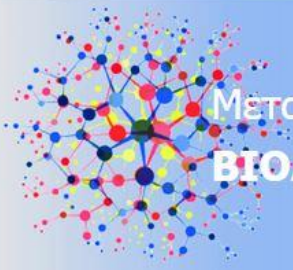




ΓΕΩΠΟΝΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
AGRICULTURAL UNIVERSITY OF ATHENS



Μεταπτυχιακές σπουδές στη
ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ



ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

«Ανάλυση του εναλλακτικού ματίσματος από δεδομένα RNA-seq στο
μετάλλαγμα *lefkothea* του φυτού *Arabidopsis*»

Αθανάσιος Σ. Μπαλτζής

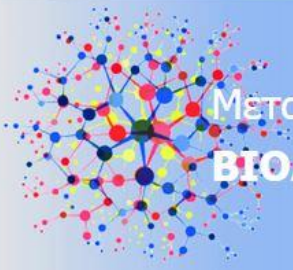
Επιβλέπων: Καθηγητής Πολυδεύκης Χατζόπουλος

Εργαστήριο Μοριακής Βιολογίας

Τμήμα Βιοτεχνολογίας

Γεωπονικό Πανεπιστήμιο Αθηνών

Αθήνα, 2018



Μεταπτυχιακές σπουδές στη ΒΙΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΩΝ



ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

«Ανάλυση του εναλλακτικού ματίσματος από δεδομένα RNA-seq στο μετάλλαγμα *lefkothea* του φυτού *Arabidopsis*»

Αθανάσιος Σ. Μπαλτζής

Επιβλέπων: Καθηγητής Πολυδεύκης Χατζόπουλος

Εργαστήριο Μοριακής Βιολογίας

Τμήμα Βιοτεχνολογίας

Γεωπονικό Πανεπιστήμιο Αθηνών

Αθήνα, 2018

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΑΤΡΙΒΗ

«Ανάλυση του εναλλακτικού ματίσματος από δεδομένα RNA-seq στο
μετάλλαγμα *lefkothea* του φυτού *Arabidopsis*»

Αθανάσιος Σ. Μπαλτζής

Επιβλέπων: Χατζόπουλος Πολυδεύκης, Καθηγητής Μοριακής Βιολογίας

Τριμελής συμβουλευτική και εξεταστική επιτροπή:

Χατζόπουλος Πολυδεύκης, Καθηγητής Μοριακής Βιολογίας (Επιβλέπων)

Μηλιώνη Δήμητρα, Αναπληρώτρια Καθηγήτρια Μοριακής Βιολογίας (Μέλος)

Ρήγας Σταμάτης, Επίκουρος Καθηγητής Μοριακής Βιολογίας (Μέλος)

Περίληψη

«Ανάλυση του εναλλακτικού ματίσματος από δεδομένα RNA-seq στο μετάλλαγμα *lefkothea* του φυτού *Arabidopsis*»

Η LEFKOTHEA είναι μία πρωτεΐνη απαραίτητη για την ομαλή ανάπτυξη του εμβρύου του φυτού *Arabidopsis thaliana*, καθώς τοποθετείται στο χλωροπλάστη και στον πυρήνα συμμετέχοντας σε δύο μηχανισμούς ματίσματος: των ιντρονίων τύπου II των χλωροπλαστικών γονιδίων και των μεταγράφων πυρηνικών γονιδίων που σχετίζονται με χλωροπλαστικές λειτουργίες.

Στην παρούσα εργασία επιχειρήθηκε η αποκάλυψη επιπλέον πληροφοριών για το ρόλο της LEFKOTHEA μέσα από τον *in silico* εντοπισμό γεγονότων εναλλακτικού ματίσματος από δεδομένα αλληλούχισης RNA σε στελέχη που φέρουν τη θνησιγόνο μετάλλαξη *lefko1* και σε στελέχη αγρίου τύπου. Συνολικά χρησιμοποιήθηκαν 5 λογισμικά εντοπισμού γεγονότων εναλλακτικού ματίσματος. Στα δεδομένα που προέκυψαν πραγματοποιήθηκε ανάλυση εμπλουτισμού γονιδίων και σύγκριση των προβλεπόμενων αλληλουχιών στις θέσεις ματίσματος. Τα αποτελέσματα για τα εξαρτώμενα από τη LEFKOTHEA γονίδια οδήγησαν στην εξαγωγή χρήσιμων συμπερασμάτων σχετικά με τις κυτταρικές λειτουργίες που ρυθμίζονται από την πρωτεΐνη.

Επιστημονική περιοχή: Επιστήμες ζωής – Βιολογία - Βιοπληροφορική

Λέξεις – κλειδιά: LEFKOTHEA, εναλλακτικό μάτισμα, αλληλούχιση RNA, *in silico* ανάλυση

Abstract

“Analysis of alternative splicing from RNA-seq data in *lefkothea* mutant of *Arabidopsis* plant”

LEFKOTHEA is an essential protein for embryo development of *Arabidopsis thaliana* plants, as it is localized both in chloroplast and nucleus participating in a splicing mechanism of type II introns in chloroplast encoded genes and in a splicing mechanism of transcripts of nuclear encoded genes related to chloroplast functions.

In this study, we have attempted to reveal additional information about the role of LEFKOTHEA via *in silico* detection of alternative splicing events from RNA-seq data in *lefko1* lethal mutant and in wild-type plants. Overall, 5 alternative splicing events detection programs have been used. Gene enrichment analysis and comparison between the predicted splice site sequences have been carried out in the produced data. The results for LEFKO-dependent genes have provided useful conclusions about cellular functions regulated by LEFKOTHEA.

Science subject: Life sciences – Biology – Bioinformatics

Keywords: LEFKOTHEA, alternative splicing, RNA-seq, *in silico* analysis

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω τον Επιβλέποντα Καθηγητή μου, Πολυδεύκη Χατζόπουλο, καθώς και τον Επίκουρο Καθηγητή Σταμάτη Ρήγα για την καθοδήγηση και τις χρήσιμες συμβουλές τους, οι οποίες αποδείχθηκαν καταλυτικές για τη διεκπεραίωση της παρούσας μεταπτυχιακής διατριβής. Επίσης, ευχαριστώ την Αναπληρώτρια Καθηγήτρια Δήμητρα Μηλιώνη για τη συμμετοχή της στην τριμελή συμβουλευτική και εξεταστική επιτροπή.

Επιπλέον, θα ήθελα να δώσω ιδιαίτερες ευχαριστίες στον μεταδιδακτορικό ερευνητή Γεράσιμο Δάρα για τη διαρκή υποστήριξη και βοήθεια, αλλά και στα υπόλοιπα μέλη του εργαστηρίου Μοριακής Βιολογίας για το ευχάριστο εργασιακό περιβάλλον.

Τέλος, θα ήθελα να αναφερθώ στην οικογένειά μου, η οποία στάθηκε για άλλη μία φορά αρωγός καθ' όλη τη διάρκεια του μεταπτυχιακού προγράμματος σπουδών στηρίζοντας έμπρακτα αυτή μου την προσπάθεια.

Περιεχόμενα

Περίληψη	3
Abstract	4
Ευχαριστίες.....	5
1. Εισαγωγή.....	7
1.1.1 Η διαδικασία του ματίσματος γονιδίων του πυρήνα και άλλων οργανιδίων..	8
1.1.2 Εναλλακτικό μάτισμα και τύποι εναλλακτικού ματίσματος.....	10
1.2 Πρωτεΐνες πολλαπλής τοποθέτησης μέσα στο κύτταρο.....	13
1.3.1 Η LEFKOTHEA ανήκει στην οικογένεια πρωτεϊνών PORR.....	13
1.3.2 Δομική ανάλυση της LEFKOTHEA.....	16
1.4 Σκοπός της διατριβής.....	19
2. Μέθοδοι ανάλυσης.....	20
2.1 Προετοιμασία δειγμάτων και αλληλούχηση RNA	21
2.2 Ανάλυση πρωτογενών δεδομένων αλληλούχησης RNA	21
2.3 Προγράμματα εντοπισμού γεγονότων εναλλακτικού ματίσματος	23
2.3.1 Splicing Adder (SplAdder)	25
2.3.2 replicate Multivariate Analysis of Transcript Splicing (rMATS).....	26
2.3.3 Junction Based Analysis of Splicing Events (JuncBASE).....	27
2.3.4 ASprofile.....	29
2.3.5 Alternative Splicing Analysis Tool Package (ASATP)	30
2.4 Ανάλυση εμπλουτισμού ομάδων γονιδίων.....	32
2.5 Έλεγχος συμφωνίας των αλληλουχιών στις θέσεις εναλλακτικού ματίσματος..	33
3. Αποτελέσματα	34
3.1 Σύγκριση των γεγονότων εναλλακτικού ματίσματος από τα 5 προγράμματα ...	35
3.2 Κοινά γεγονότα και για τους 4 τύπους εναλλακτικού ματίσματος	37
3.3 Κοινά γεγονότα παράλειψης εξωνίου	39
3.4 Κοινά γεγονότα εναλλακτικής 3' θέσης ματίσματος.....	40
3.5 Κοινά γεγονότα εναλλακτικής 5' θέσης ματίσματος.....	42
3.6 Κοινά γεγονότα διατήρησης ιντρονίου	43
4. Συμπεράσματα και Συζήτηση	47
5. Βιβλιογραφία	52
6. Παράρτημα 1.....	56
7. Παράρτημα 2	63
8. Παράρτημα 3	71

1. Εισαγωγή

1.1.1 Η διαδικασία του ματίσματος γονιδίων του πυρήνα και άλλων οργανιδίων

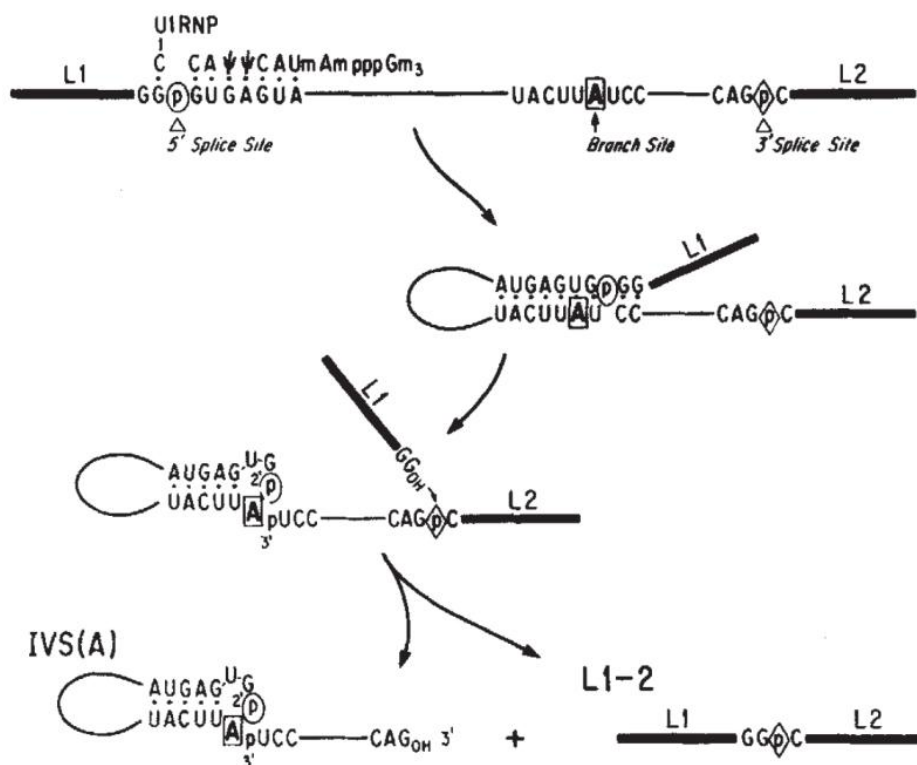
Η ωρίμανση του mRNA (μάτισμα) αποτελεί ένα από τα βασικά στάδια επεξεργασίας που απαιτούνται για την έκφραση των γονιδίων και την παραγωγή πρωτεϊνών και άλλων ειδών RNA στα ευκαρυωτικά συστήματα. Κατά τη διαδικασία αυτή, πραγματοποιείται στα πρώιμα μετάγραφα αφαίρεση των ιντρονίων και συρραφή των εξωνίων με αποτέλεσμα την δημιουργία ώριμων mRNAs, τα οποία μπορούν να επιτελέσουν το βιολογικό τους ρόλο.

Το μάτισμα του RNA ανακαλύφθηκε για πρώτη φορά τη δεκαετία του 1970 σε αλληλουχίες RNA αδενοϊών που μολύνουν κύτταρα θηλαστικών (Berget *et al.*, 1977; Chow *et al.*, 1977), καθώς και σε άλλα ευκαρυωτικά γονίδια, όπως της αιμοσφαιρίνης και των ανοσοσφαιρινών (Darnell, 1978).

Σε μοριακό επίπεδο, ο μηχανισμός ματίσματος στα γονίδια του πυρήνα έχει μελετηθεί σε ένα πλήθος βιολογικών συστημάτων και περιλαμβάνει την αποκοπή των ιντρονίων από τα πρώιμα μετάγραφα σε καλά συντηρημένες αλληλουχίες που ονομάζονται αλληλουχίες ματίσματος (splice sites). Οι θέσεις αυτές βρίσκονται στα 5' και 3' άκρα του εκάστοτε ιντρονίου και περιέχουν, κατά κύριο λόγο, τα δινουκλεοτίδια GU και AG αντίστοιχα (κανόνας GT-AG). Βέβαια, έχει αποδειχθεί ότι σε ένα πολύ μικρό ποσοστό ιντρονίων οι αλληλουχίες στις θέσεις ματίσματος απαρτίζονται από τα δινουκλεοτίδια AU στην αρχή και AC στο τέλος. Επίσης, σημαντικό ρόλο κατέχει και το λεγόμενο σημείο διακλάδωσης, μία αλληλουχία που εντοπίζεται 18 με 40 νουκλεοτίδια ανοδικά του 3' άκρου του ιντρονίου και περιέχει σχεδόν πάντα αδενίνη (A) στην προτελευταία θέση.

Το μάτισμα συμβαίνει σε πολλά στάδια και απαιτεί τη δράση μικρών πυρηνικών ριβονουκλεοπρωτεϊνών (snRNPs) (Εικόνα 1.1) (Clancy, 2008). Σε πρώτο στάδιο, το πρώιμο mRNA αποκόπτεται στο 5' άκρο του ιντρονίου και η ριβονουκλεοπρωτεΐνη U1 προσδένεται στην συμπληρωματική της αλληλουχία στο ιντρόνιο. Έπειτα, ακολουθεί η σύνδεση του κομμένου 5' άκρου στο σημείο διακλάδωσης με την ένωση μιας γουανίνης (G) με την συντηρημένη αδενίνη (A) του σημείου διακλάδωσης, με αποτέλεσμα το σχηματισμό μιας θηλιάς. Η ένωση αυτή πραγματοποιείται με μία αντίδραση μετεστεροποίησης, κατά την οποία μία ομάδα υδροξυλίου ενός ατόμου άνθρακα της αδενίνης προσβάλλει το δεσμό της γουανίνης στη θέση ματίσματος, με συνέπεια την αποκοπή της από την αλυσίδα του RNA και την πρόσδεση μέσω συμπληρωματικότητας με την αδενίνη. Στο επόμενο στάδιο, οι

ριβονουκλεοπρωτεΐνες U2 και U4/U6 ισχυροποιούν την επαφή μεταξύ του 5' άκρου και του σημείου διακλάδωσης, ενώ η ριβονουκλεοπρωτεΐνη U5 καταλύει την αποκοπή του 3' άκρου του ιντρονίου και την ένωσή του με το 5' άκρο με μία αντίδραση μετεστεροποίησης. Τέλος, τα παρακείμενα εξώνια συρράπτονται μεταξύ τους και η θηλιά απελευθερώνεται έχοντας προσδεμένες τις U2, U5 και U6.



Εικόνα 1.1: Διαγραμματική απεικόνιση της διαδικασίας του ματίσματος.

Όλη η διαδικασία του ματίσματος συμβαίνει σε ένα μεγάλο πρωτεϊνικό σύμπλοκο που ονομάζεται σωματίο ματίσματος (spliceosome) και αποτελείται από τον συνδυασμό των snRNPs μαζί με άλλες πρωτεΐνες. Μέχρι σήμερα έχουν ανακαλυφθεί δύο τύποι σωματίων ματίσματος: το μείζον σωματίο ματίσματος (U2 spliceosome ή major spliceosome), το οποίο περιγράφηκε παραπάνω και καταλύει την αποκοπή της πλειονότητας των πυρηνικών ιντρονίων (U2-type introns) και το πιο σπάνιο U12 σωματίο ματίσματος (minor spliceosome) που σχετίζεται με το μάτισμα ενός μικρού ποσοστού ιντρονίων (U12-type introns) και βασίζεται στη ριβονουκλεοπρωτεΐνη U12 (Patel *et al.*, 2003).

