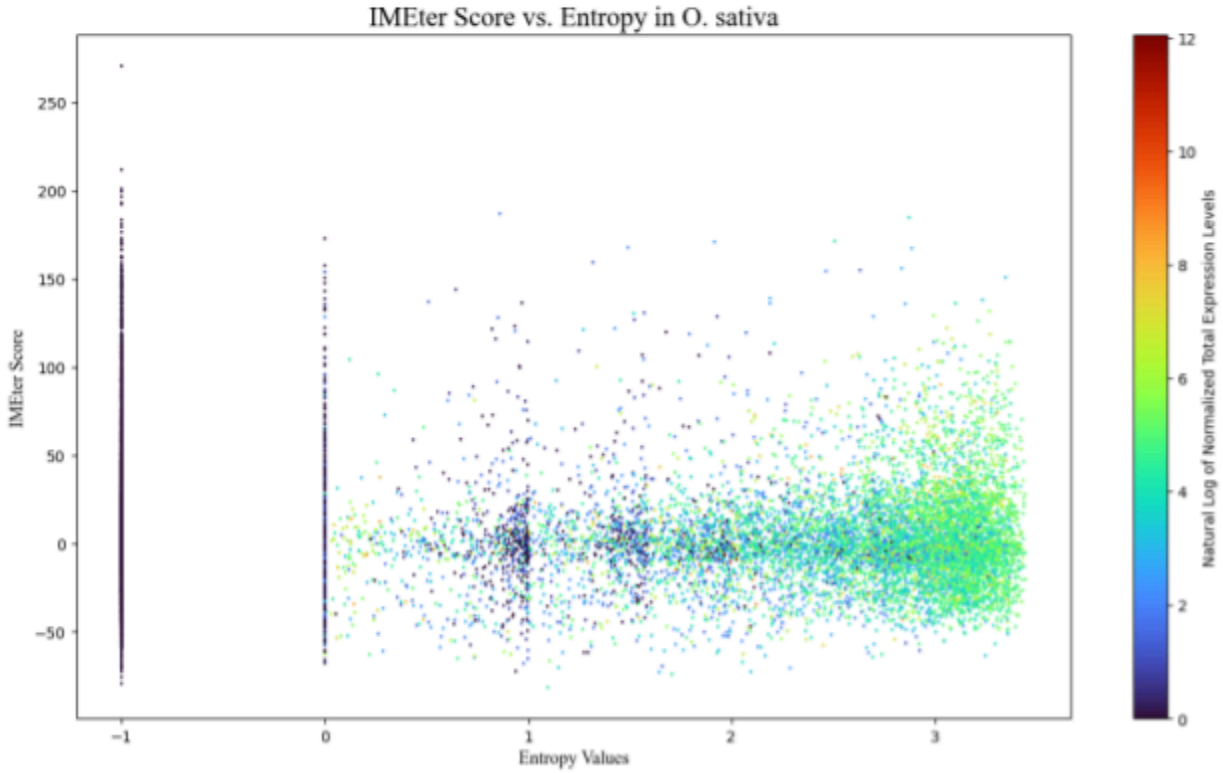


**Figure 2** Graph of IMEter score vs. entropy values in *O. sativa*. The trend is less clear as compared to *A. thaliana*, but it is obvious that genes with higher entropy continue to have higher IMEter scores in general than those with lower entropy.

The experiment was repeated with the introns and expression levels of genes in *O. sativa*, a common rice strain. Although not as noticeable as *A thaliana*, the same trends persisted in our second organism. Notably, total expression levels continued to increase as genes became more broadly expressed and their entropy scores increased. A slight increase in IMEter score was also observed as expression broadened, although the trend was less noticeable than in *A. thaliana*.

To ensure that the trends observed were true and not a result of an increased number of genes, histograms were made of the number of genes with respect to IMEter scores, entropy values, and total expression.

