Intron Retention Propensity

Michael Killeen Jones

Genomic Medicine MSc, 2024 – 2025

Supervised by Jernej Ule

Word Count: words

Abstract: words

Abstract - less than 250 words

List of Contents

1.1 Introduction

2.1 Results

3.1 Discussion

4.1 Future Work

5.1 Methods

List of Figures and Tables

Abbreviations (separate page)

Appendix (or simply link Git repo?)

* 1. **Introduction**

Once considered errors of splicing, recent work suggests cell-type and developmental-stage specific intron retention (IR) is a subtle lever for modulating gene expression, implicated in cancer and neurodegenerative disease1. Alternative polyadenylation and “cassette” exon skipping are well known alternative splicing mechanisms for transcriptional regulation. Despite IR being well-documented in the Plant kingdom, its evolutionary divergence in higher mammals has contributed to significant diversity across species and tissue types, especially the brain.

IR contributes to regulating differentiation of granulocytes2 and erythroblasts3, ion channel isoform diversity4 and localisation of transcripts to dendrites in neurons5. IR events can be translated, included in stable non-coding circular RNAs6 or can target a transcript for exosome degradation7, controlled nonsense mediated decay, or nuclear detention8 for tightly regulated gene expression. Beyond transcriptional regulation of individual genes, IR may act more broadly to regulate RNA binding proteins (RBP) networks, for example cytoplasmic intron-retaining transcripts (CIRTs) are noted to possess high binding affinity to mislocalised RBPs in amyotrophic lateral sclerosis (ALS)9, where mistiming of IR has been demonstrated as a molecular hallmark10. Given that IR and readthrough also increase with age11, interrogating splicing regulatory motifs may yield insights regarding aging and neurodegeneration.

Despite the widespread observation of IR as a regulatory mechanism, we understand little about the *cis* and *trans* features contributing to the spatiotemporal changes that underlie retention. The probability of IR is associated with sequence features including GC content, intron length, splice site composition, and presence of cryptic splice sites12. Highly expressed nervous system genes, with lengthy introns spanning up to 100 kilobases13, are highly enriched for IR, which is more evolutionarily conserved in brain tissue14 (Figure 1). These introns are often regulated by multivalent RBP interactions14,15; quantified by a scoring algorithm developed by the Ule lab16. Subnuclear localisation may also impact or be impacted by these interactions 17.

Integrating Ule lab datasets with other studies measuring: sequence multivalency scores, RBP binding sites (CLIP), IR (RNA-Seq), subcellular localisation (TSA-Seq18) and splicing kinetics (metabolic RNA-Seq19) will explore how *cis/trans* elements and binding complexity contribute to IR. Exploring the sequence characteristics of physiologic IR, we aim to holistically evaluate their deleterious effects in neurodegenerative disease using a novel approach.

**Objectives**

Q1. What common sequence features guide IR programmes across diverse cell and tissue types

Q2. Can IR propensity be predicted by known *cis* and *trans* features?

1. Gather and uniformly process high-throughput sequencing datasets that will provide insight into intron retention
2. Produce an IR propensity score for all human introns
3. Quantify weighted value of *cis* and *trans* features associated with IR probability across cell types by building a machine-learning (ML) predictive model

**2.1 Results**

VastDB exploration

* Intron Length vs Max PSI (Hexbin)
* Median PSI by Intron Length Deciles
* Distribution of All PSI Values and Max PSI values
* Events with PSI > 15 across samples
* Total IR Events per sample with PSI > 15
* Heatmap, expression filtered
* PSI vs cRPKM Across All Tissues by Cluster (Hexbin clusters 1-4)

**Introduction**. This should provide the reader with the background material necessary to understand your research project. This should include a brief review of the published literature together with a discussion of the specific work, published and unpublished, that led to your own research project. This is not an essay or a general review, just the specific background to your project. There is no need to spend a lot of time reviewing basic knowledge of Human Genetics or Molecular Biology. The final part of it should introduce the specific question addressed by your research work. You can include separate sub-sections in the final part on hypothesis, aims and research plan as you see fit.