Τα μιτοχονδριακά και χλωροπλαστικά γονιδιώματα περιέχουν, επίσης, ένα πλήθος ιντρονίων, τα οποία διαχωρίζονται στις ομάδες I, II και III με βάση συντηρημένα

χαρακτηριστικά της πρωτοταγούς δομής και την προβλεπόμενη δευτεροταγή τους δομή (Schmitz and Barkan, 2007). Στα μιτοχόνδρια και τους χλωροπλάστες των περισσότερων φυτών συναντώνται, κυρίως ιντρόνια της ομάδας II καθώς και ένα μικρό ποσοστό της ομάδας I. Για παράδειγμα, το χλωροπλαστικό γονιδίωμα του *Arabidopsis thaliana* περιέχει 17 ιντρόνια της ομάδας II και 1 ιντρόνιο της ομάδας I. Αντίθετα, τα ιντρόνια της ομάδας III είναι ιδιαίτερα σπάνια και έχουν βρεθεί μόνο στα Ευγληνοειδή (Christofer and Hallick, 1989).

Όπως έχει αποδειχθεί, το μάτισμα των χλωροπλαστικών ιντρονίων της ομάδας II επιτελείται με τη δράση πολλών πρωτεϊνών που κωδικοποιούνται από τον πυρήνα (de Longevialle *et al.*, 2010). Στα αγγειόσπερμα το γεγονός αυτό επιβεβαιώνεται και από την απουσία φαινομένων αυτο-ματίσματος. Ωστόσο, οι πρωτεΐνες αυτές φαίνεται ότι δεν εμφανίζουν ομοιότητες με τους αντίστοιχους πυρηνικούς παράγοντες ματίσματος παρά την προφανή εξελικτική σχέση μεταξύ του ματίσματος των ιντρονίων της ομάδας II και του ματίσματος των πυρηνικών ιντρονίων (Valadkhan, 2007). Η πλειονότητά των χλωροπλαστικών παραγόντων ματίσματος ανήκει στις οικογένειες πρωτεϊνών που διαθέτουν συγκεκριμένες επικράτειες πρόσδεσης RNA (CRM, PPR, PORR και APO) και τοποθετούνται πρωταρχικά στα μιτοχόνδρια ή τους χλωροπλάστες.

1.1.2 Εναλλακτικό μάτισμα και τύποι εναλλακτικού ματίσματος

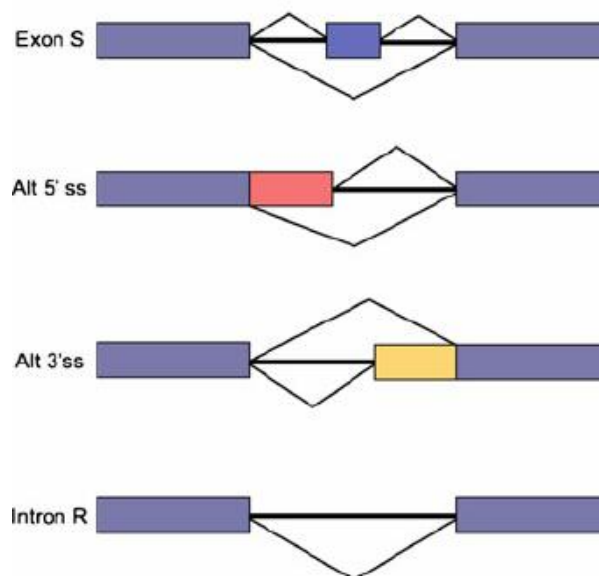
Όπως αναφέρθηκε στην προηγούμενη ενότητα, η διαδικασία του ματίσματος οδηγεί στην παραγωγή ενός ώριμου mRNA από ένα γονίδιο με την αποκοπή των ιντρονίων και τη συρραφή των εξωνίων. Ωστόσο, στους ανώτερους ευκαρυωτικούς οργανισμούς παρατηρείται το φαινόμενο από ένα γονίδιο να προκύπτουν δύο ή/και περισσότερα ώριμα μετάγραφα, το οποίο ονομάζεται εναλλακτικό ή διαφορικό μάτισμα. Το εναλλακτικό μάτισμα αποτελεί έναν σημαντικό μηχανισμό ρύθμισης της γονιδιακής έκφρασης που προσδίδει ποικιλία στις πρωτεΐνες που συντίθεται στο κύτταρο και ευθύνεται σε μεγάλο βαθμό για την πολυπλοκότητα των ανώτερων οργανισμών παρά τον μικρό αριθμό γονιδίων που διαθέτουν.

Η πρώτη αναφορά στο εναλλακτικό μάτισμα χρονολογείται το 1977 όταν οι επιστήμονες απέδειξαν με πειράματα σε αδενοϊούς πως ένα πρώιμο μόριο mRNA μπορεί να υποστεί μάτισμα σε διαφορετικές θέσεις παράγοντας ποικίλα mRNAs,

καθένα από τα οποία περιέχει διαφορετικό συνδυασμό εξωνίων (Berget *et al.*, 1977). Ακολούθησε ο εντοπισμός γεγονότων εναλλακτικού ματίσματος και σε άλλα κυτταρικά γονίδια, όπως το γονίδιο που κωδικοποιεί την ανοσοσφαιρίνη *IgM* σε ποντίκια (Early *et al.*, 1980).

Αναφορικά με τους διαφορετικούς τύπους εναλλακτικού ματίσματος, εμφανίζονται τέσσερις κύριες κατηγορίες (Εικόνα 1.2) (Keren *et al.*, 2010):

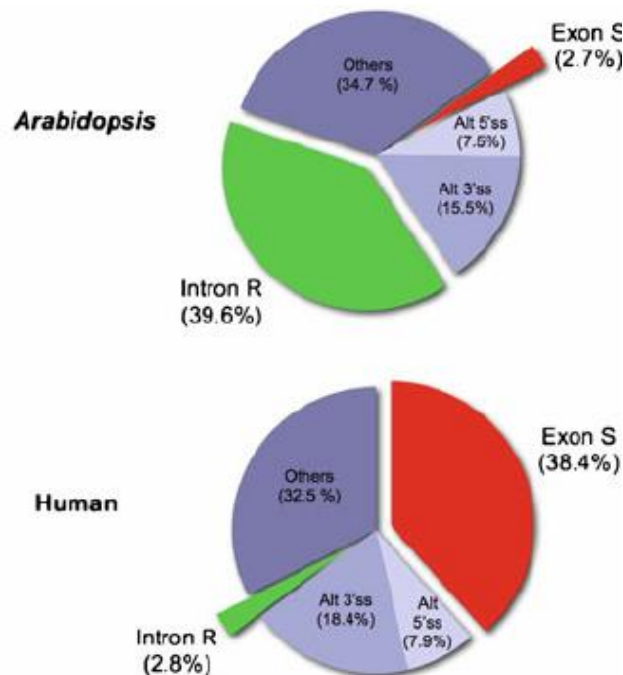
- Η παράλειψη εξωνίου (exon skipping), όπου ένα εξώνιο αποκόπτεται από το εναλλακτικό μετάγραφο
- Η διατήρηση ιντρονίου (intron retention), όπου ένα ιντρόνιο συμπεριλαμβάνεται στο εναλλακτικό mRNA
- Η εναλλακτική 5' θέση ματίσματος (alternative donor site), κατά την οποία αλλάζει το 3' όριο του ανοδικού εξωνίου
- Η εναλλακτική 3' θέση ματίσματος (alternative acceptor site), κατά την οποία αλλάζει το 5' όριο του καθοδικού εξωνίου



Εικόνα 1.2: Απεικόνιση των τεσσάρων κύριων τύπων εναλλακτικού ματίσματος. Με μπλε χρώμα εμφανίζονται τα εξώνια, με μαύρες γραμμές τα ιντρόνια και με κόκκινο και κίτρινο χρώμα οι περιοχές που υφίστανται εναλλακτικό μάτισμα (Carvalho *et al.*, 2013).

Παρολ' αυτά, αξίζει να σημειωθεί η ύπαρξη πιο σύνθετων αλλά λιγότερο συχνών γεγονότων εναλλακτικού ματίσματος, όπως ο αμοιβαίος αποκλεισμός εξωνίων (mutually exclusive exons), η εναλλακτική θέση εκκινήτη (alternative promoter) και η εναλλακτική θέση πολυαδενυλίωσης (alternative poly(A)).

Η συχνότητα εμφάνισης των τύπων εναλλακτικού ματίσματος διαφέρει ανάμεσα στους ευκαρυωτικούς οργανισμούς. Στα σπονδυλωτά ο επικρατέστερος τύπος είναι η παράλειψη εξωνίου, ενώ ο σπανιότερος η διατήρηση ιντρονίου (Kim et al., 2007). Αντίθετα, τα γονιδιώματα των φυτών και των κατώτερων ευκαρυωτών εμφανίζουν γεγονότα διατήρησης ιντρονίου με μεγαλύτερη συχνότητα, όπως φαίνεται και στην σύγκριση των τύπων εναλλακτικού ματίσματος μεταξύ του ανθρώπου και του *Arabidopsis* που παρουσιάζεται στην Εικόνα 1.3.



Εικόνα 1.3: Κατανομή τύπων εναλλακτικού ματίσματος στο γονιδίωμα του *Arabidopsis* και του ανθρώπου (Carvalho et al., 2013).

1.2 Πρωτεΐνες πολλαπλής τοποθέτησης μέσα στο κύτταρο

Ο διπλασιασμός κυτταρικών λειτουργιών λόγω της εξέλιξης είναι ένα φαινόμενο που συναντάται συχνά στα ευκαρυωτικά κύτταρα και έχει ως συνέπεια διαφορετικές ισομορφές πρωτεϊνών να δρουν σε διαφορετικά υποκυτταρικά διαμερίσματα πραγματοποιώντας παρόμοιες λειτουργίες (Krause and Krupinska, 2009). Η συνθετάση του tRNA της ιστιδίνης αποτέλεσε ένα από τα πρώτα παραδείγματα πρωτεΐνης διπλής τοποθέτησης (Natsoulis *et al.*, 1986) και από τότε έως σήμερα η ανακάλυψη αντιστοιχών πρωτεϊνών συνεχίζεται διαρκώς παρέχοντας χρήσιμες πληροφορίες για το ρόλο τους στην επικοινωνία μεταξύ των υποκυτταρικών διαμερισμάτων.

Δύο βασικοί μηχανισμοί φαίνεται να εμπλέκονται στη διπλή τοποθέτηση των πρωτεϊνών στα μιτοχόνδρια και τους χλωροπλάστες (Peeters and Small, 2001; Mackenzie, 2005): ο πρώτος (twin presequences) βασίζεται στην ύπαρξη δύο διαδοχικών πεπτιδίων συνθήματος, καθένα από τα οποία αναγνωρίζεται εκλεκτικά από το σύστημα υποδοχέων του οργανιδίου-στόχου. Απαραίτητη προϋπόθεση είναι η κωδικοποίηση δύο ανεξάρτητων ισομορφών της πρωτεΐνης, ώστε η μία να φέρει στο αμινοτελικό της άκρο το πρώτο πεπτιδίο συνθήματος και η άλλη ισομορφή το δεύτερο. Αναφορικά με το δεύτερο μηχανισμό (ambiguous presequence), η ύπαρξη ενός μοναδικού πεπτιδίου συνθήματος στο αμινοτελικό άκρο της πρωτεΐνης, το οποίο αναγνωρίζεται από τους υποδοχείς εισόδου και των δύο οργανιδίων, εξασφαλίζει την ταυτόχρονη είσοδο της πρωτεΐνης τόσο στα μιτοχόνδρια όσο και στους χλωροπλάστες.

Ωστόσο, παραδείγματα διπλής τοποθέτησης πρωτεϊνών στον πυρήνα και τον χλωροπλάστη καταδεικνύουν έναν διαφορετικό μηχανισμό, ο οποίος επιτυγχάνεται με την ύπαρξη ενός πεπτιδίου συνθήματος στο αμινοτελικό άκρο για την είσοδο στο χλωροπλάστη, αλλά και ενός σινιάλου αναγνώρισης για το σύστημα εισόδου στον πυρηνικό φάκελο (NLS) σε διαφορετική θέση στην αμινοξική ακολουθία.

1.3.1 Η LEFKOTHEA ανήκει στην οικογένεια πρωτεϊνών PORR

Η επικράτεια αναγνώρισης RNA φυτικών οργανιδίων (PORR), γνωστή και ως Domain of Unknown Function 860 (DUF860), αποτελεί ένα κοινό χαρακτηριστικό μιας οικογένειας πρωτεϊνών που στοχεύουν στον χλωροπλάστη ή το μιτοχόνδριο με

παράλληλη τοποθέτηση στον πυρήνα με βάση βιοπληροφορικές αναλύσεις και συμμετέχουν στο μάτισμα των ιντρονίων της ομάδας II των χλωροπλαστικών γονιδίων.

Η πρώτη καλά χαρακτηρισμένη πρωτεΐνη της οικογένειας PORR είναι η WTF1 από τον οργανισμό *Zea mays*, η οποία καταλύει το μάτισμα περίπου των μισών ιντρονίων των χλωροπλαστικών γονιδίων και σταθεροποιεί τα ριβο-πρωτεϊνικά σύμπλοκα με τα ιντρόνια (Kroeger, 2009). Ένα δεύτερο παράδειγμα PORR πρωτεΐνης είναι η WTF9 στο *Arabidopsis*. Υπολογιστικές προβλέψεις καθώς και αποτελέσματα πειραμάτων απέδειξαν την τοποθέτησή της WTF9 στα μιτοχόνδρια, όπου απαιτείται για το μάτισμα των μιτοχονδριακών ιντρονίων *rpl2* και *ccmF_c* (Frans-Small, 2012).

Συνολικά από το γονιδίωμα του *Arabidopsis thaliana* κωδικοποιούνται 15 PORR πρωτεΐνες (Kroeger, 2009), μια εκ των οποίων είναι και η LEFKOTHEA. Η πρωτεΐνη αυτή αναγνωρίστηκε από την εύρεση μιας μετάλλαξης (*lefko2*) στο γονίδιο AT5G62990 που οδηγεί στην εμφάνιση λευκών κοτυληδόνων (Εικόνα 1.4). Με βάση τον φαινότυπο αυτό έγινε και η επιλογή του ονόματος “Λευκοθέα”, η οποία σύμφωνα με τη μυθολογία είναι μία από τις λευκές νηρηίδες της θάλασσας.

Το μετάλλαγμα *lefko2* προέρχεται από χημική μεταλλαξογένεση με EMS στο γονίδιο AT5G62990, η οποία προκαλεί αλλαγή της γουανίνης σε αδενίνη στη θέση 1188bp από το κωδικόνιο έναρξης της μετάφρασης, με συνέπεια την υποκατάσταση του καταλοίπου γλυκίνης στη θέση 373 σε ασπαρτικό οξύ (G→D).

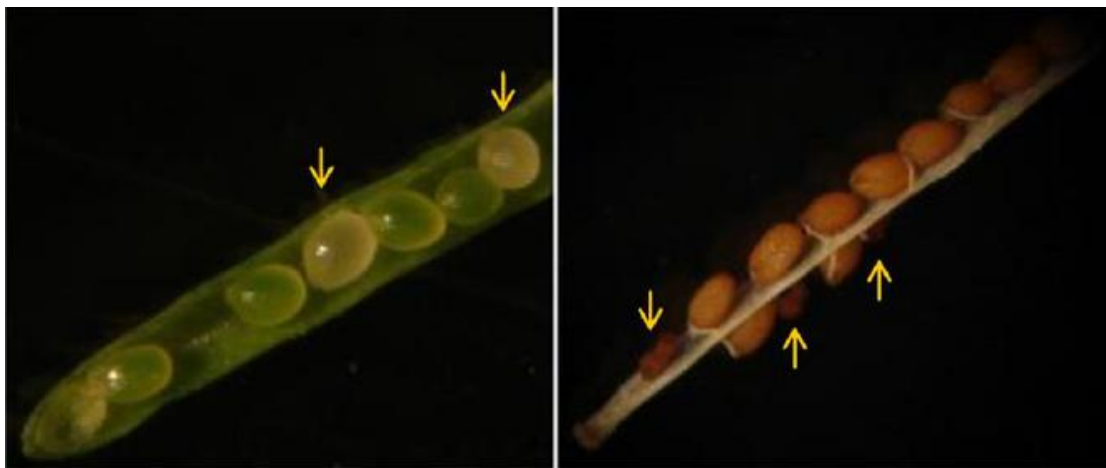
Αντίθετα με τις κοτυληδόνες, τα φύλλα του μεταλλάγματος *lefko2* είναι πράσινα, γεγονός που καταδεικνύει την μη φυσιολογική ανάπτυξη του στα πρώτα στάδια μετά τη βλαστικότητα των σπερμάτων. Ο σημαντικός ρόλος της LEFKOTHEA στην μορφογένεση των ιστών με εμβρυακές καταβολές (κοτυληδόνες), άρα και στην ανάπτυξη του εμβρύου, αποκαλύφθηκε και από την παρατήρηση μορφολογικών αλλοιώσεων των χλωροπλαστών, όπως ο μη φακοειδής σχηματισμός και η οργανωμένη δομή θυλακοειδών, σε κύτταρα κοτυληδόνων φυτών *lefko2*.



