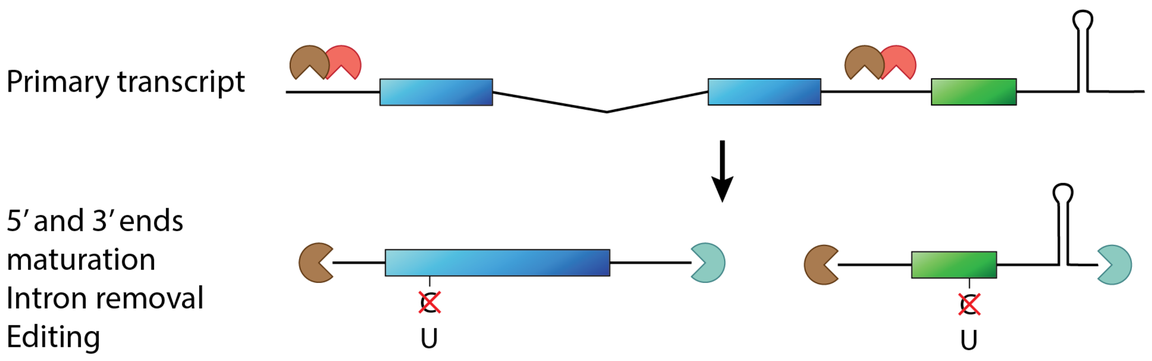
Using contrast to study RNA transcripts co-maturations

Benjamin VACUS[1,2], Arnaud LIEHRMANN[1,2,3], Guillem RIGAILL[1,2,3], Benoit CASTANDET[1,2], and Etienne DELANNOY[1,2]

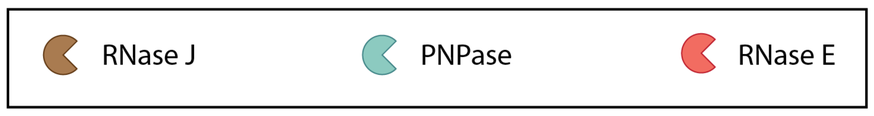
[1]Institute of Plant Sciences Paris-Saclay (IPS2), Université Paris-Saclay, CNRS, INRAE, Université Evry, 91405, Orsay, France
[2]Institute of Plant Sciences Paris-Saclay (IPS2), Université Paris Cité, CNRS, INRAE, 91405, Orsay, France
[3]Laboratoire de Mathématiques et de Modélisation d'Evry (LaMME)

Introduction

Inside plants chloroplasts, RNA transcripts undergo a complex set of maturation events including splicing, editing and processing of their extremities [1]. The Nanopore sequencing technology recently allowed us to study the dependencies between two maturation events - sometimes separated by several thousand bases - on the same transcript in *A. thaliana* [2]. We assessed the dependencies using a Fisher test, thus ignoring several important features of the sequencing data (count, dispersion and replicates variability).



To better model the data we propose to plug a specific GLM model in DeSeq2.