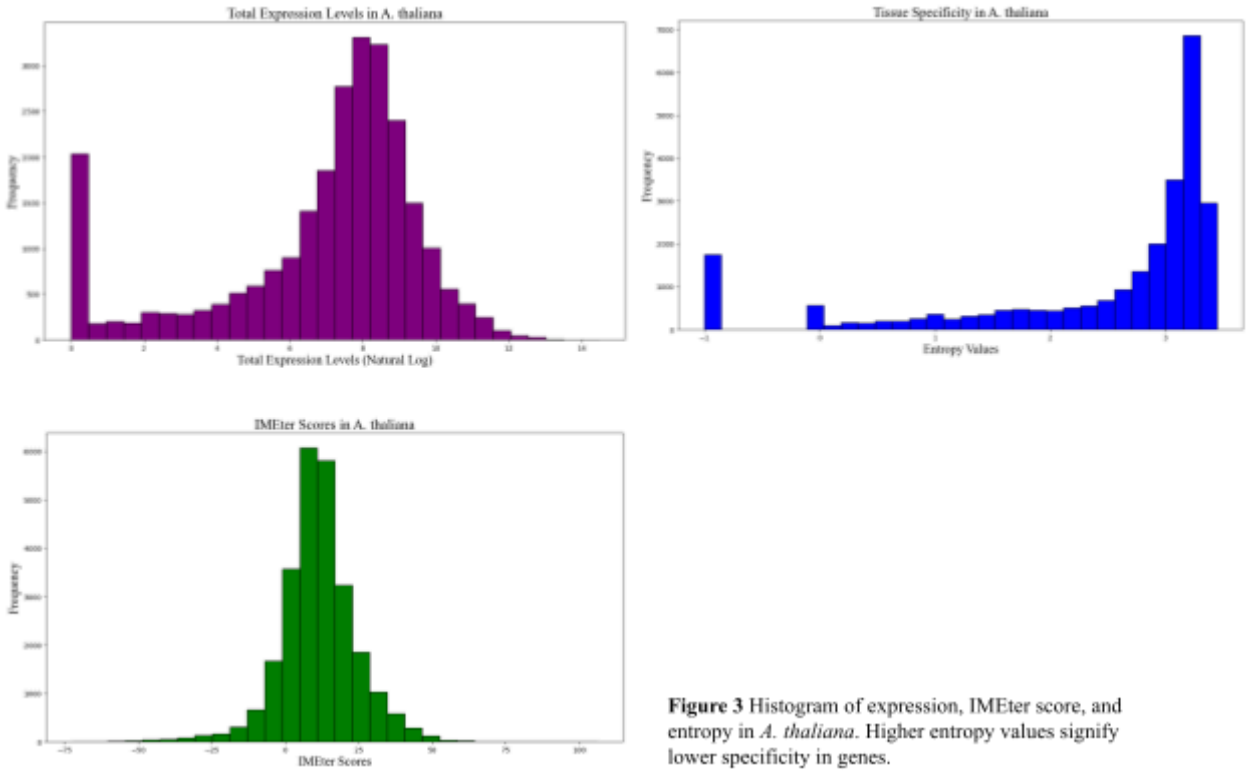


**Figure 3** Histogram of expression, IMEter score, and entropy in *A. thaliana*. Higher entropy values signify lower specificity in genes.

From our analysis, we see that despite there being a dearth of genes with high IMEter scores, the majority of high-scoring genes were concentrated in the higher-entropy regions of the graph and contained higher expression, verifying that our results were not the result of an overwhelming number of broadly expressed genes. Rather, the number of highly-scoring genes with high total expression was low but made up an overwhelming number of broadly expressed genes. The same graph was reproduced for *O. sativa*.