Εικόνα 1.4: Φαινότυπος φυτών *lefko2* σε σύγκριση με στελέχη *Col-o* (αγρίου τύπου). Α. Σπορόφυτα (*Col-o* αριστερά και *lefko2* δεξιά) Β. Κοτυληδόνες φυτών 5 ημερών (*Col-o* πάνω και *lefko2* κάτω).

Μετά από βιοπληροφορική αναζήτηση βρέθηκε ακόμη ένα μετάλλαγμα (*lefko1*) στο γονίδιο που κωδικοποιεί την πρωτεΐνη LEFKOTHEA. Πιο αναλυτικά, η μετάλλαξη *lefko1* προκαλείται από T-DNA ένθεση 30bp ανοδικά του κωδικονίου έναρξης στο γονίδιο AT5G62990 και είναι θνησιγόνος σε εμβρυική ηλικία. Το στέλεχος *lefko1* είχε ήδη εντοπιστεί και ονομασθεί emb1692 στα πλαίσια έρευνας για μεταλλάγματα που επηρεάζουν την εμβρυογένεση στο *Arabidopsis* (Tzafrir, 2004). Παρατήρηση των καρποταξιών του *lefko1* επιβεβαίωσε τη διάσχιση του θνησιγόνου φαινοτύπου σε ομοζυγωτικά για το μεταλλαγμένο αλληλόμορφο φυτά (Εικόνα 1.5).

Επίσης, πειράματα με τη μέθοδο της ημιποσοτικής RT-PCR σε ολικό RNA που απομονώθηκε από ομοζυγωτικά *lefko1* θνησιγενή έμβρυα και από έμβρυα αγρίου τύπου *Col-o* αποκάλυψαν την παρουσία μη ώριμων μεταγράφων πυρηνικών γονιδίων που τοποθετούνται στο χλωροπλάστη, όπως τα AT1G03160 και AT1G67660, στα μεταλλαγμένα στελέχη. Συνεπώς, η LEFKOTHEA εμπλέκεται τόσο στο μάτισμα των ιντρονίων της ομάδας II στον χλωροπλάστη μέσω του τμήματος PORR όσο και στο μάτισμα ορισμένων γονιδίων του πυρήνα που σχετίζονται με χλωροπλαστικές λειτουργίες (Δάρας, 2011).



Εικόνα 1.5: Καρποταξίες ετερόζυγων ως προς την ένθεση φυτών *lefko1*. Με βέλη σημειώνονται τα ομοζυγώτα σπέρματα σε μη ώριμη (αριστερά) και ώριμη (δεξιά) καρποταξία.

1.3.2 Δομική ανάλυση της LEFKOTHEA

Κάνοντας χρήση βιοπληροφορικών μέσων και αντλώντας πληροφορίες από την υπάρχουσα βιβλιογραφία, προσδιορίστηκαν τα δομικά χαρακτηριστικά της πρωτεΐνης LEFKOTHEA από την αμινοξική της ακολουθία (Εικόνα 1.6). Οι πρωτεϊνικές επικράτειες, από τις οποίες απαρτίζεται, είναι τρεις: η επικράτεια του αμινο-τελικού άκρου μέχρι το κατάλοιπο 60, η επικράτεια PORR μεγέθους 319 αμινοξέων (61-380) καθώς και η επικράτεια του καρβοξυ-τελικού άκρου με μέγεθος 114 αμινοξέα (381-494).



Εικόνα 1.6: Οι επικράτειες της πρωτεΐνης LEFKOTHEA.

Σύμφωνα με τα αποτελέσματα του προγράμματος ChloPro, τα 69 πρώτα κατάλοιπα αποτελούν το πεπτιδίο-σύνθημα για την αναγνώριση της πρωτεΐνης από τους χλωροπλάστες που αποκόπτεται κατά την είσοδο. Περαιτέρω βιοπληροφορικές αναλύσεις με χρήση νευρωνικών δικτύων και βάσεων δεδομένων προβλέπουν την ύπαρξη ακολουθιών που σχετίζονται με τη στόχευση της LEFKOTHEA στον πυρήνα (Πίνακας 1.1). Πιο συγκεκριμένα, τα προγράμματα NucPred, PredictNLS και NetNES ανίχνευσαν την παρουσία αλληλουχιών-σινιάλων εισόδου (NLS) και αλληλουχιών-σινιάλων εξόδου από τον πυρήνα (NES) (Εικόνα 1.7).

Πίνακας 1.1: Αποτελέσματα βιοπληροφορικών αναλύσεων με νευρωνικά δίκτυα και βάσεις δεδομένων για την υποκυτταρική τοποθέτηση της LEFKOTHEA.

At5g26860	Nucleus	Chloroplast	Prediction Program
484 a.a.	3	7	PSORT
	N.D.	0.49	PREDOTAR
	N.D.	0.687	TARGETP
	N.D.	0.553	CHLOROP
	N.D.	0.576	SLP-Local
	N.D.	1.0	BACELLO
	5.0	N.D.	SUBLOC
	0.925	N.D.	CELLO
	0.74	N.D.	NUCLEO
	1.0	N.D.	NUC-PLOC
	0.69	N.D.	YLOC

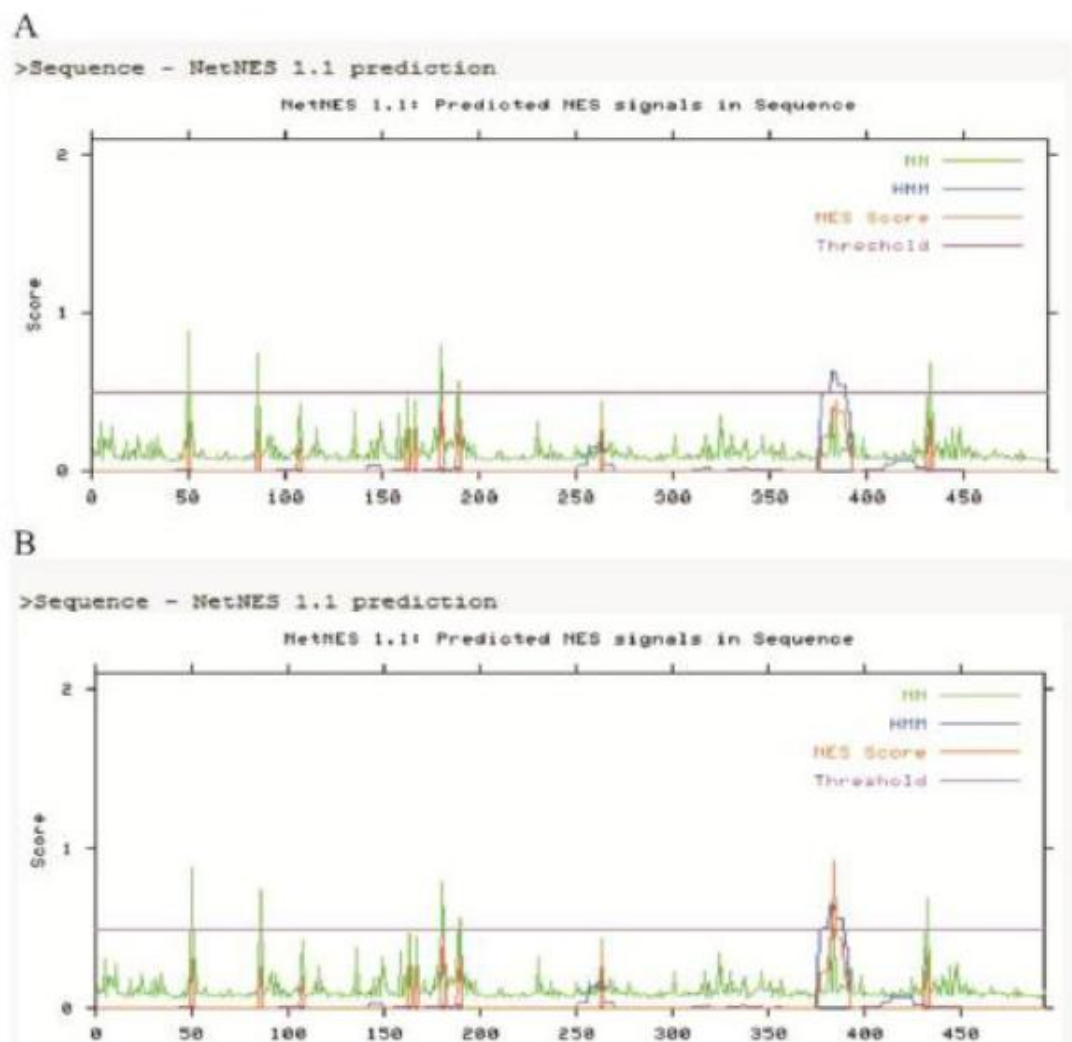
Το κυριότερο μονομερές σινιάλο εισόδου στον πυρήνα είναι το “KRIRRERRK” στη θέση 396-405 που περιέχει τα χαρακτηριστικά κατάλοιπα λυσίνης και αργινίνης για τις αλληλουχίες NLS (Boulikas, 1994). Επιπλέον, εντοπίζονται τα διμερή σινιάλα εισόδου στον πυρήνα στις θέσεις 259-272 [KR(X)₉KRR] και 404-419 [RK(X)₁₁KRK], τα οποία συμφωνούν με τα ήδη γνωστά μοτίβα (Kosugi *et al.*, 2009).

Αναφορικά με τα σινιάλα εξόδου από τον πυρήνα, η LEFKOTHEA περιέχει την αλληλουχία “GDLLDKDERLVIRERLIDLI” στη θέση 373-392, η οποία ταιριάζει σε μεγάλο βαθμό με το πρότυπο των NES [L(X)₂₋₃L(X)₁₋₂L] και συγκεντρώνει τα περισσότερα χαρακτηριστικά τους, όπως η ύπαρξη των αμινοξέων γλουταμινικό οξύ (E), ασπαρτικό οξύ (D) και σερίνη (S), το αρνητικό φορτίο και η έκθεση από την υπόλοιπη δομή της πρωτεΐνης (la Cour *et al.*, 2004).



Εικόνα 1.7: Απεικόνιση των δομικών χαρακτηριστικών υποκυτταρικής τοποθέτησης της LEFKOTHEA πάνω στην αμινοξική ακολουθία. TP: πεπτίδιο στόχευσης στο χλωροπλάστη, NLS: σινιάλο εισόδου στον πυρήνα, NES: σινιάλο εξόδου από τον πυρήνα.

Όπως γίνεται φανερό και μέσα από την Εικόνα 1.7, η υποκατάσταση του καταλοίπου γλυκίνης σε ασπαρτικό οξύ στη θέση 373 (μετάλλαξη *lefko2*) συμβαίνει μέσα στην αλληλουχία NES. Με τη βοήθεια του προγράμματος NetNES προβλέφθηκε ότι η πιθανότητα για λειτουργικό σινάλο εξαγωγής από τον πυρήνα είναι αρκετά αυξημένη στη μεταλλαγμένη σε σχέση με τη φυσιολογική πρωτεΐνη (Εικόνα 1.8). Η τάση αυτή εξαγωγής της *Lefko2* από τον πυρήνα του κυττάρου φαίνεται να σχετίζεται με προβλήματα στο μάτισμα των mRNAs πυρηνικών γονιδίων, τα οποία είναι απαραίτητα για την έκφραση σημαντικών χλωροπλαστικών γονιδίων και, με αυτόν τον τρόπο, προκαλούνται οι ανωμαλίες στην ανάπτυξη του φυτού που αναφέρθηκαν παραπάνω (Δάρας, 2011).



Εικόνα 1.8: Γραφική αναπαράσταση της πρόβλεψης του προγράμματος NetNES για αμινοξική ακολουθία που πιθανόν να αποτελεί σινάλο εξόδου της LEFKOTHEA από τον πυρήνα. Α. Αποτελέσματα για τη φυσιολογική αμινοξική αλληλουχία της πρωτεΐνης. Β. Αποτελέσματα για την αλληλουχία της πρωτεΐνης με την αμινοξική υποκατάσταση της γλυκίνης σε ασπαρτικό οξύ στη θέση 373.

1.4 Σκοπός της διατριβής

Με βάση τα υπάρχοντα πειραματικά δεδομένα καθώς και τα αποτελέσματα βιοπληροφορικών αναλύσεων, έχουν αποκαλυφθεί βασικές πληροφορίες σχετικά με την διπλή υποκυτταρική τοποθέτηση και την δράση της πρωτεΐνης LEFKOTHEA στο *Arabidopsis*. Συγκεκριμένα, έχει αποδειχθεί πως η LEFKOTHEA εντοπίζεται τόσο στους χλωροπλάστες συμμετέχοντας ενεργά στο μηχανισμό ματίσματος των ιντρονίων της ομάδας II όσο και στον πυρήνα όπου απαιτείται για την επεξεργασία πρώιμων μεταγράφων γονιδίων, τα οποία σχετίζονται με κύριες χλωροπλαστικές λειτουργίες.

Ωστόσο, ιδιαίτερο ενδιαφέρον παρουσιάζει η εύρεση των γονιδίων που ελέγχονται από τη δράση της LEFKOTHEA, και κατ' επέκταση η πρόβλεψη των σηματοδοτικών μονοπατιών και των κυτταρικών λειτουργιών στις οποίες εμπλέκεται. Για την επίτευξη των παραπάνω στόχων ενδείκνυται η χρήση των ομικών τεχνολογιών, και ειδικότερα της αλληλούχησης RNA (RNA-seq).

Στην παρούσα διατριβή επιχειρείται η αποσαφήνιση του ρόλου της LEFKOTHEA στα φυτικά κύτταρα του *Arabidopsis* με μία ολιστική προσέγγιση. Πιο αναλυτικά, έγινε απομόνωση και αλληλούχηση των μεταγράφων εμβρυικών κυττάρων από το στέλεχος αγρίου τύπου (*Col*) καθώς και από το στέλεχος *lefko1*, το οποίο φέρει τη θνησιγόνο μετάλλαξη της πρωτεΐνης. Τα δεδομένα που προέκυψαν από τις δύο αλληλουχίσεις χρησιμοποιήθηκαν για την *in silico* αναγνώριση γεγονότων εναλλακτικού ματίσματος μεταξύ των δύο δειγμάτων. Στόχος είναι μέσω της σύγκρισης αυτής να βρεθούν και να χαρακτηριστούν τα γονίδια, το μάτισμα των οποίων μεταβάλλεται όταν η LEFKOTHEA είναι μη λειτουργική, και, στη συνέχεια να ομαδοποιηθούν, ώστε να είναι εφικτή η εξαγωγή γενικότερων συμπερασμάτων για τις κυτταρικές διεργασίες που επηρεάζονται.

2. Μέθοδοι ανάλυσης

2.1 Προετοιμασία δειγμάτων και αλληλούχηση RNA

Το πρώτο βήμα για την πραγματοποίηση της αλληλούχησης ήταν η εξαγωγή του ολικού RNA από εμβρυικά κύτταρα *Arabidopsis* στο σφαιρικό/καρδιόσχημο στάδιο που φέρουν τη μετάλλαξη *lefko1* αλλά και από αγρίου τύπου (*Col-o*). Για το σκοπό αυτό χρησιμοποιήθηκε το kit NucleoSpin Plant RNA της Macherey-Nagel μαζί με DNase σύμφωνα με τις οδηγίες των κατασκευαστών. Η ποσότητα και η ποιότητα του RNA που απομονώθηκε εκτιμήθηκε με τη χρήση ενός φασματοφωτόμετρου NanoDrop 1000 και ηλεκτροφόρηση σε πηκτή αгарόζης.

Έπειτα, δημιουργήθηκαν οι απαιτούμενες βιβλιοθήκες αλληλούχησης RNA με το kit TruSeq Low Input της Illumina ακολουθώντας τις προβλεπόμενες οδηγίες. Η αλληλούχηση έγινε σε ένα μηχάνημα Illumina HiSeq 2000 στο Ινστιτούτο Γονιδιοματικής του Πεκίνου (BGI).