**Materials and Methods.** This section should describe the reagents, cells, etc. that you used and the methods that you carried out. It should include descriptions of the human subjects if needed. It must give sufficient detail for someone to read the method and be able to repeat the experiment. Commercial reagents should have their source identified (i.e. the company and country) in brackets after they are mentioned for the first time, but not on subsequent occasions, e.g. Low melting-point agarose (Biorad, UK). The country should only be given the first time a company is mentioned. On subsequent occasions the name of the company is sufficient, e.g. Agarose, Molecular Biology Grade (Biorad). Remember to include sections for bioinformatics and statistics, where appropriate. You do not need to include detailed methods that you did not carry out yourself, if needed a brief description and suitable reference is appropriate. Please remember that methods were all originally reported for the first time in the literature and so most have associated references. Where possible include references to methods.

**Results.** The way in which you present your results will depend upon the nature of your project but the following general rules apply to all studies. Your data should be concisely described in the text. Data should be presented as Figures (e.g. graphs, gel images) and Tables. Figures and Tables should be numbered sequentially (e.g. Figure 1, Figure 2, etc., Table 1, Table 2, etc.) and should be referred to in the appropriate position in the text. Figures and Tables should be placed as close as possible to the relevant text so that the reader does not have to flick backwards and forwards to read the work. Each Figure or Table should be fully labelled and accompanied by a detailed legend, which is text describing the figure or table so that it can be understood without reading the main text. Avoid presenting the same data in more than one way unless there is a good reason, e.g. basic raw data in a Table and the analysed data as a graph is fine but presenting the raw data in a table and as a graph is not. Statistical analysis should be applied wherever appropriate.

**Discussion.** This section is probably the most important part of the report as it is where you can demonstrate the quality of your scientific thought, your critical analysis skills and your appreciation of the position of your work in the context of the scientific literature. There are two aspects to a discussion of the work you have undertaken: technical and academic. For the technical part you should discuss the advantages and disadvantages of the techniques that you used. You should discuss the problems (there are always some!) that you encountered, why you think these arose and how you tried to solve them. For the academic part you should summarise the major findings of your research and discuss your interpretation of these findings and their significance in the context of the published literature. It is important to be critical in your discussion of both your data and the literature. Critical does NOT just mean negative, but rather that you should point out both the positive and negative aspects of your project results.

Finally, you should discuss future work that could be done to follow on from your project, and the direction in which you think this research might lead. This may be presented as part of the Discussion or as a separate section.

**References.** When you write your thesis you will need to cite previously published work. Wherever possible, every statement should be backed up by a suitable reference, which should be a primary source, ideally an original article, though a secondary source such as a review or, possibly, a book, may be appropriate. In general, it is not a good idea to cite textbooks. The quantity, relevance and formatting of your referencing will be assessed. Please ensure that you use up-to-date references where possible, because fields such as human genetics are very fast-moving. Make sure that all web-based material is of suitable academic quality (e.g. not Wikipedia) and be sure to give the date that you accessed it together with the URL.

**Citations**

Referencing should be in Harvard or Vancouver style and are briefly described below.

For a more detailed guide on referencing see the SGUL Library Resources guide: <https://libguides.sgul.ac.uk/Harvard/citing_referencing>

**Harvard**

Harvard uses an author-date style of referencing which includes the author name and year of the publication when cited in the main text.

Harvard Style Citation

A reference should be cited with a number at the relevant point in the sentence.

      Many students have dreams of winning a Nobel Prize: for less than 0.1% of students these dreams may be realised (Fish, 2022).

Or

Fish (2022) observed that many students have dreams of winning a Nobel Prize: for less than 0.1% of students these dreams may be realised.

If you want to cite several pieces of work in the same sentence, you will need to include the citation for each piece of work. The sources should be cited in alphabetical order and a semi-colon should be used to separate each publication:

Many students have dreams of winning a Nobel Prize: for less than 0.1% of students these dreams may be realised (Aardvark, 2022; Bird, 2020; Cat, 1999).

**Vancouver**

Vancouver uses consecutive numbers for citations in the main text and then orders the reference list on the basis of the order they appear in the main text.

Vancouver Style Citation

A reference should be cited with a number at the relevant point in the sentence.

      Many students have dreams of winning a Nobel Prize: for less than 0.1% of students these dreams may be realised (1).