R pipeline

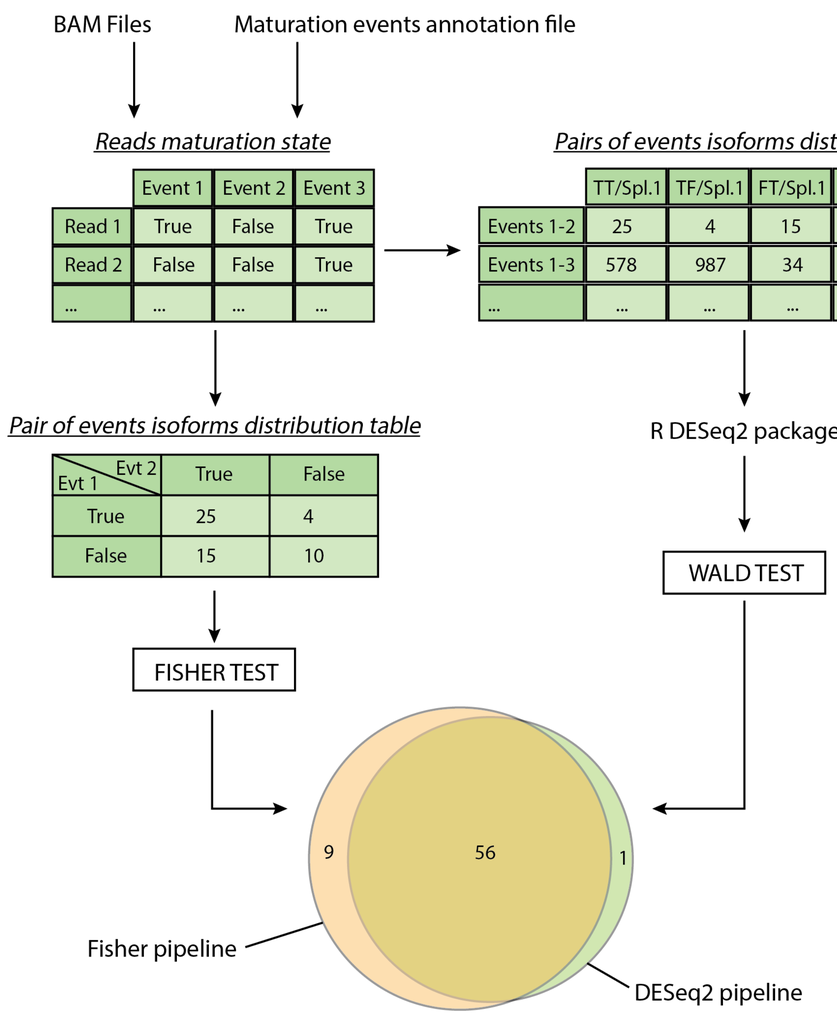
DESeq2 [3] to test dependencies

i. **Annotation** : for every sample, each read, and every maturation we get a state

ii. **Count matrix :** A matrix with one row for each pair of event is built: it contains the counts in each maturation state for every pair of event and every sample.

iii. **DESeq2 :** is used to estimate an interaction effect between the two events.

iv. **Multiple-testing :** The interaction is tested using a Wald test followed by FDR control.



	Event 1	Event 2	Event 3
Read 1	True	False	True
Read 2	False	False	True
...	...	...	...

	TT/Spl.1	TF/Spl.1	FT/Spl.1	FF/Spl.1	TT/Spl.2	...
Events 1-2	25	4	15	10	10	...
Events 1-3	578	987	34	102	102	...
...	...	...	...	...	...	...

Evt 1 \ Evt 2	True	False
True	25	4
False	15	10

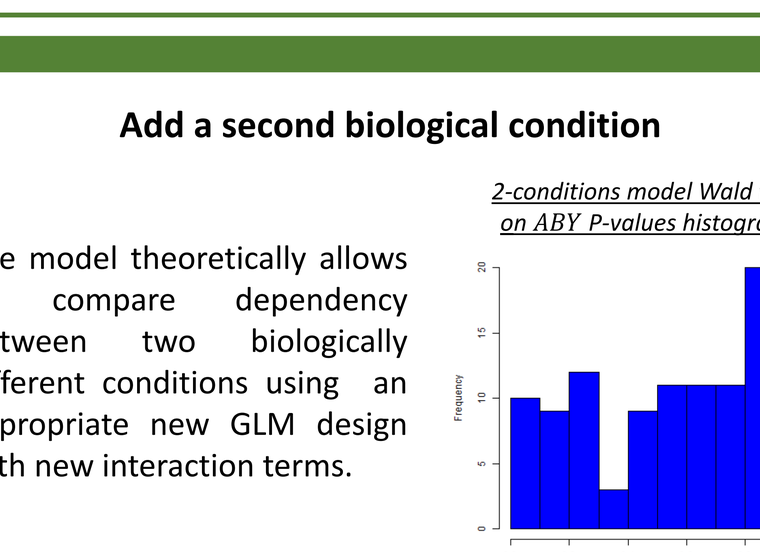
The GLM uses the following design:

$$\log_{10} \boldsymbol{E}(c) = \mu + A + B + AB$$

Baseline
If first site is matured
If second site is matured
If both sites are matured

The non-nullity of the interaction term AB is tested.