2.2 Ανάλυση πρωτογενών δεδομένων αλληλούχησης RNA

Η αλληλούχηση των μεταγράφων των δύο δειγμάτων (Albin και Col) οδήγησε στην παραγωγή τεσσάρων αρχείων fastq, δύο για κάθε δείγμα, που αντιστοιχούν στα διαβάσματα και των δύο αλυσίδων των μορίων RNA (paired-end reads).

Σε γενικές γραμμές ο τύπος αρχείων fastq χρησιμοποιείται ευρέως για την αποθήκευση δεδομένων αλληλούχησης, καθώς εκτός από τις αλληλουχίες νουκλεϊκών οξέων συμπεριλαμβάνει μία αριθμητική τιμή για τον έλεγχο της ποιότητας του κάθε νουκλεοτιδίου (Cock *et al.*, 2010). Ένα αρχείο fastq περιέχει πολλά διαβάσματα (reads), καθένα από τα οποία αποτελείται από τέσσερις διαφορετικές γραμμές (Εικόνα 2.1):

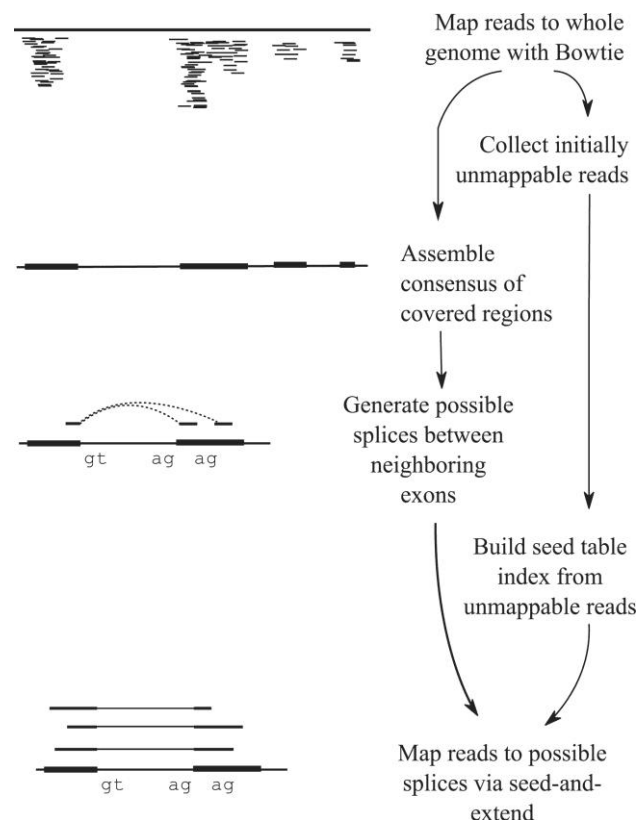
- Η πρώτη γραμμή ξεκινά με τον ειδικό χαρακτήρα “@” και περιλαμβάνει τον αναγνωριστικό τίτλο του διαβάσματος καθώς και προαιρετική περιγραφή
- Η δεύτερη γραμμή αντιστοιχεί στην νουκλεοτιδική αλληλουχία
- Η τρίτη γραμμή, η οποία σηματοδοτεί τη λήξη της αλληλουχίας, ξεκινά με τον ειδικό χαρακτήρα “+” και συνήθως περιέχει μια πλήρη επανάληψη των πληροφοριών της πρώτης γραμμής
- Τέλος, στην τέταρτη γραμμή αναγράφονται οι τιμές που καθορίζουν την ποιότητα του κάθε νουκλεοτιδίου της αλληλουχίας με τη μορφή εκτυπώσιμων χαρακτήρων ASCII (κυρίως του υποσυνόλου 33-126 ASCII)

```
@FCC4U32ACXX:8:1102:3425:43501#/1
ATTCTTGGATTTATGAAAGACGAACAACCTGCGAAAGCATTTGCCAAGGATGTTTTTCATTAATCAAGAACGAAAGTTGGGGGCTCGAAGACGATCAGATAC
+
bbbeeeegfgggfhnhiibghfghf`ghiihnehhihihihiehieibfghiiiaagh__bdY^b^bbaw^]`bZ_]acaacaccb_[ba]^^ab`b_Y
```

Εικόνα 2.1: Παράδειγμα ενός διαβάσματος αλληλούχισης RNA όπως αναπαριστάται σε ένα αρχείο τύπου fastq.

Αφού έγινε έλεγχος της ποιότητας των διαβασμάτων στα αρχεία fastq, ακολούθησε η στοίχισή τους στο γονιδίωμα του *Arabidopsis thaliana* με βάση την έκδοση TAIR10 χρησιμοποιώντας το πρόγραμμα TopHat (Trapnell *et al.*, 2009).

Το TopHat είναι ένα λογισμικό πακέτο, το οποίο αναγνωρίζει θέσεις ματίσματος *ab initio* χαρτογραφώντας σε μεγάλη κλίμακα διαβάσματα αλληλούχισης RNA σε ήδη γνωστά γονιδιώματα οργανισμών. Η ροή εργασίας που ακολουθεί παρουσιάζεται στην Εικόνα 2.2. Αρχικά, επιχειρείται η χαρτογράφηση των διαβασμάτων RNA σε ολόκληρο το γονιδίωμα του εκάστοτε οργανισμού με χρήση του προγράμματος Bowtie (Langmead *et al.*, 2009). Τα διαβάσματα που δεν χαρτογραφήθηκαν συλλέγονται και αποθηκεύονται σε ένα πίνακα, ενώ με τα χαρτογραφημένα διαβάσματα γίνεται απόπειρα συναρμολόγησης γονδιωματικών περιοχών επικάλυψης και δημιουργίας πιθανών θέσεων ματίσματος μεταξύ γειτονικών εξωνίων. Τέλος, τα μη χαρτογραφημένα διαβάσματα στοιχίζονται στις παραπάνω θέσεις ματίσματος.



Εικόνα 2.2: Διάγραμμα ροής εργασίας του προγράμματος TopHat (Trapnell *et al.*, 2009).

Στην ανάλυσή μας από τα 470 εκατομμύρια διαβάσματα μήκους 101 ζευγών βάσεων που δημιουργήθηκαν κατά μέσο όρο για κάθε δείγμα, το 90% στοιχήθηκε με επιτυχία στην αλληλουχία του γονιδιώματος του *Arabidopsis*.

Τα αποτελέσματα της στοίχισης αυτής με το TopHat αποθηκεύτηκαν σε ένα αρχείο τύπου BAM (Binary Alignment Map) για κάθε δείγμα. Τα αρχεία BAM αποτελούν τη δυαδική αναπαράσταση των αρχείων SAM (Sequence Alignment Map) και συμπίεζονται με τη βιβλιοθήκη BGZF. Ο τύπος αρχείων SAM χρησιμοποιείται ευρέως για την αποθήκευση στοίχισεων διαβασμάτων έναντι γονιδιωμάτων αναφοράς υποστηρίζοντας διαβάσματα μικρού και μεγάλου μεγέθους από διαφορετικές πλατφόρμες αλληλούχισης. Συνοπτικά, κάθε στοίχιση στο αρχείο SAM αποτελείται από το τμήμα της κεφαλίδας, οι γραμμές του οποίου ξεκινάν με τον ειδικό χαρακτήρα “@”, και το καθαυτό τμήμα της στοίχισης. Επίσης, αξίζει να αναφερθεί πως τα αρχεία SAM/BAM χρησιμοποιούνται συνήθως ταξινομημένα (sorted) με βάση τη θέση και κατατάσσονται σε πίνακα (indexing) με στόχο την ταχύτερη εύρεση αλληλεπικαλυπτόμενων στοίχισεων στην ίδια χρωμοσωμική περιοχή. Η διαχείριση και η ανάλυση των αρχείων SAM και BAM πραγματοποιείται με το λογισμικό SAMtools (Li *et al.*, 2009).

2.3 Προγράμματα εντοπισμού γεγονότων εναλλακτικού ματίσματος

Το κύριο μέρος της ανάλυσης περιλάμβανε την πρόβλεψη γεγονότων εναλλακτικού ματίσματος για τους 4 βασικούς τύπους (διατήρηση ιντρονίου, παράλειψη εξωνίου, εναλλακτική 5’ και 3’ θέση ματίσματος) με υπολογιστικές μεθόδους. Πιο συγκεκριμένα, χρησιμοποιήθηκαν 5 διαφορετικά προγράμματα, τα οποία μπορούν να ταξινομηθούν σε 2 κατηγορίες με βάση τον τύπο αρχείων εισόδου: εκείνα που δέχονται αρχεία τύπου BAM και εκείνα που δέχονται αρχεία τύπου GTF (Gene Transfer Format). Στην πρώτη κατηγορία ανήκουν τα προγράμματα SplAdder (Kahles *et al.*, 2016), rMATS (Shen *et al.*, 2014) και JuncBASE (Brooks *et al.*, 2011), ενώ στη δεύτερη τα ASprofile (Florean *et al.*, 2013) και ASATP (Ziding Zhang’s Laboratory, 2016).

Αναφορικά με τα GTF, είναι αρχεία κειμένου βασισμένα στον τύπο αρχείων GFF (General Feature Format) και χρησιμοποιούνται, όπως και τα GFF, για την αποθήκευση πληροφοριών σχετικά με την περιγραφή γονιδιωματικών περιοχών.

Κάθε γραμμή του αρχείου αντιστοιχεί σε ένα χαρακτηριστικό και περιέχει 9 στήλες δεδομένων με την ακόλουθη σειρά (Εικόνα 2.3):

1. Όνομα χρωμοσώματος
2. Πηγή προέλευσης του χαρακτηριστικού
3. Είδος χαρακτηριστικού π.χ. γονίδιο
4. Θέση έναρξης
5. Θέση λήξης
6. Τιμή σκορ
7. Κατεύθυνση αλυσίδας (+ ή -)
8. Πλαίσιο ανάγνωσης (τιμές : 0,1 ή 2)
9. Πρόσθετα στοιχεία π.χ. όνομα γονιδίου, όνομα μεταγράφου

Chr1	TAIR10	exon	3631	3913	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	3996	4276	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	4486	4605	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	4706	5095	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	5174	5326	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	5439	5899	.	+	.	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	3760	3913	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	3996	4276	.	+	2	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	4486	4605	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	4706	5095	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	5174	5326	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	CDS	5439	5630	.	+	0	transcript_id "AT1G01010.1"; gene_id "AT1G01010"; gene_name "AT1G01010";
Chr1	TAIR10	exon	5928	6263	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	6437	7069	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7157	7232	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7384	7450	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7564	7649	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7762	7835	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	7942	7987	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	8236	8325	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	8417	8464	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	exon	8571	8737	.	-	.	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	6915	7069	.	-	2	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7157	7232	.	-	0	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7384	7450	.	-	1	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7564	7649	.	-	0	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";
Chr1	TAIR10	CDS	7762	7835	.	-	2	transcript_id "AT1G01020.1"; gene_id "AT1G01020"; gene_name "AT1G01020";

Εικόνα 2.3: Τμήμα του αρχείου GTF των χαρακτηρισμένων γονιδίων του *Arabidopsis thaliana*.

Όπως γίνεται κατανοητό, σε αρχεία GTF/GFF περιέχονται και οι χαρακτηρισμένες περιοχές που αναγνωρίστηκαν μετά από την αλληλούχιση ολόκληρων των γονιδιωμάτων των οργανισμών. Στην εργασία αυτή χρησιμοποιήθηκαν τα αρχεία αναγνωρισμένων γονιδίων του *Arabidopsis thaliana* από τη βάση δεδομένων TAIR (The Arabidopsis Information Resource) (Berardini *et al.*, 2015), και συγκεκριμένα

από την έκδοση του γονιδιώματος TAIR10, ώστε να επιτευχθεί το ταίριασμα των διαβασμάτων των αλληλουχίσεων RNA με τα ήδη γνωστά γονίδια.

2.3.1 Splicing Adder (SplAdder)

Το λογισμικό πακέτο SplAdder συνιστά ένα καινοτόμο εργαλείο για την εύρεση και ποσοτικοποίηση γεγονότων εναλλακτικού ματίσματος από δεδομένα αλληλούχισης RNA. Ειδικότερα, δημιουργήθηκε στην γλώσσα προγραμματισμού python και οι τύποι εναλλακτικού ματίσματος που υποστηρίζει είναι η παράλειψη εξωνίου, η διατήρηση ιντρονίου, οι εναλλακτικές 5' και 3' θέσεις ματίσματος, η παράλειψη πολλαπλών εξωνίων και τα αμοιβαία αποκλειόμενα εξώνια. Ο αλγόριθμος που χρησιμοποιεί απαρτίζεται από πολλαπλά βήματα (Kahles *et al.*, 2016):

- Αρχικά δημιουργεί ένα γράφο ματίσματος από τα δεδομένα του αρχείου αναγνωρισμένων γονιδίων του οργανισμού ενδιαφέροντος
- Ακολουθεί ο εμπλουτισμός του γράφου με τα δεδομένα ματίσματος από τα αρχεία BAM των δειγμάτων που αλληλουχήθηκαν
- Στη συνέχεια γίνεται αναγνώριση των γεγονότων εναλλακτικού ματίσματος από τον εμπλουτισμένο γράφο και ποσοτικοποίησή τους με βάση τα δεδομένα της αλληλούχισης RNA για κάθε δείγμα

Στην ανάλυσή μας εκτελέστηκε το πρόγραμμα SplAdder δύο φορές, μία για το *lefko1* δείγμα (Albin αρχείο BAM) και μία για το δείγμα αγρίου τύπου (Col αρχείο BAM), κάνοντας και στις δύο περιπτώσεις χρήση του αρχείου GFF αναγνωρισμένων γονιδίων του *A. thaliana* από την έκδοση TAIR10. Τα αποτελέσματα αποθηκεύτηκαν σε αρχεία κειμένου με στήλες, ένα για κάθε τύπο εναλλακτικού ματίσματος για κάθε δείγμα. Κάθε γραμμή αντιπροσωπεύει ένα γεγονός εναλλακτικού ματίσματος και περιέχει πληροφορίες για το όνομα του γονιδίου, την κατεύθυνση της αλυσίδας, το χρωμόσωμα, τις συντεταγμένες καθώς και ποσοτικά δεδομένα (Εικόνα 2.4).

contig	strand	event_id	gene_name	exon_pre_start	exon_pre_end	exon_start	exon_end	exon_aft_start	exon_aft_end	Albin.sort:exon_pre_cov	Albin.sort:exon_aft_cov	Albin.sort:intron_pre_conf	Albin.sort:intron_aft_conf	Albin.sort:intron_skip_conf	Albin.sort:intron_pre_cov	Albin.sort:intron_aft_cov	Albin.sort:intron_skip_cov
bin.sort:psi																	
Chr1	+	exon_skip_1	AT1G01550	199639	199763	199890	199959	200511	201775	27.9	72.5	275.8	61	68	4	0.94	
Chr1	+	exon_skip_3	AT1G03365	826569	826737	827179	829182	829721	830383	6.9	6.0	35.5	3	30	4	0.80	
Chr1	+	exon_skip_6	AT1G05000	1426016	1426109	1426235	1426330	1427899	1427956	44.9	4.9	50.5	4	4	50	0.07	
Chr1	+	exon_skip_12	AT1G07170	2200123	2200219	2200463	2200539	2200638	2201265	26.0	4.8	80.9	4	5	67	0.06	
Chr1	+	exon_skip_17	AT1G08110	2535328	2535385	2535570	2535585	2535675	2535794	30.1	39.2	175.6	15	44	63	0.32	
Chr1	+	exon_skip_22	AT1G14630	5020620	5020671	5020832	5020924	5021174	5021537	18.0	19.6	29.0	13	28	3	0.87	
Chr1	+	exon_skip_23	AT1G15160	5216472	5216558	5216638	5216694	5216798	5217036	66.1	64.8	86.6	64	54	11	0.84	
Chr1	+	exon_skip_24	AT1G16020	5499192	5499375	5499475	5499513	5499641	5499750	6.5	4.8	13.0	6	3	6	0.43	
Chr1	+	exon_skip_30	AT1G20920	7285240	7288843	7289000	7289309	7289386	7289778	13.1	16.8	46.2	34	10	3	0.88	

Εικόνα 2.4: Πρώτες γραμμές του αρχείου εξόδου για γεγονότα παράλειψης εξωνίου στο δείγμα Albin από το πρόγραμμα SplAdder.