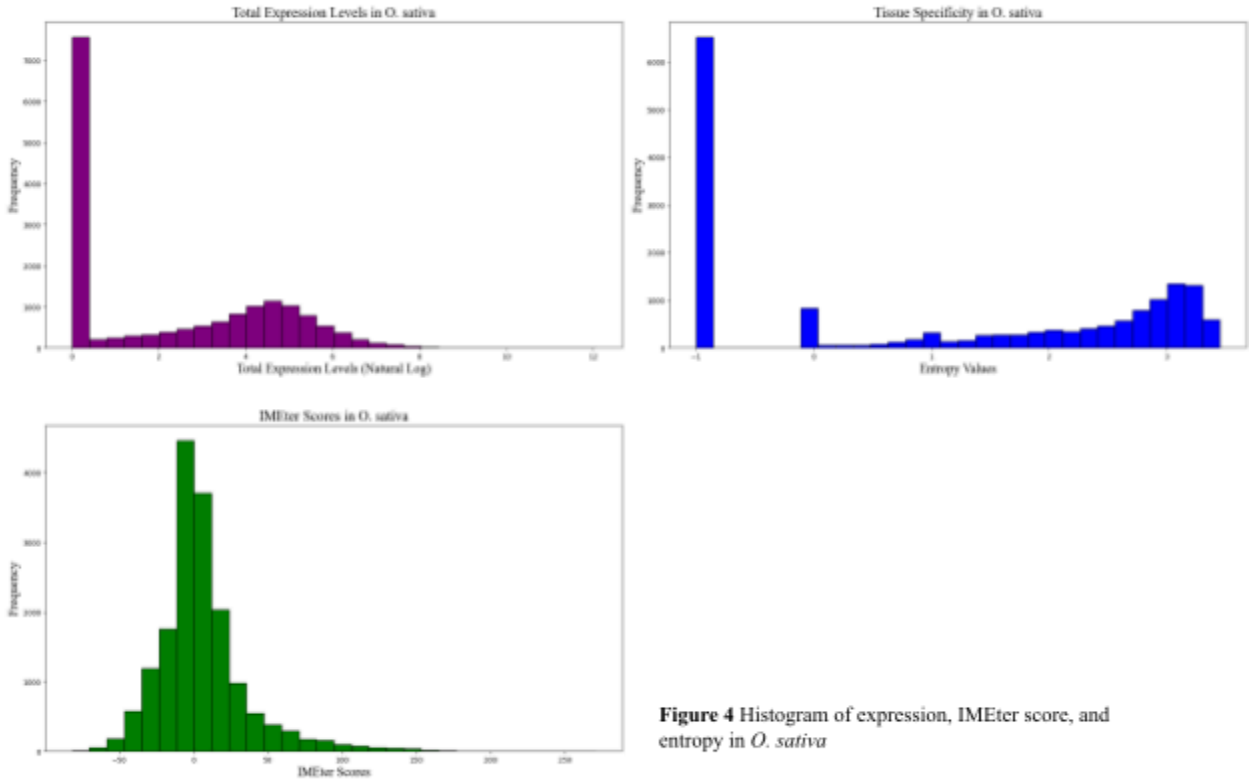


**Figure 4** Histogram of expression, IMEter score, and entropy in *O. sativa*

# Discussion

The use of genes in biotechnological applications requires complete understanding of the inner workings of genes, including the ways in which genes are expressed and the various factors that contribute to the relative levels of gene expressions in various tissues and, indeed, individual cells. To do so, we sought to explore how introns, once considered waste products and redundancies in DNA, play an integral part in controlling the gene expression across the organism.

The underlying basis of this experiment was the question raised by inserting powerful intron sequences with high IME content to create transgenes (Emami et al., 2013). The resulting phenomenon of overexpressed genes that had previously been tissue-specific caused us to wonder whether IME content was specific to broadly expressed genes and anthema to tissue-specific genes, no matter their expression levels. The experiment above confirmed this hypothesis, namely that IME content is nearly absent from genes with tissue-specific expression, regardless of their total expression levels.

Such findings raise further questions about gene expression that are harder to answer. In plants, tissue-specific genes that are highly expressed (examples include the LEC1 gene, which is responsible for embryo development and accumulation of nutrients necessary for seed germination and seedling growth) must have other mechanisms that increase total expression without sacrificing gene specificity. To study such mechanisms, future research should focus on genes with low IMEter scores and therefore introns with low IME content, as IME reduces tissue specificity. Outliers with low IMEter scores and entropy values but express high total expression buck the trend and may be singled out as potential examples of mechanisms in tissue-specific genes.