DESeq2 pipeline quality checks
Dispersion plot and P-values histogram



Results are coherent with those found using the Fisher test – differences lie in the most ambiguous cases (p-value close to the 5% threshold).

Conclusion

(1) Our pipeline retrieved most dependencies found by the Fisher test. DESeq2 quality checks look good. It is thus our opinion that the dependencies found only by the Fisher test are false positives (because dispersion and replicates variability were not taken into account).

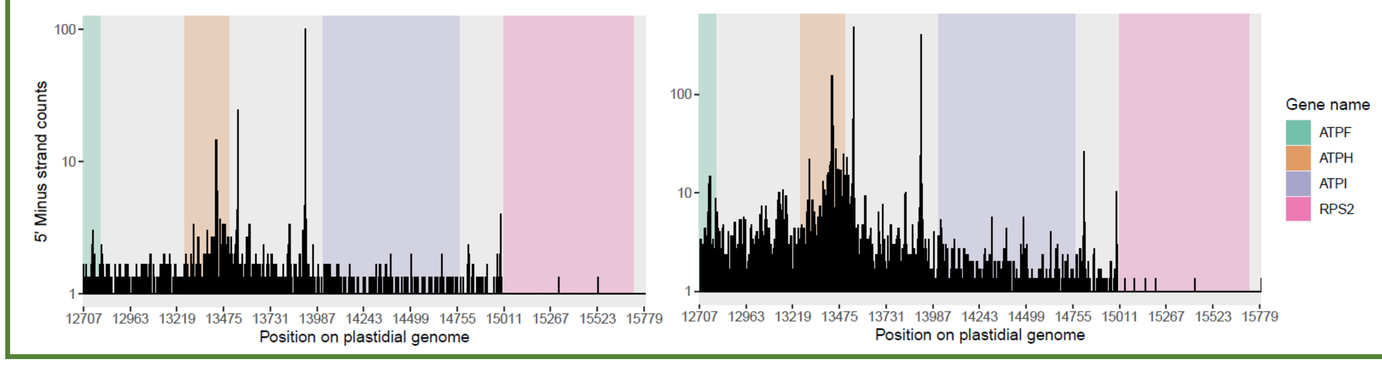
(2) Our pipeline could be applied to any other long-read dataset studying the dependencies between pairs of maturation event.

Further work

Add "processing of extremities" events

Chloroplast RNAs termini are also processed and processing could be linked to editing and splicing. It is however harder to integrate in our pipeline as the observed extremities often define broad peaks making it difficult to establish a clear-cut rule to annotate them.

Distribution of extremities of reads covering one editing site in a matured state (right) or native (left)



Add a second biological condition

The model theoretically allows to compare dependency between two biologically different conditions using an appropriate new GLM design with new interaction terms.

2-conditions model Wald test on ABY P-values histogram

$$\log_{10} \boldsymbol{E}(c) = \mu + A + B + AB + Y + AY + BY + ABY$$

References
[1] Stern, David B., Michel Goldschmidt-Clermont, and Maureen R. Hanson. Chloroplast RNA metabolism. Annual review of plant biology 61: 125-155, 2010.
[2] Guilcher, M., Liehrmann, A., Seyman, C., Blein, T., Rigaill, G., Castandet, B., & Delannoy, E.. Full length transcriptome highlights the coordination of plastid transcript processing. International journal of molecular sciences, 22(20), 11297, 2021.
[3] Love, Michael I., Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome biology 15.12: 1-21, 2014.