Για τα δεδομένα που προέκυψαν ακολουθήθηκε η εξής διαδικασία για τους 4 κύριους τύπους εναλλακτικού ματίσματος (Παράρτημα 1 και 2):

- Σύγκριση των γεγονότων εναλλακτικού ματίσματος κάθε τύπου ξεχωριστά μεταξύ των 2 δειγμάτων
- Απομόνωση των γεγονότων που εμφανίζονται ειδικά στο δείγμα *lefko1* και αυτών που εμφανίζονται ειδικά στο δείγμα αγρίου τύπου
- Τροποποίηση των 8 παραχθέντων αρχείων, ώστε να αποκτήσουν δομή 4 στηλών με τη σειρά: όνομα γονιδίου (AGI) – χρωμόσωμα – θέση έναρξης γεγονότος – θέση λήξης γεγονότος

2.3.2 replicate Multivariate Analysis of Transcript Splicing (rMATS)

Το υπολογιστικό πρόγραμμα rMATS περιλαμβάνει ένα στατιστικό μοντέλο σχεδιασμένο για τον εντοπισμό διαφορικού εναλλακτικού ματίσματος από δεδομένα αλληλούχισης RNA δίνοντας τη δυνατότητα ενσωμάτωσης και αντιγράφων στην ανάλυση. Είναι γραμμένο στη γλώσσα προγραμματισμού python και αναγνωρίζει όλους τους κύριους τύπους εναλλακτικού ματίσματος (διατήρηση ιντρονίου, παράλειψη εξωνίου, 5' και 3' εναλλακτικές θέσεις ματίσματος, αμοιβαία αποκλειόμενα εξώνια).

Αναφορικά με τη μέθοδό του, το rMATS χρησιμοποιεί ένα ιεραρχικό μοντέλο για τον ταυτόχρονο υπολογισμό της αβεβαιότητας των δειγμάτων σε μεμονωμένα μεταγράφα και της διασποράς μεταξύ των μεταγράφων. Επιπλέον, παρέχεται η επιλογή χρήσης του παραπάνω μοντέλου τόσο για την ανάλυση μη ταιριασμένων αντιγράφων όσο και ταιριασμένων αντιγράφων μεταξύ δύο ομάδων δειγμάτων (Shen *et al.*, 2014).

Στα πλαίσια της έρευνάς μας, το rMATS εκτελέστηκε μία φορά συμπεριλαμβάνοντας τις 2 ομάδες δειγμάτων (*lefko1* και αγρίου τύπου). Τα αρχεία εξαγωγής του προγράμματος, ένα για κάθε τύπο εναλλακτικού ματίσματος, έχουν παρόμοια δομή με τα αντίστοιχα του SplAdder, μόνο που σ' αυτήν την περίπτωση περιέχουν και πληροφορίες σχετικά με τη σύγκριση μεταξύ των 2 δειγμάτων λόγω της διαφορικής ανάλυσης, όπως η διαφορά των επιπέδων συμπερίληψης (IncLevel) (Εικόνα 2.5).

ID	GeneID	geneSymbol	chr	strand	riExonStart_0base	riExonEnd	upstreamES	upstreamEE	downstreamES	downstreamEE	ID
IC_SAMPLE_1	SC_SAMPLE_1	IC_SAMPLE_2	SC_SAMPLE_2	IncFormLen	SkipFormLen	PValue	FDR	IncLevel1	IncLevel2	IncLevelDI	
662	"AT4G39260"	"AT4G39260"	Chr4	-	18273898	18274567	18273898	18274185	18274416	18274567	662
2829	78	3600	47	330	99	4.6331663922e-05	0.011281760165	0.916	0.958	-0.042	
1081	"AT2G21940"	"AT2G21940"	Chr2	+	9352490	9352733	9352490	9352569	9352668	9352733	1081
0.000122830568458			0.0180525395566	0.333	0.0	0.333					12
588	"AT1G65980"	"AT1G65980"	Chr1	-	24559305	24559777	24559305	24559649	24559701	24559777	588
5679	102	4700	43	151	99	0.000148275478904	0.0180525395566	0.973	0.986	-0.013	
961	"AT5G66240"	"AT5G66240"	Chr5	+	26466203	26466443	26466203	26466274	26466363	26466443	961
72	8	79	37	188	99	0.000223513517356	0.0217702165905	0.826	0.529	0.297	
1019	"AT1G77765"	"AT1G77765"	Chr1	+	29245728	29246237	29245728	29245887	29245965	29246237	1019
0	18	4	1	177	99	0.000332358590242	0.026976438908	0.0	0.691	-0.691	
1074	"AT3G61600"	"AT3G61600"	Chr3	+	22797127	22798320	22797127	22797937	22798053	22798320	1074
378	32	171	2	215	99	0.000450688661826	0.0274356722887	0.845	0.975	-0.13	
515	"AT3G50670"	"AT3G50670"	Chr3	-	18827357	18828411	18827357	18827439	18828348	18828411	515
54	28	4	15	1008	99	0.000406746447963	0.0274356722887	0.159	0.026	0.133	
54	"AT4G12780"	"AT4G12780"	Chr4	+	7513758	7515229	7513758	7513910	7513991	7515229	54
0.00104497364538			0.0565446850333	1.0	0.216	0.784					12
									0	3	6
										180	99

Εικόνα 2.5: Πρώτες γραμμές του αρχείου εξόδου για γεγονότα διατήρησης ιντρονίου από το πρόγραμμα rMATS.

Η διαδικασία μετα-ανάλυσης των δεδομένων από το rMATS για τους 4 κύριους τύπους εναλλακτικού ματίσματος ήταν η ακόλουθη (Παράρτημα 1):

- Απομόνωση των γεγονότων που εμφανίζονται ειδικά στο δείγμα Albin και αυτών που εμφανίζονται ειδικά στο δείγμα Col με βάση τα υπολογισμένα επίπεδα συμπερίληψης
- Τροποποίηση των 8 παραχθέντων αρχείων, ώστε να αποκτήσουν δομή 4 στηλών με τη σειρά: όνομα γονιδίου (AGI) – χρωμόσωμα – θέση έναρξης γεγονότος – θέση λήξης γεγονότος.

2.3.3 Junction Based Analysis of Splicing Events (JuncBASE)

Το JuncBASE αποτελεί ένα λογισμικό που δημιουργήθηκε στη γλώσσα προγραμματισμού python και στοχεύει στην αναγνώριση και ταξινόμηση γεγονότων εναλλακτικού ματίσματος με βάση τις θέσεις ματίσματος των διαβασμάτων αλληλούχησης RNA και τις ήδη γνωστές συντεταγμένες των εξωνίων στο γονιδίωμα του οργανισμού ενδιαφέροντος. Το πρόγραμμα πραγματοποιεί διαφορική ανάλυση μεταξύ 2 ή περισσότερων δειγμάτων χρησιμοποιώντας τον αριθμό των διαβασμάτων για την ποσοτικοποίηση της σχετικής έκφρασης κάθε ισομορφής του ίδιου γονιδίου και δίνει τη δυνατότητα χαρακτηρισμού αναγνωρισμένων και μη αναγνωρισμένων γεγονότων εναλλακτικού ματίσματος διαφορετικών τύπων (διατήρηση ιντρονίου, παράλειψη εξωνίου, 5' και 3' εναλλακτικές θέσεις ματίσματος, αμοιβαία αποκλειόμενα εξώνια, εναλλακτικό πρώτο και τελευταίο εξώνιο) (Brooks *et al.*, 2011).

Η εκτέλεση του JuncBASE περιλαμβάνει μια σειρά βημάτων, τα οποία παραθέτονται παρακάτω:

- Δημιουργία μίας βάσης δεδομένων των αναγνωρισμένων γονιδίων από το αντίστοιχο αρχείο GTF (TAIR10)
- Επεξεργασία του κάθε αρχείου BAM
- Αναγνώριση των θέσεων ματίσματος
- Δημιουργία των αρχείων καταμέτρησης των θέσεων ματίσματος εξωνίων-ιντρονίων
- Δημιουργία ενός εικονικού δείγματος για τον καθορισμό όλων των θέσεων ματίσματος
- Αναγνώριση και ποσοτικοποίηση των γεγονότων εναλλακτικού ματίσματος κάθε δείγματος
- Δημιουργία πινάκων με τις πρωτογενείς και κανονικοποιημένες μετρήσεις των διαβασμάτων για τις αποκλειόμενες και συμπεριλαμβανόμενες ισομορφές

Το αποτέλεσμα της διαδικασίας αυτής είναι η παραγωγή 2 αρχείων κειμένου για όλα τα γεγονότα εναλλακτικού ματισμάτος, ένα με τις πρωτογενείς και ένα με τις κανονικοποιημένες μετρήσεις (Εικόνα 2.6). Η μετα-ανάλυση του αρχείου με τα πρωτογενή δεδομένα περιλάμβανε (Παράρτημα 1 και 2):

- Διόρθωση του προθέματος chrChr σε Chr στην στήλη που περιέχει το χρωμόσωμα του κάθε γεγονότος εναλλακτικού ματίσματος
- Διαχωρισμός σε ξεχωριστά αρχεία των γεγονότων των 4 κύριων τύπων εναλλακτικού ματίσματος
- Απομόνωση των γεγονότων που εμφανίζονται ειδικά στο Albin και ειδικά στο Col
- Διατήρηση των γεγονότων, τα οποία αφορούν ήδη χαρακτηρισμένα γονίδια
- Τροποποίηση των 8 παραχθέντων αρχείων, ώστε να αποκτήσουν δομή 4 στηλών με τη σειρά: όνομα γονιδίου (AGI) – χρωμόσωμα – θέση έναρξης γεγονότος – θέση λήξης γεγονότος

#Contains_Novel_or_Only_Known(Annotated)_Junctions	as_event_type	gene_name	chr	strand	exclusion_junctions	inclusion_junctions	exclu
sion_exons_inclusion_exons_intron-exon_junctions	neighboring_constitutive_exons	Albin	Col				
N jcn_only_AA	chrChr3 -	chrChr3:4952233-4952328	chrChr3:4952239-4952328		chrChr3:4952233-4952238	chrChr3:4952232-4952233	
13;6	7;2						
N jcn_only_AA	chrChr3 -	chrChr3:14198480-14198875;chrChr3:14198814-14198875;chrChr3:14198823-14198875			chrChr3:14198221-14198875		
0;0	575;70						
N jcn_only_AD	chrChr3 -	chrChr3:21925613-21925727	chrChr3:21925613-21925719		chrChr3:21925720-21925727		chrCh
r3:21925727-21925728	14;3	1;0					
N jcn_only_AD	chrChr2 +	chrChr2:2055968-2056300;chrChr2:2055962-2056300;chrChr2:2055959-2056300	chrChr2:2055935-2056300				
32;9	0;0						
N jcn_only_AD	chrChr3 -	chrChr3:18742695-18742903	chrChr3:18742695-18742882		chrChr3:18742883-18742903		chrCh
r3:18742903-18742904	62;2	54;1					
N intron_retention	chrChr4 -	chrChr4:13609733-13612069			chrChr4:13609732-13609733;chrChr4:13612069-13		
612070	13;383573	0;567					
N intron_retention	chrChr4 -	chrChr4:13612130-13614012			chrChr4:13612129-13612130;chrChr4:13614012-13		
614013	5;383295	0;558					
K intron_retention	AT3G14690.1	chrChr3 +	chrChr3:4938813-4938884		chrChr3:4938812-4938813;chrChr3:4938884-49388		
85	32;20	10;0					
N alternative_acceptor	AT5G13825.1	chrChr5 -	chrChr5:4465204-4465699	chrChr5:4465218-4465699	chrChr5:4465204-4465217	chrChr5:44652	
03-4465204	0;0	3;2					

Εικόνα 2.6: Πρώτες γραμμές του αρχείου εξόδου με τις πρωτογενείς μετρήσεις από το πρόγραμμα JuncBASE.

2.3.4 ASprofile

Το ASprofile αποτελεί ένα σύνολο προγραμμάτων που αναπτύχθηκαν στις γλώσσες προγραμματισμού C και perl με στόχο την εξαγωγή, ποσοτικοποίηση και σύγκριση γεγονότων εναλλακτικού ματίσματος από δεδομένα αλληλούχησης RNA. Πιο συγκεκριμένα, το πρόγραμμα “extract-as” δέχεται αρχεία εισόδου τύπου GTF και συγκρίνει όλα τα ζεύγη μεταγράφων εντός ενός γονιδίου για να προσδιορίσει διαφορές στη δομή εξωνίων-ιντρονίων. Οι τύποι εναλλακτικού ματίσματος που αναγνωρίζονται είναι η παράλειψη εξωνίου, η εναλλακτική θέση έναρξης και λήξης, η διατήρηση ιντρονίου και οι εναλλακτικές θέσεις ματίσματος (5’ και 3’) στα άκρα ενός εξωνίου. Τα προγράμματα “extract-as-fpkm” και “collect-fpkm” χρησιμοποιούνται αντίστοιχα για την ποσοτικοποίηση του κάθε γεγονότος εναλλακτικού ματίσματος υπολογίζοντας τον δείκτη FPKM (Fragments Per Kilobase Million) των μεταγράφων και για την σύγκριση μεταξύ διαφορετικών δειγμάτων αλληλούχησης RNA με βάση τις υπολογισμένες τιμές FPKM (Florencia *et al.*, 2013).

Η ανάλυση των δεδομένων μας με το ASprofile έγινε ως εξής (Παράρτημα 1 και 2):

- Μετατροπή των 2 αρχείων BAM σε GTF με το λογισμικό Cufflinks (Trapnell *et al.*, 2010), με το οποίο επιτυγχάνεται η συναρμολόγηση μεταγράφων από χαρτογραφημένα διαβάσματα αλληλούχησης RNA
- Διατήρηση των μεταγράφων με επιβεβαιωμένη έκφραση στο κάθε δείγμα (FPKM > 1)
- Εκτέλεση του προγράμματος “extract-as” 2 φορές και παραγωγή 2 αρχείων κειμένου (Εικόνα 2.7) με τα αποτελέσματα για τα 2 δείγματα ξεχωριστά

- Τροποποίηση των παραπάνω αρχείων, ώστε να αποκτήσουν δομή 5 στηλών με τη σειρά: τύπος εναλλακτικού ματίσματος – όνομα γονιδίου (AGI) – χρωμόσωμα – θέση έναρξης γεγονότος – θέση λήξης γεγονότος
- Διαχωρισμός σε ξεχωριστά αρχεία για κάθε δείγμα των γεγονότων των 4 κύριων τύπων εναλλακτικού ματίσματος (τα γεγονότα 5' και 3' εναλλακτικών θέσεων ματίσματος εντάσσονται σε μία κοινή κατηγορία από το ASprofile άρα τα συνολικά παραχθέντα αρχεία ήταν 6)
- Σύγκριση και απομόνωση των γεγονότων που εμφανίζονται ειδικά στο δείγμα *lefko1* και αυτών που εμφανίζονται ειδικά στο δείγμα αγρίου τύπου

#event_id	event_type	gene_id	chrom	event_start	event_end	event_pattern	strand	txpt_id	txpt_numexons	this_exon	num_alt_exons
1000001	TSS	AT1G01010	Chr1	3631	3913	+	1	1	3631,3996,4486,4706,5174,5439	3913,4276,460	
5,5095,5326,5899	0.845542	1688									
1000002	TTS	AT1G01010	Chr1	5439	5899	+	6	1	3631,3996,4486,4706,5174,5439	3913,4276,460	
5,5095,5326,5899	0.845542	1688									
1000003	TSS	AT1G01020	Chr1	8571	8737	-	1	1	6790,7157,7564,7762,7942,8236,8417,8571	7069,	
7450,7649,7835,7987,8325,8464,8737	1.239101	1085									
1000004	TTS	AT1G01020	Chr1	5928	6263	-	10	1	5928,6437,7157,7384,7564,7762,7942,8236,8417,		
8571	6263,7069,7232,7450,7649,7835,7987,8325,8464,8737	1.203476									
1000005	IR_ON	AT1G01020	Chr1	7232	7384	-	8	6	1	6790,7157,7564,7762,7942,8236,8417,85	
71	7069,7450,7649,7835,7987,8325,8464,8737	1.239101									
1000006	IR_OFF	AT1G01020	Chr1	7232	7384	-	10	7	1	5928,6437,7157,7384,7564,7762	
,7942,8236,8417,8571	6263,7069,7232,7450,7649,7835,7987,8325,8464,8737	1.203476									
1000007	TSS	AT1G01030	Chr1	13335	13714	-	1	1	11649,13335	13173,13714	0.026443
1905											
1000008	TTS	AT1G01030	Chr1	11649	13173	-	2	1	11649,13335	13173,13714	0.026443
1905											
1000009	TSS	AT1G01040	Chr1	23146	24451	+	1	1	23146,24542,24752,25041,25524,25825,26081,262		
92,26543,26862,27099,27372,27618,27803,28708,28890,29160,30147,30410,30902	24451,24655,24962,25435,25743,25997,26203,26452,26776,27012,27281,27533,27713,2843	1.218454									
1,28805,29080,30065,30311,30816,31227	6251										

Εικόνα 2.7: Πρώτες γραμμές του αρχείου εξόδου για το δείγμα Albin από το πρόγραμμα ASprofile.