Or

Many students have dreams of winning a Nobel Prize: for less than 0.1% of students these dreams may be realised1.

If you want to cite several pieces of work in the same sentence, you will need to include the citation number for each piece of work. A hyphen should be used to link numbers which are consecutive, and a comma used where numbers are not consecutive.

The following is an example where works 5,6,7,9,13,14 and 15 have been cited in the same place in the text.

Many students have dreams of winning a Nobel Prize: for less than 0.1% of students these dreams may be realised (5-7,9,13-15).

If you cite a reference which has the same author and was written in the same year as an earlier citation, each work will have a different number.

**Reference List**

All the references cited in your document should be collected together at the end of the report and arranged in alphabetical order for Harvard or in numerical order for Vancouver style. Reference manager applications such as EndNote, Mendeley, Qiqqa, Zotero etc. will help you organise your references and in automatically creating the bibliography. All the details, including all authors, full title, full journal name (in italics), volume number (bold text) and first and last page numbers, should be given as below.

Student, U. and Supervisor, M.E. (2012) “Our path to the Nobel Prize, riches and fame”, *Nature Fantasy* **233**:17-43

When you want to refer to a chapter in a book:

Student, U. and Supervisor, A., (2015). A life in science: the agony and the ecstasy. In: A. Bright and M.I. Breezy (eds.) The Microenvironment of the University System. St George’s Press, London, pp 5-25.

When you want to refer to a whole book:

Old, M.I. and Weary, A. (2020). Sticky ends – where is molecular biology leading us? St George’s Press, London.

When you need to refer to a website, give the URL, the author (where possible) and the date accessed.

General Guidelines

If you are quoting exactly what was said in a paper or book, put the extract in italics and within quotation marks, and give the reference, for example:

“*Copying sentences from papers or other sources without making it clear that you are quoting the author is a form of plagiarism, which is detected by Turnitin even if small changes to the wording are made, and which earns severe penalties*” (Supervisor, 2015).

Please note that is extremely rare to see even one quotation in a scientific report of a lab-based project.

Do not quote a reference that you have not read. A critical assessment of the full paper is needed to decide whether you agree with a paper’s conclusions.

Do not put a reference in the Reference section of your report unless it is mentioned in the text.

**Harvard vs. Vancouver:**

- Advantages of author-style (Harvard):

* 1. You can recognize the referenced work from just its in-text citation.
* 2. The first author is more visibly credited.
* 3. Dates help readers establish the chronology of prior work and quickly identify outdated citations.
* 4. Citations and references can be prepared independently,

-  Advantages of numeric-style (Vancouver):

* 1. Numeric-style is more space efficient.
* 2. Numeric-style is less distracting.
* 3. Numeric lookup of references is easiest.
* 4. Numeric-style citations do not degrade when citing works without first authors.

For further discussion of the different styles see: <https://blog.dhimmel.com/citation-styles/>

**Appendix.** Occasionally, students may wish to include supplementary material that is not essential for the report, and that would disrupt the flow of the main text too much, but which provides the reader with useful information. This can be added in the form of an appendix, or very occasionally appendices. An appendix should appear as the very last part of the report. As a guideline, a maximum of 10 pages should be used for appendices, if more is required then careful consideration should be given to including it in the main text. Examples of appendices are long lists of SNPs, details of PCR primer sequences, gene sequences, examples of raw outputs from assays that are too long to include in the main text, etc.

**Plagiarism.** Please remember that your work will be checked for plagiarism using Turnitin. Evidence of plagiarism will be treated very seriously and is likely to attract heavy penalties, for example, a mark of zero for the submitted document. For more details on this topic, please read the appropriate section in the Course Handbook and the Plagiarism Policy document available in the Course Information section on Canvas

The markers should include in their assessment the following: consideration for overall presentation, use of illustrations, presentation and analysis of results, clarity of argument and structure as well as content, although not all of these factors have equal weighting when deciding the overall mark.

Each section of the submission should be graded on a scale of 100 and the final mark must reflect the balance of these grades, weighted as necessary by the marking scheme.