Experiments may also focus on the other side of the spectrum, mainly outlier genes that express exceptionally high IMEter scores but are tissue-specific. Such genes may be found in the upper-left quadrant of the graph, and although few and far between, do exist. These genes likely contain special conditions that explain its interesting position in the graph.

The experiment was also limited by the IMEter algorithm it employed. Rather than scoring introns exactly based on their IME likelihood, the algorithm was based on the straightforward assumption that promoter-proximal introns were more likely to contain IME content while promoter-distal introns had less of an effect on gene expression. Although various modifications made to the algorithm, such as implementing a threshold cutoff for k-mers rather than entire introns during training, made the data slightly more accurate, not much of a change was observed. Specifically, the IMEter model used in this experiment did not consider searching for common motifs in IME-inducing introns, such as the common "TTNGATYTG," which was found to convert non-stimulating introns into IME introns (Gallegos & Rose, 2019). Editing the IMEter to take motifs into account could make the trend more accurate and further differentiate between genes that follow the observed trend and outliers worth investigating.

# Acknowledgements

The author of this paper would like to thank Dr. Korf for his guidance throughout the project and for providing the inspiration and encouragement for this paper. The author would also like to extend his thanks to the YSP program and to Xochitl for providing guidance and encouragement. Lastly, the author would like to thank his lab partners Merdy and Nicole for their help and encouragement throughout the duration of this paper.

# References

Bartlett, J. G., Snape, J. W., & Harwood, W. A. (2009). Intron-mediated enhancement as a method for increasing transgene expression levels in barley. *Plant Biotechnology Journal*, *7*(9), 856–866. https://doi.org/10.1111/j.1467-7652.2009.00448.x

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis (New York, N.Y.: 2000)*, *53*(8), 474–485. https://doi.org/10.1002/dvg.22877

Emami, S., Arumainayagam, D., Korf, I., & Rose, A. B. (2013). The effects of a stimulating intron on the expression of heterologous genes in Arabidopsis thaliana. *Plant Biotechnology Journal*, *11*(5), 555–563. https://doi.org/10.1111/pbi.12043

Gallegos, J. E., & Rose, A. B. (2015). The enduring mystery of intron-mediated enhancement. *Plant Science*, *237*, 8–15. https://doi.org/10.1016/j.plantsci.2015.04.017

Gallegos, J. E., & Rose, A. B. (2018). *Intron-mediated enhancement is not limited to introns* (p. 269852). bioRxiv. https://doi.org/10.1101/269852

Gallegos, J. E., & Rose, A. B. (2019). An intron-derived motif strongly increases gene expression from transcribed sequences through a splicing independent mechanism in Arabidopsis thaliana. *Scientific Reports*, *9*(1), 13777. https://doi.org/10.1038/s41598-019-50389-5

Long, M., de Souza, S. J., & Gilbert, W. (1995). Evolution of the intron-exon structure of eukaryotic genes. *Current Opinion in Genetics & Development*, *5*(6), 774–778. https://doi.org/10.1016/0959-437x(95)80010-3

Maas, C., Laufs, J., Grant, S., Korfhage, C., & Werr, W. (1991). The combination of a novel

stimulatory element in the first exon of the maize Shrunken-1 gene with the following intron 1 enhances reporter gene expression up to 1000-fold. *Plant Molecular Biology*, *16*(2), 199–207. https://doi.org/10.1007/BF00020552

Maniatis, T., & Tasic, B. (2002). Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, *418*(6894), 236–243. https://doi.org/10.1038/418236a

Rose, A. B. (2002). Requirements for intron-mediated enhancement of gene expression in Arabidopsis. *RNA*, *8*(11), 1444–1453. https://doi.org/10.1017/S1355838202020551

Rose, A. B., Elfersi, T., Parra, G., & Korf, I. (2008). Promoter-Proximal Introns in Arabidopsis thaliana Are Enriched in Dispersed Signals that Elevate Gene Expression. *The Plant Cell*, *20*(3), 543–551. https://doi.org/10.1105/tpc.107.057190

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621