2.3.5 Alternative Splicing Analysis Tool Package (ASATP)

Το ASATP περιλαμβάνει μια σειρά εργαλείων που αναπτύχθηκαν στη γλώσσα προγραμματισμού perl και χρησιμοποιούνται για τον εντοπισμό και την οπτικοποίηση γεγονότων εναλλακτικού ματίσματος, τον έλεγχο αλλαγών στο ανοιχτό αναγνωστικό πλαίσιο (ORF), την εκτίμηση της έκφρασης των εναλλακτικών μεταγράφων καθώς και την σύγκριση μεταξύ δειγμάτων μέσω στατιστικής ανάλυσης (Ziding Zhang's Laboratory for Protein Bioinformatics, 2016). Ειδικότερα, τα εργαλεία αυτά είναι τα ακόλουθα:

- ASRecovist: αναγνωρίζει γεγονότα εναλλακτικού ματίσματος παράλειψης εξωνίου, διατήρησης ιντρονίου, αμοιβαία αποκλειόμενων εξωνίων,

εναλλακτικών 5' και 3' θέσεων ματίσματος και εναλλακτικού πρώτου και τελευταίου εξωνίου

- ASQuantityDiff: εκτελεί ποσοτική σύγκριση της έκφρασης των μεταγράφων μεταξύ διαφορετικών δειγμάτων
- ASaffectORF: ελέγχει την εμφάνιση γεγονότων εναλλακτικού ματίσματος μέσα σε κωδικές περιοχές του γονιδιώματος
- runASATP: συνδυάζει τα 3 παραπάνω προγράμματα
- τα βοηθητικά προγράμματα asp2bit, bit2asp και splitGtf

Για την εύρεση των γεγονότων εναλλακτικού ματίσματος από τα δεδομένα του πειράματός μας πραγματοποιήθηκαν τα παρακάτω βήματα (Παράρτημα 1 και 2):

- Μετατροπή των 2 αρχείων BAM σε GTF με το λογισμικό Cufflinks (Trapnell *et al.*, 2010)
- Διατήρηση των μεταγράφων με επιβεβαιωμένη έκφραση στο κάθε δείγμα (FPKM > 1)
- Εκτέλεση του ASReconist μέσω του διακομιστή δικτύου (web server) του ASATP με αρχεία εισόδου τα GTF που δημιουργήθηκαν παραπάνω τόσο για το δείγμα *lefko1* όσο και για το δείγμα αγρίου τύπου
- Σύγκριση και απομόνωση των γεγονότων που εμφανίζονται ειδικά στο Albin και αυτών που εμφανίζονται ειδικά στο Col
- Διαχωρισμός σε ξεχωριστά αρχεία για κάθε δείγμα των γεγονότων των 4 κύριων τύπων εναλλακτικού ματίσματος
- Τροποποίηση των παραπάνω αρχείων, ώστε να αποκτήσουν δομή 4 στηλών με τη σειρά: όνομα γονιδίου (AGI) – χρωμόσωμα – θέση έναρξης γεγονότος – θέση λήξης γεγονότος

Gene	Chromosome	Strand	Transcript1	Transcript2	AS Pattern	AS Event Type	AS Event Span Unit	AS Bit Span Unit
AT4G25880	Chr4	+	AT4G25880.1	AT4G25880.2	4	A3SS	13157237-13157351	13156110-13157236,13157237-13157333,13157334-13157351
,13157352-13157475	Chr4	+	AT4G25880.1	AT4G25880.3	4	A3SS	13157237-13157351	13156110-13157236,13157237-13157333,13157334-13157351
AT4G25880	Chr4	+	AT4G25880.1	AT4G25880.3	8	A5SS	13158744-13159055	13158572-13158743,13158744-13158760,13158761-13159055
,13159056-13159382	Chr4	+	AT4G25880.2	AT4G25880.3	8	A5SS	13158744-13159055	13158572-13158743,13158744-13158760,13158761-13159055
AT4G25880	Chr4	+	AT4G29950.1	AT4G29950.2	4	A3SS	14657888-14657975	14657784-14657887,14657888-14657971,14657972-14657975
,13159056-13159382	Chr4	+	AT4G29950.1	AT4G29950.2	4	A3SS	14657888-14657975	14657784-14657887,14657888-14657971,14657972-14657975
AT4G29950	Chr4	+	AT4G29950.1	AT4G29950.2	4	A3SS	14657888-14657975	14657784-14657887,14657888-14657971,14657972-14657975
,14657976-14658148	Chr5	-	AT5G07360.1	AT5G07360.2	4	A3SS	2328146-2328272	2328273-2328427,2328167-2328272,2328146-2328166,2328030-23281
AT5G07360	Chr5	-	AT5G07360.1	AT5G07360.2	4	A3SS	2328146-2328272	2328273-2328427,2328167-2328272,2328146-2328166,2328030-23281
45	Chr3	-	AT3G46440.1	AT3G46440.2	2	IR	17091636-17091724	17091725-17091826,17091636-17091724,17091534-17091635
AT3G46440	Chr3	-	AT5G03560.1	AT5G03560.2	17T2	ALE	900838-902411	902412-902748,902236-902411,901323-902235,901169-901322,90083
8-901168	Chr1	-	AT1G76280.1	AT1G76280.3	4	A3SS	28621769-28621954	28621955-28622142,28621781-28621954,28621769-28621780
AT1G76280	Chr1	-	AT1G76280.1	AT1G76280.3	16	CE	28620161-28620705	28620706-28621149,28620364-28620705,28620268-28620363
,28621673-28621768	Chr1	-	AT1G76280.1	AT1G76280.3	16	CE	28620161-28620705	28620706-28621149,28620364-28620705,28620268-28620363
AT1G76280	Chr1	-	AT1G76280.1	AT1G76280.3	16	CE	28620161-28620705	28620706-28621149,28620364-28620705,28620268-28620363
,28620161-28620267,28619987-28620160								

Εικόνα 2.8: Πρώτες γραμμές του αρχείου εξόδου για το δείγμα Albin από το πρόγραμμα ASReconist του πακέτου ASATP.

2.4 Ανάλυση εμπλουτισμού ομάδων γονιδίων

Η ανάλυση εμπλουτισμού ομάδων γονιδίων συνιστά μία ευρέως χρησιμοποιούμενη μέθοδο για την εξαγωγή βιολογικών πληροφοριών από μία λίστα γονιδίων, η οποία προέκυψε από αναλύσεις μεγάλης κλίμακας όπως η αλληλούχηση RNA. Πιο συγκεκριμένα, με την ανάλυση εμπλουτισμού ερευνάται η υπερεμφάνιση βιολογικών όρων στη λίστα αυτή με βάση τα ήδη γνωστά δεδομένα για τα γονίδια του οργανισμού ενδιαφέροντος που προέρχονται από βάσεις δεδομένων σχετιζόμενες συνήθως με οντολογία γονιδίων (Gene Ontology – GO), αλλά και με επιπλέον βιολογικές κατηγορίες ανάλογα με τη μέθοδο χρήσης, όπως η πρωτεϊνική βάση UniProt (UP) και οι βάσεις SMART και INTERPRO που χρησιμοποιούνται για εύρεση και ταξινόμηση πρωτεϊνικών επικρατειών και άλλων σημαντικών μοτίβων.

Από τα διάφορα διαθέσιμα λογισμικά για την ανάλυση εμπλουτισμού γονιδίων επιλέχθηκαν στην ερευνητική αυτή εργασία τα εργαλεία της βάσης δεδομένων Βιοπληροφορικής DAVID (Database for Annotation, Visualization and Integrated Discovery) λόγω τόσο των πολλαπλών εφαρμογών που παρέχουν όσο και της πρόσφατης ανανέωσής τους. Ειδικότερα, σε σύγκριση με παρόμοια προγράμματα ανάλυσης εμπλουτισμού, ο αλγόριθμος της DAVID αντλεί δεδομένα από περισσότερες από 40 κατηγορίες αναγνωρισμένων βιολογικών πληροφοριών, όπως όροι οντολογίας γονιδίων, αλληλεπιδράσεις πρωτεϊνών, λειτουργικές επικράτειες πρωτεϊνών, βιο-μονοπάτια, χαρακτηριστικά αλληλουχίας και βιβλιογραφικές πηγές, επιτρέποντας μ' αυτόν τον τρόπο την πολύπλευρη εξαγωγή βιολογικών πληροφοριών για τα εξεταζόμενα γονίδια (Huang *et al.*, 2009). Παρακάτω παραθέτονται οι βασικές δυνατότητες των εργαλείων DAVID για μία λίστα γονιδίων:

- Αναγνώριση εμπλουτισμένων βιολογικών όρων
- Αποκάλυψη λειτουργικά σχετιζόμενων ομάδων γονιδίων
- Δισδιάστατη διαγραμματική απεικόνιση των σχετιζόμενων γονιδίων με τους αντίστοιχους βιολογικούς όρους
- Έλεγχος για την ύπαρξη επιπλέον σχετιζόμενων γονιδίων εκτός της λίστας εισαγωγής
- Οπτικοποίηση γονιδίων σε χάρτες μονοπατιών των βάσεων δεδομένων KEGG (Kanehisa and Goto, 2000) και BioCarta (Nishimura, 2004)
- Εμφάνιση αλληλεπιδρώντων πρωτεϊνών καθώς και λειτουργικών πρωτεϊνικών επικρατειών και μοτίβων
- Ανακατεύθυνση σε σχετιζόμενα βιβλιογραφικά δεδομένα

2.5 Έλεγχος συμφωνίας των αλληλουχιών στις θέσεις εναλλακτικού ματίσματος

Το τελευταίο μέρος της ανάλυσης στην παρούσα διατριβή περιλάμβανε την εύρεση και σύγκριση των αλληλουχιών των θέσεων ματίσματος για τα κοινά γεγονότα εναλλακτικού ματίσματος που εντοπίστηκαν. Ειδικότερα, δόθηκε έμφαση στα γεγονότα διατήρησης ιντρονίου, καθώς αυτά εμφανίζονται με μεγαλύτερη συχνότητα στους φυτικούς οργανισμούς. Μ' αυτόν τον τρόπο επιχειρήθηκε να αποσαφηνιστεί εάν τα ιντρόνια, στο μάτισμα των οποίων συμμετέχει η πρωτεΐνη LEFKOTHEA, ανήκουν στην κατηγορία των U2 ιντρονίων και ή αποτελούν εξαίρεση και ανήκουν είτε στην κατηγορία των U12 ιντρονίων είτε σε κάποια ξεχωριστή κατηγορία.

Η διαδικασία που εκτελέστηκε ήταν η εξής (Παράρτημα 1 και 2):

- Ενσωμάτωση της κατεύθυνσης της αλυσίδας (+ ή -) στα ήδη επεξεργασμένα αρχεία κειμένου που περιέχουν τα γεγονότα διατήρησης ιντρονίου
- Διαχωρισμός των γεγονότων με κατεύθυνση 5' -> 3' (+) και αυτών με κατεύθυνση 3' -> 5' (-)
- Άντληση των αλληλουχιών DNA που αντιστοιχούν στην αρχή και το τέλος του ιντρονίου από το αρχείο fasta του γονιδιώματος του *A. thaliana* και για τις 2 παραπάνω κατηγορίες με τη βοήθεια της εντολής getfasta του προγράμματος BEDTools (Quinlan and Hall, 2010)
- Δημιουργία των ανάστροφων συμπληρωματικών αλληλουχιών με το script reverse_negative_strand.pl για τις αλληλουχίες με κατεύθυνση 3' -> 5'
- Ένωση όλων των αλληλουχιών σε ένα αρχείο fasta και απεικόνιση των αποτελεσμάτων με το λογισμικό WebLogo (Crooks *et al.*, 2004; Schneider and Stephens, 1990)

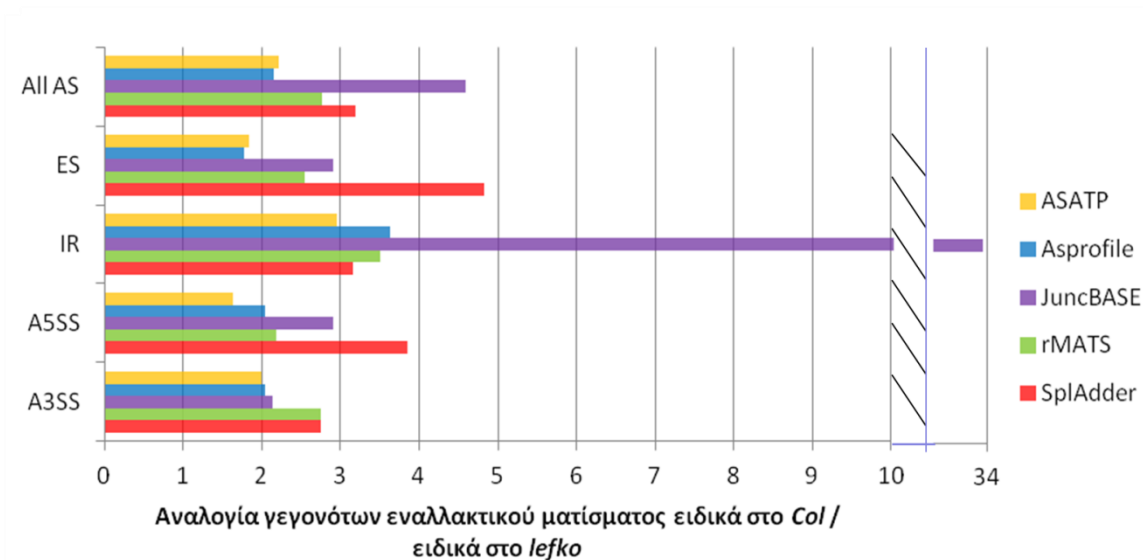
3. Αποτελέσματα

3.1 Σύγκριση των γεγονότων εναλλακτικού ματίσματος από τα 5 προγράμματα

Το πρώτο βήμα για την αξιοποίηση των αποτελεσμάτων που παρήγαγαν τα 5 λογισμικά εντοπισμού γεγονότων εναλλακτικού ματίσματος περιλάμβανε την συγκριτική τους ανάλυση υπολογίζοντας για καθέναν από τους 4 κύριους τύπους εναλλακτικού ματίσματος το λόγο (λ) εμφάνισης γεγονότων ειδικά στο δείγμα αγρίου τύπου προς τα γεγονότα που εμφανίζονται ειδικά στο *lefko* (Πίνακας 3.1 και Εικόνα 3.1).

Πίνακας 3.1: Γεγονότα εναλλακτικού ματίσματος ειδικά στο δείγμα αγρίου τύπου / ειδικά στο δείγμα *lefko* και οι υπολογιζόμενες αναλογίες (λ).

	A3SS	A5SS	IR	ES	All AS
SplAdder	106 / 292 (1 : 2.75)	42 / 162 (1 : 3.85)	62 / 196 (1 : 3.16)	12 / 58 (1 : 4.83)	222 / 708 (1 : 3.19)
rMATS	32 / 88 (1 : 2.75)	21 / 46 (1 : 2.19)	46 / 161 (1 : 3.5)	90 / 230 (1 : 2.55)	189 / 525 (1 : 2.77)
JuncBASE	174 / 372 (1 : 2.13)	92 / 267 (1 : 2.90)	23 / 778 (1 : 33.82)	53 / 154 (1 : 2.90)	342 / 1571 (1 : 4.59)
ASprofile	279 / 570 (1 : 2.04)		47 / 171 (1 : 3.63)	102 / 182 (1 : 1.78)	428 / 923 (1 : 2.15)
ASATP	291 / 582 (1 : 2)	165 / 272 (1 : 1.64)	252 / 747 (1 : 2.96)	76 / 140 (1 : 1.84)	784 / 1741 (1 : 2.22)



Εικόνα 3.1: Διαγραμματική απεικόνιση των λόγων (λ) εμφάνισης γεγονότων εναλλακτικού ματίσματος για τους 4 κύριους τύπους ειδικά στο *Col* προς ειδικά στο *lefko*.

Όπως γίνεται φανερό, τα γεγονότα εναλλακτικού ματίσματος είναι σε όλες τις περιπτώσεις περισσότερα στο μεταλλαγμένο στέλεχος σε σχέση με το στέλεχος αγρίου τύπου με τις αναλογίες να κυμαίνονται από 1:1.64 έως και 1:33.82. Ειδικότερα, ο αριθμός των γεγονότων και των 4 τύπων εναλλακτικού ματίσματος είναι περίπου διπλάσιος με τα προγράμματα ASprofile και ASATP και τριπλάσιος με τα SplAdder και rMATS, ενώ με το JuncBASE η αναλογία αυτή προσεγγίζει το 1:5.