The overall length (excluding abstract, tables, figure legends and references) should not exceed 10,000 words

**To achieve and maintain consistency in marking, examiners are asked to consider the following points:**

***Abstract (weighting 5%)***

This should summarize in no more than 250 words the research problem, the methods, the core results and their significance – much as in a journal article.

***Introduction & literature (weighting 20%)***

This should introduce the topic and area of the project and should thus include an appropriately detailed review of the literature. Introduction should critically analyse current literature, systematically and comprehensively and be up-to-date, setting the background for the study. Diagrams are encouraged and should be clearly explained. Any ethical issues should be addressed. Any areas of agreement/disagreement and lack of clarity within the literature should be highlighted. This section should lead, naturally and logically, to the Aims and Objectives.

***Aims & objectives (weighting 5%)***

There should be a clear hypothesis. The aims/objectives can be a series of bullet points, presenting the key ideas as a series of questions/statements to be answered, and an explanation of how these will be achieved

***Materials & methods (weighting 10%)***

This section should be presented well enough to allow others to reproduce the protocols (details can be referenced or put in an appendix at the end of the report). For high marks, the text should justify why particular methods were used. Markers should also consider if the methods are appropriate for the objectives and suitable for the study.

***Results (weighting 30%)***

The results should be presented as a logical progression - how one experiment leads to another – accompanied by internally consistent controls. The text provides a commentary of the experimental findings. The data should be presented as properly formatted Figures and Tables and statistically analysed as appropriate with supporting figure legends. The data collection should be reliable, reproducible and appropriate to the type and nature of the data. The conclusions drawn should be justified by the data and summarize the findings accurately and succinctly, showing originality. It is immaterial whether the student is presenting negative or positive results, they should be marked on how they are presented.

***Discussion & Future work (weighting 25%)***

The discussion section should describe, in a critical way, how analysis of the experimental evidence answers the questions laid out in the aims. This may include the presentation of caveats, limitations and problems encountered which prevented a fully satisfying set of results from being achieved. It should include discussion of what went wrong (and why), what could be done next if continuing with this project, and what the future prospects are. The issues of validity, reliability and generalisability can also be discussed. The discussion should relate back to the context (the objectives and hypothesis) in which the work was set. The results should be systematically and critically evaluated in relation to the literature.

***General Presentation (weighting 5%)***

Is the overall presentation clear, logically structured and detailed enough to form a judgement? Marks are awarded for general appearance, quality of images and legends (e.g. proper magnifications, explanations of sample lanes etc.), and well-organised references. Appropriate referencing including accurate citations of original articles and reviews is a pre-requisite for marking. A report without an acceptable standard of references will not be assessed and should be resubmitted.

No specific referencing style is required but whatever style is used must allow the reader to easily find the full reference in the reference list and then find the original paper online or in a library.

Class Mark Description

Distinction 90-100 Faultless and brilliant. The report could be published in a good journal or submitted as a grant application exactly as it stands. Only the very best reports fall into this category!

Distinction 80-89 Excellent in every respect. Clear evidence of critical judgement in selection and discussion of relevant material. In-depth knowledge and understanding of topic, clear presentation and analysis of results, logical structure and argument with evidence of original thought.

Distinction 70-79 The basic criterion for a Distinction is that the work as a whole be of a high professional standard and show originality; it should be readily understood, and enjoyed, by an interested examiner who is not an expert in the topic of the report. Numerical data must be analysed using appropriate statistics. Appropriate referencing including accurate citations of original articles and reviews.

Merit 65-69 Very good. Material is relevant and well-presented and analysed. Good critical grasp and discussion of subject. Minor defects are allowable if compensated by other factors, but referencing should be correct.

Merit 60-64 Good. Less critical grasp than above but still clear and well-structured with sufficient referencing.

Pass 55-59 Reasonably good in most areas. Some satisfactory material but there may be gaps, errors and/or misunderstandings. Referencing should be correctly and consistently formatted.

Pass 50-54 Fair. Some reasonable material but shortcomings in relevance, understanding, structure and/or clarity. The minimum for an MSc.

Near Fail 40-49 Poor. Multiple shortcomings with few good features. Not acceptable at MSc level.

Poor Fail 0-39 Bad. Serious misunderstandings and little evidence of effort.