Παρατηρώντας τους λόγους για καθέναν από τους επιμέρους τύπους εναλλακτικού ματίσματος διαπιστώνονται τα εξής:

- Τόσο για τα γεγονότα παράλειψης εξωνίου όσο και για τα γεγονότα εναλλακτικής 5' θέσης ματίσματος, τα προγράμματα rMATS, JuncBASE, ASprofile και ASATP σχηματίζουν μία ομάδα με παρόμοιες αναλογίες. Το SplAdder αποκλίνει από την ομάδα αυτή, καθώς η τιμή της αναλογίας του κινείται ελαφρώς ανοδικά και στις δύο περιπτώσεις.
- Στα γεγονότα εναλλακτικής 3' θέσης ματίσματος οι λόγοι και των 5 προγραμμάτων είναι σχεδόν οι ίδιοι.
- Αναφορικά με τα γεγονότα διατήρησης ιντρονίου, η τιμή που εμφανίζει το JuncBASE ανέρχεται περίπου στο 39, σε αντίθεση με τα υπόλοιπα προγράμματα, οι τιμές των οποίων συγκεντρώνονται γύρω από το 3.

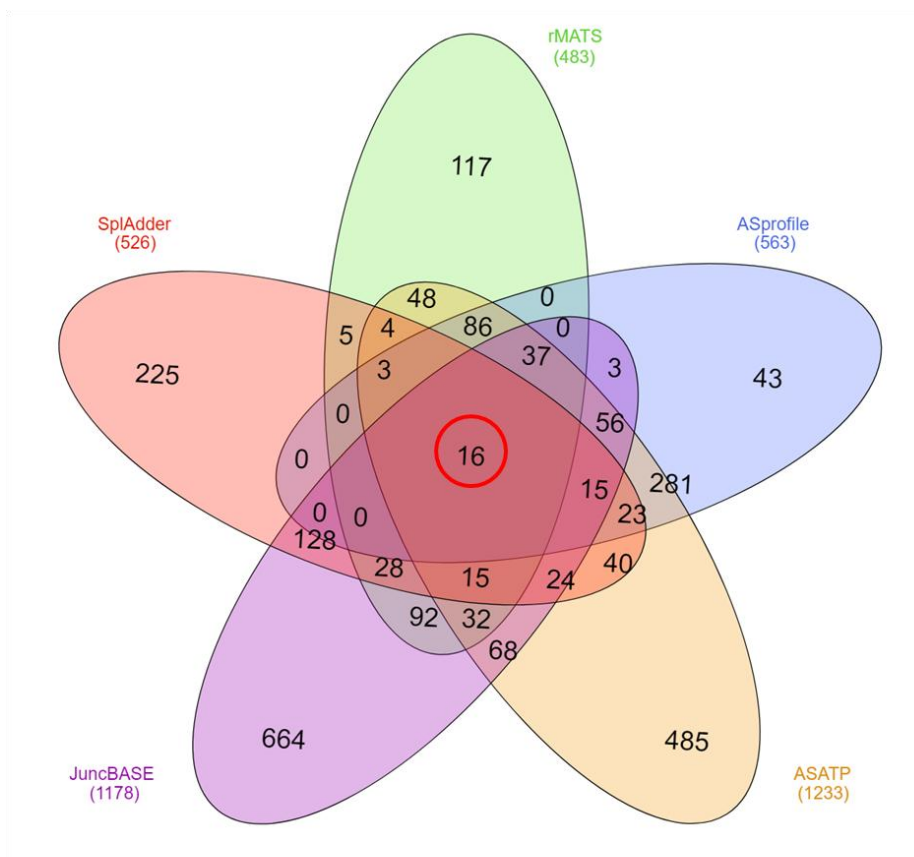
Ένα επιπρόσθετο κριτήριο που χρησιμοποιήθηκε για την αξιολόγηση της σημαντικότητας του κάθε εντοπισμένου γεγονότος εναλλακτικού ματίσματος είναι η τιμή του επιπέδου συμπερίληψης (inclusion level – percent spliced in – PSI/Ψ). Γενικά, η τιμή PSI υπολογίζεται ως ο λόγος των διαβασμάτων, στα οποία

περιλαμβάνεται το ιντρόνιο ή το εξώνιο ή το εναλλακτικό μετάγραφο, προς το συνολικό αριθμό των διαβασμάτων (Schafer *et al.*, 2015).

3.2 Κοινά γεγονότα και για τους 4 τύπους εναλλακτικού ματίσματος

Το επόμενο βήμα της ανάλυσης περιλάμβανε την εύρεση αλληλεπικαλυπτόμενων γεγονότων εναλλακτικού ματίσματος μεταξύ των προγραμμάτων ειδικά στο δείγμα που φέρει τη μετάλλαξη *lefko1* με στόχο τον χαρακτηρισμό των γονιδίων, το μάτισμα των οποίων εξαρτάται από την λειτουργία της LEFKOTHEA. Γι' αυτό το λόγο δημιουργήθηκαν, διαγράμματα Venn με το διαδικτυακό εργαλείο Interactivenn (Heberle *et al.*, 2015).

Σε πρώτη φάση πραγματοποιήθηκε σύγκριση των γεγονότων εναλλακτικού ματίσματος και των 5 λογισμικών και για τους 4 διαφορετικούς τύπους με βάση μόνο τα αναγνωριστικά γονιδίων (AGI), ώστε να εξαχθούν γενικότερα συμπεράσματα για το βαθμό αλληλεπικαλύψεων μεταξύ των προγραμμάτων. Το διάγραμμα που παρουσιάζεται στην Εικόνα 3.2 αναδεικνύει την ύπαρξη 16 κοινών γονιδίων, τα οποία εμφανίζουν κάποιο γεγονός εναλλακτικού ματίσματος και στα 5 προγράμματα. Η λίστα των παραπάνω γονιδίων παρατίθεται στον Πίνακα 3.2. Επιπλέον, παρατηρείται μικρότερος αριθμός αλληλεπικαλυπτόμενων γονιδίων μεταξύ του SplAdder και των υπόλοιπων λογισμικών.



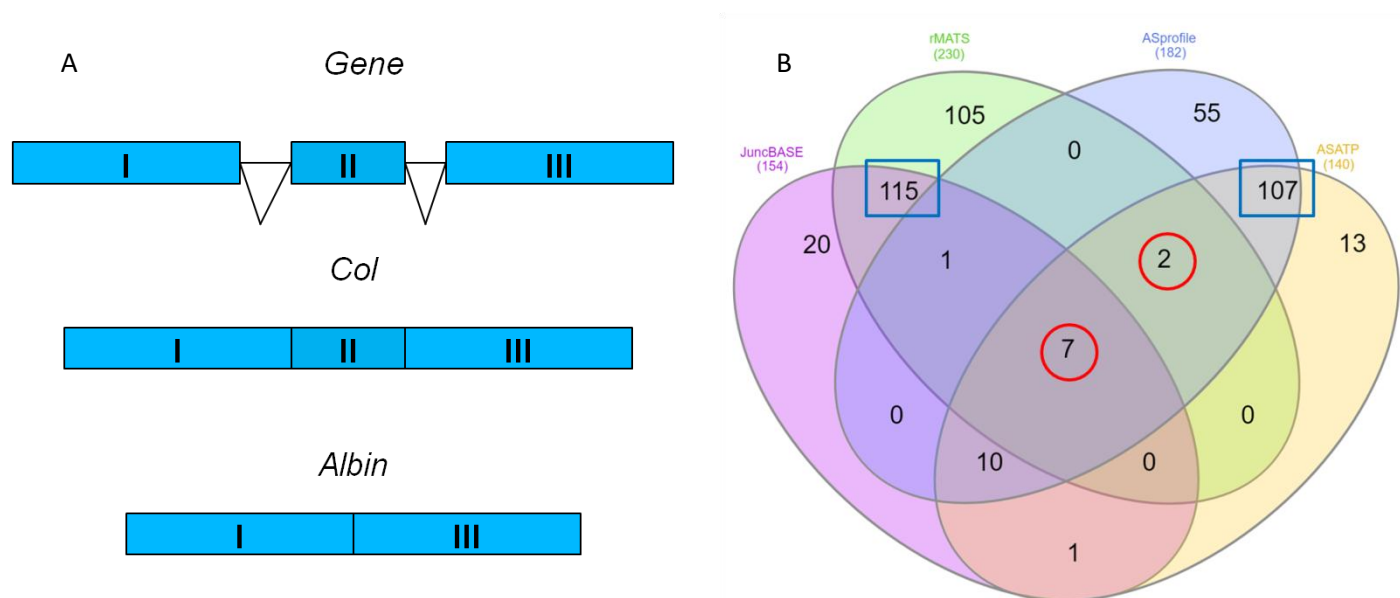
Εικόνα 3.2: Απεικόνιση σε διάγραμμα Venn των αλληλεπικαλυπτόμενων γεγονότων εναλλακτικού ματίσματος και των 5 προγραμμάτων και για τους 4 κύριους τύπους με βάση μόνο τα αναγνωριστικά γονιδίων (AGI) ειδικά στο *lefko*.

Πίνακας 3.2: Τα 16 κοινά γονίδια που εμφανίζουν κάποιο γεγονός εναλλακτικού ματίσματος και στα 5 προγράμματα ειδικά στο δείγμα *lefko*.

ID	Όνομα γονιδίου
AT1G02850	beta glucosidase 11(BGLU11)
AT1G07420	sterol 4-alpha-methyl-oxidase 2-1(SMO2-1)
AT1G61150	LisH and RanBPM domains containing protein(AT1G61150)
AT1G70620	cyclin-like protein(AT1G70620)
AT2G38880	nuclear factor Y, subunit B1(NF-YB1)
AT2G44470	beta glucosidase 29(BGLU29)
AT3G58030	RING/U-box superfamily protein(AT3G58030)
AT4G17150	alpha/beta-Hydrolases superfamily protein(AT4G17150)
AT4G19160	transglutaminase family protein(AT4G19160)
AT4G20380	LSD1 zinc finger family protein(LSD1)
AT5G01670	NAD(P)-linked oxidoreductase superfamily protein(AT5G01670)
AT5G19400	Telomerase activating protein Est1(SMG7)
AT5G43930	Transducin family protein / WD-40 repeat family protein(AT5G43930)
AT5G57060	60S ribosomal L18a-like protein(AT5G57060)
AT5G65050	AGAMOUS-like 31(AGL31)
AT5G65070	K-box region and MADS-box transcription factor family protein(MAF4)

3.3 Κοινά γεγονότα παράλειψης εξωνίου

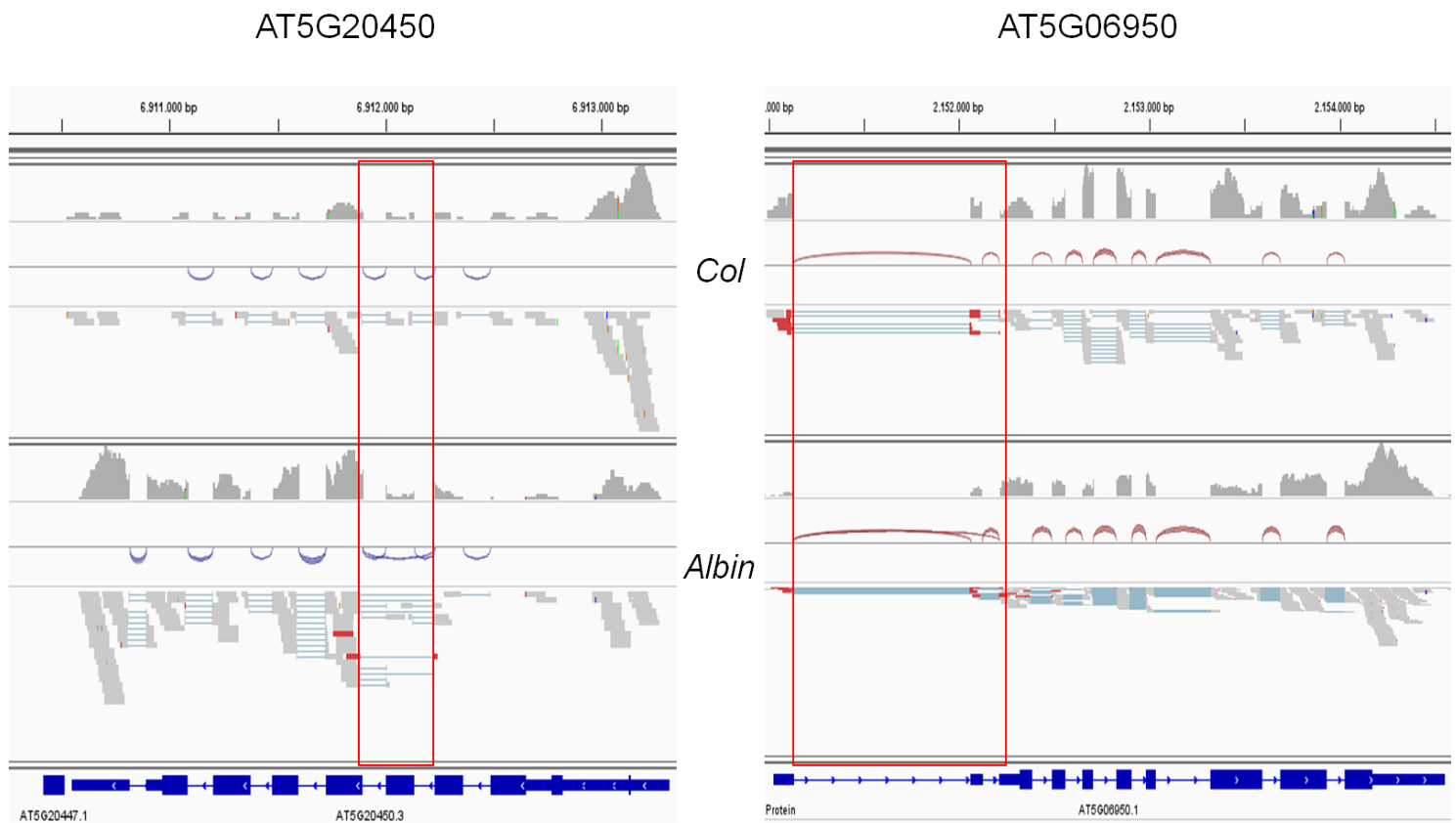
Στη συνέχεια, εξετάστηκε η ύπαρξη αλληλεπικαλυπτόμενων γεγονότων εναλλακτικού ματίσματος σε καθέναν από τους 4 τύπους ειδικά στο δείγμα *lefko1*. Όσον αφορά τα γεγονότα παράλειψης εξωνίου, στην ανάλυση χρησιμοποιήθηκαν τα δεδομένα (αναγνωριστικά γονιδίων και συντεταγμένες γεγονότων) των προγραμμάτων rMATS, JuncBASE, ASprofile και ASATP, ενώ δεν συμπεριλήφθηκαν αυτά του SplAdder με κριτήριο τις υπολογισμένες τιμές του λόγου του αριθμού γεγονότων ειδικά στο δείγμα *Col* προς τον αντίστοιχο αριθμό ειδικά στο δείγμα *Albin* (Πίνακας 3.1 και Εικόνα 3.1).



Εικόνα 3.3: Α. Σχηματική αναπαράσταση ενός γεγονότος παράλειψης εξωνίου. Β. Απεικόνιση σε διάγραμμα Venn των αλληλεπικαλυπτόμενων γεγονότων παράλειψης εξωνίου (ES) των προγραμμάτων rMATS, JuncBASE, ASprofile και ASATP με βάση τα αναγνωριστικά γονιδίων (AGI) και τις συντεταγμένες των γεγονότων (χρωμόσωμα-έναρξη εξωνίου-τέλος εξωνίου) ειδικά στο *lefko*.

Στα αποτελέσματα (Εικόνα 3.3) εμφανίζονται 7 κοινά γεγονότα μεταξύ των 4 λογισμικών και επιπλέον 2 μεταξύ των rMATS-ASprofile-ASATP. Η τριάδα αυτή επιλέχθηκε εξαιτίας του μικρότερου εύρους των τιμών λ. Τα 9 γεγονότα καταγράφονται στο Παράρτημα 3 μαζί με τις αντίστοιχες τιμές PSI, όπως προέκυψαν από τα προγράμματα rMATS και ASprofile. Επίσης, άξια αναφοράς κρίνεται η εμφάνιση πλήθους επιπρόσθετων αλληλεπικαλυπτόμενων γεγονότων παράλειψης εξωνίου στις δυάδες rMATS-JuncBASE και ASprofile-ASATP (115 και 107 αντίστοιχα).

Ενδεικτικά παραθέτονται στην Εικόνα 3.4 τα 2 σημαντικότερα κοινά γεγονότα παράλειψης εξωνίου. Η απεικόνιση έγινε με το πρόγραμμα IGV (Robinson et al., 2011; Thorvaldsdottir et al., 2013).



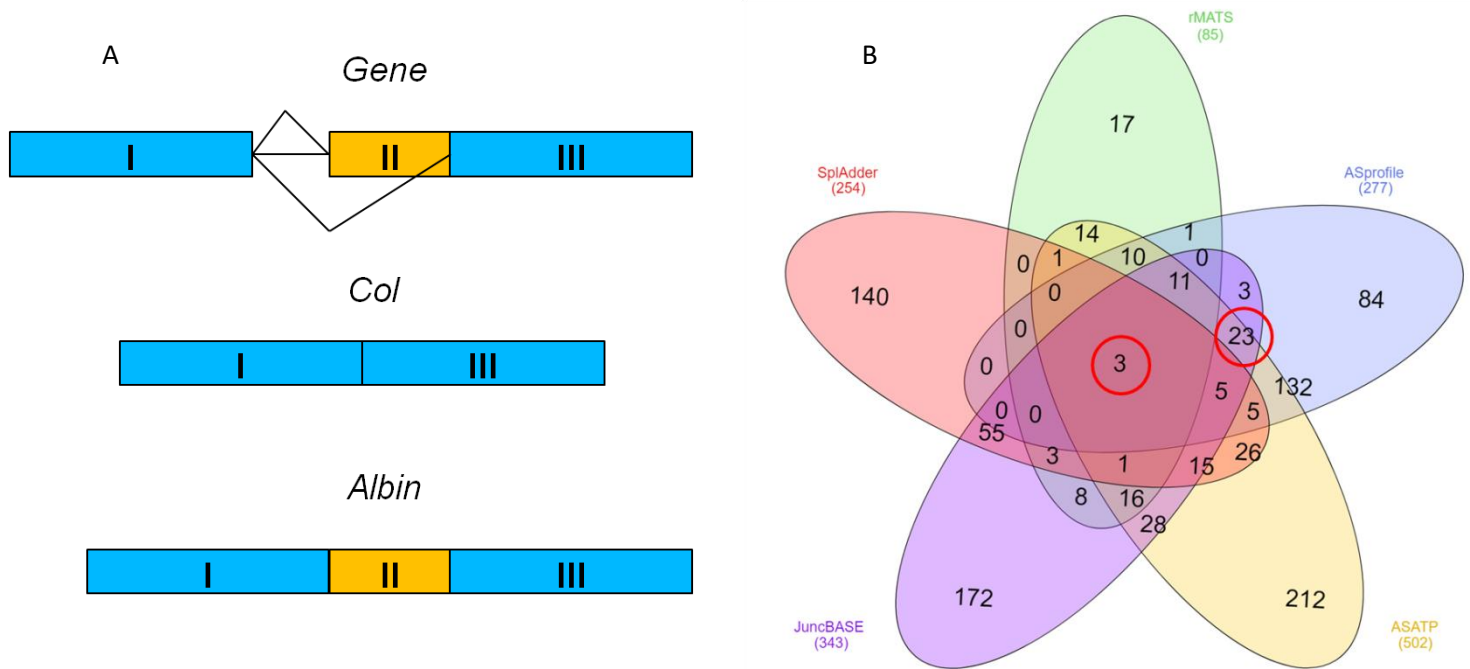
Εικόνα 3.4: Απεικόνιση των 2 σημαντικότερων κοινών γεγονότων παράλειψης εξωνίου με το πρόγραμμα IGV (Integrative Genomics Viewer).

3.4 Κοινά γεγονότα εναλλακτικής 3' θέσης ματίσματος

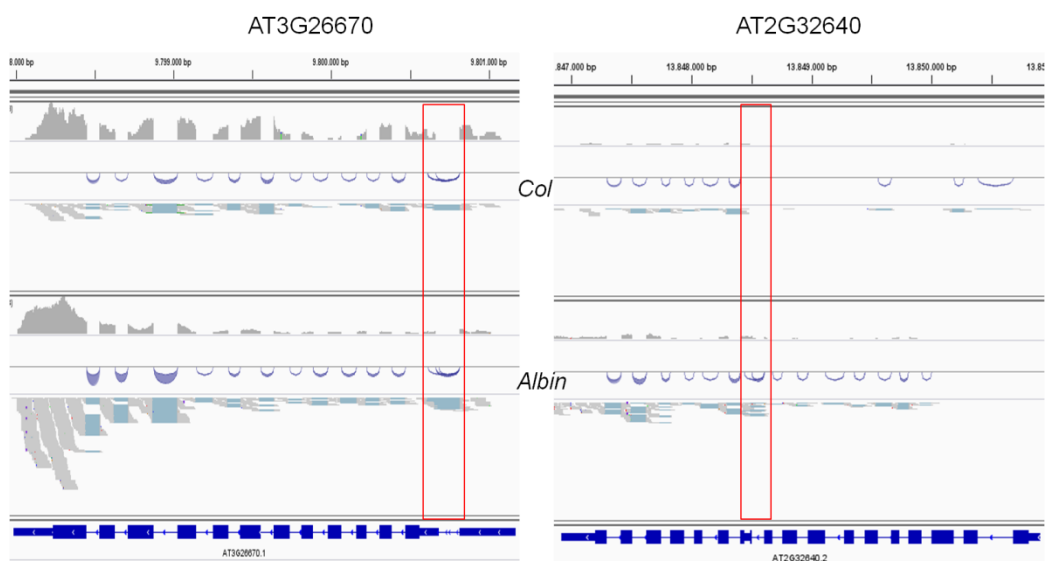
Έπειτα, ακολούθησε η εύρεση κοινών γεγονότων εναλλακτικής 3' θέσης ματίσματος. Στην περίπτωση αυτή η ανάλυση έγινε χρησιμοποιώντας τα αποτελέσματα και των 5 προγραμμάτων λόγω των παραπλήσιων αναλογιών τους όπως προαναφέρθηκαν στην ενότητα 3.1 και βασίστηκε μόνο στα αναγνωριστικά γονιδίων (AGI) εξαιτίας της διαφορετικής απόδοσης των συντεταγμένων από κάθε πρόγραμμα.

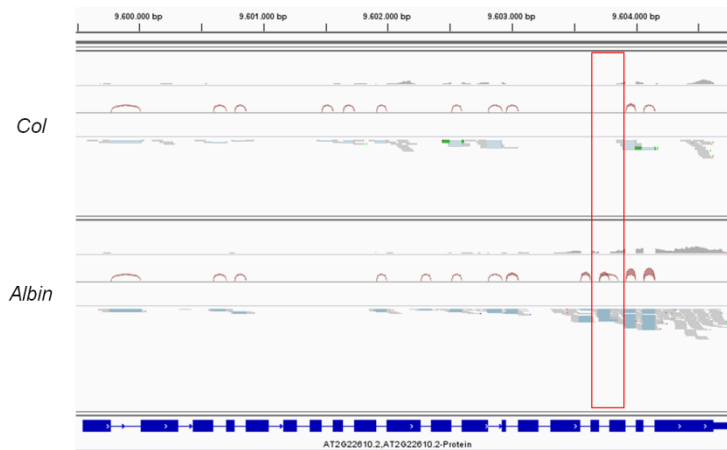
Στο διάγραμμα Venn της Εικόνας 3.5 παρατηρούνται 3 κοινά γονίδια και για τα 5 λογισμικά και επιπλέον 23 γονίδια μεταξύ των JuncBASE-ASprofile-ASATP. Η επιλογή των τριών αυτών προγραμμάτων έγινε με βάση τις παρόμοιες τιμές του λόγου λ, οι οποίες και στις 3 περιπτώσεις είναι περίπου 2. Ωστόσο, μετά από οπτικοποίηση των δεδομένων των αλληλουχίσεων για τα 26 αυτά γονίδια με το πρόγραμμα IGV, διαπιστώθηκε ότι μόνο στα 18 εμφανίζονται γεγονότα εναλλακτικής

3' θέσης ματίσματος. Η λίστα των 18 αυτών τελικών γονιδίων μαζί με τις τιμές του δείκτη PSI όπως υπολογίστηκαν από τα λογισμικά ASprofile και JuncBASE είναι καταχωρημένα στο Παράρτημα 3. Στην Εικόνα 3.6 παρουσιάζονται 3 από τα 18 κοινά γεγονότα εναλλακτικής 3' θέσης ματίσματος με τις υψηλότερες τιμές επιπέδων συμπερίληψης και τη μεγαλύτερη κάλυψη διαβασμάτων.



Εικόνα 3.5: Α. Σχηματική αναπαράσταση ενός γεγονότος εναλλακτικής 3' θέσης ματίσματος. Β. Απεικόνιση σε διάγραμμα Venn των αλληλεπικαλυπτόμενων γεγονότων εναλλακτικής 3' θέσης ματίσματος και των 5 προγραμμάτων με βάση μόνο τα αναγνωριστικά γονιδίων (AGI) ειδικά στο *lefko*.



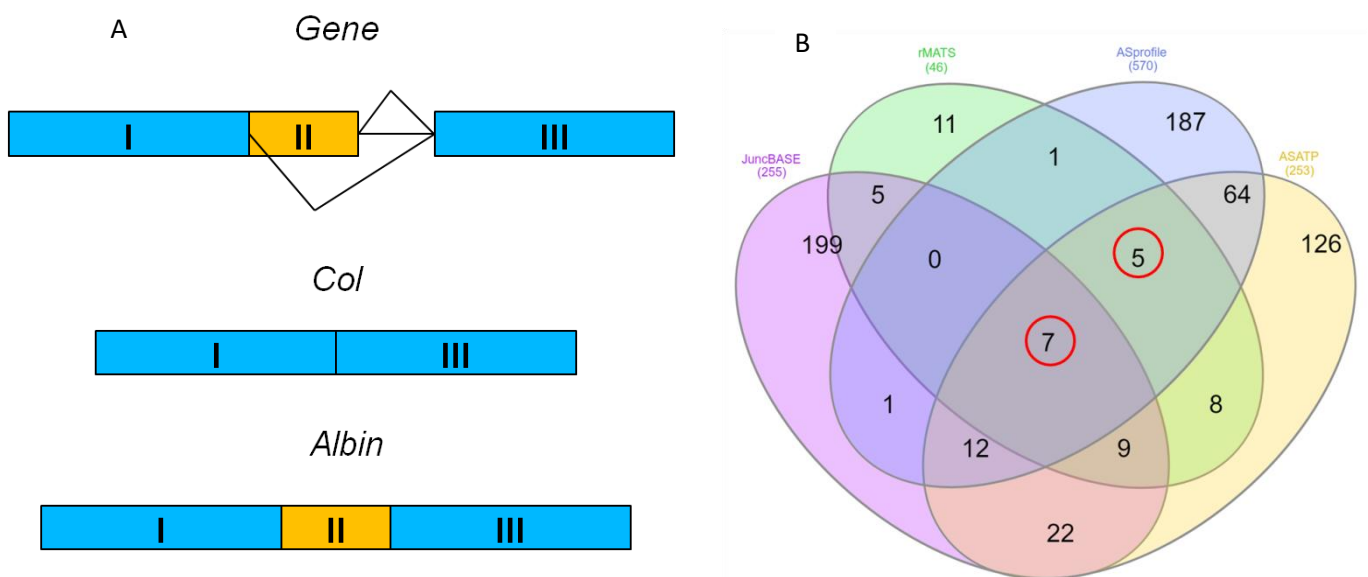


Εικόνα 3.6: Απεικόνιση των 3 σημαντικότερων κοινών γεγονότων εναλλακτικής 3' θέσης ματίσματος με το πρόγραμμα IGV (Integrative Genomics Viewer).

3.5 Κοινά γεγονότα εναλλακτικής 5' θέσης ματίσματος

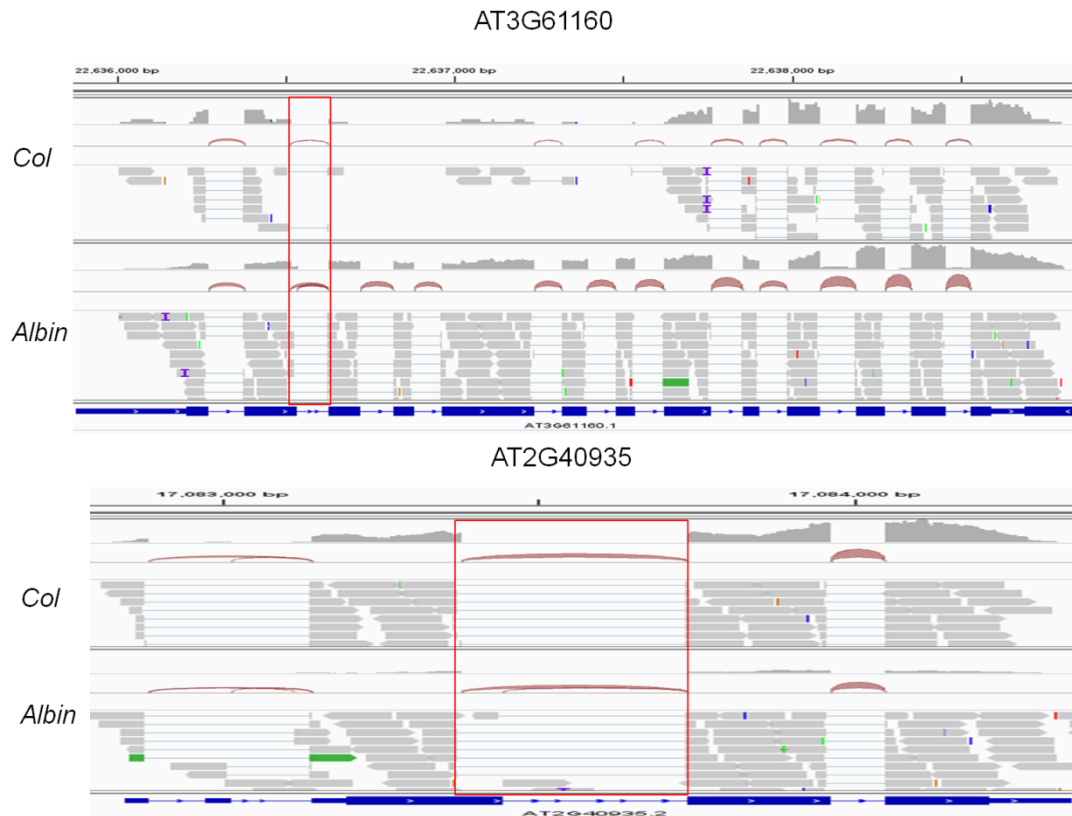
Για τον εντοπισμό των κοινών γεγονότων εναλλακτικής 5' θέσης ματίσματος πραγματοποιήθηκε η ίδια διαδικασία με τα γεγονότα εναλλακτικής 3' θέσης ματίσματος με εξαίρεση τη χρησιμοποίηση των δεδομένων από τα προγράμματα rMATS-JuncBASE-ASprofile-ASATP και όχι αυτών που παράχθηκαν από το SplAdder λαμβάνοντας υπόψη το κριτήριο των αναλογιών.

Όπως προκύπτει από την Εικόνα 3.7, τα αλληλεπικαλυπτόμενα γονίδια μεταξύ των 4 λογισμικών είναι 7, ενώ μεταξύ των rMATS-ASprofile-ASATP (μικρότερο εύρος τιμών του λ) είναι επιπλέον 5. Τα 12 αυτά γεγονότα εναλλακτικής 5' θέσης ματίσματος συνοψίζονται στον πίνακα του Παραρτήματος 3, στον οποίο δίνονται και οι τιμές των επιπέδων συμπερίληψης από το rMATS και το ASprofile.



Εικόνα 3.7: Α. Σχηματική αναπαράσταση ενός γεγονότος εναλλακτικής 5' θέσης ματίσματος. Β. Απεικόνιση σε διάγραμμα Venn των αλληλεπικαλυπτόμενων γεγονότων εναλλακτικής 5' θέσης ματίσματος των προγραμμάτων rMATS-JuncBASE-ASprofile-ASATP με βάση μόνο τα αναγνωριστικά γονιδίων (AGI) ειδικά στο *lefko*.

Η Εικόνα 3.8 αποδίδει γραφικά 2 παραδείγματα γεγονότων εναλλακτικής 5' θέσης ματίσματος με τη βοήθεια του λογισμικού IGV. Σημειώνεται ότι τα 2 συγκεκριμένα γεγονότα εμφανίζουν τις υψηλότερες τιμές PSI μεταξύ των 7 εντοπισμένων γεγονότων.

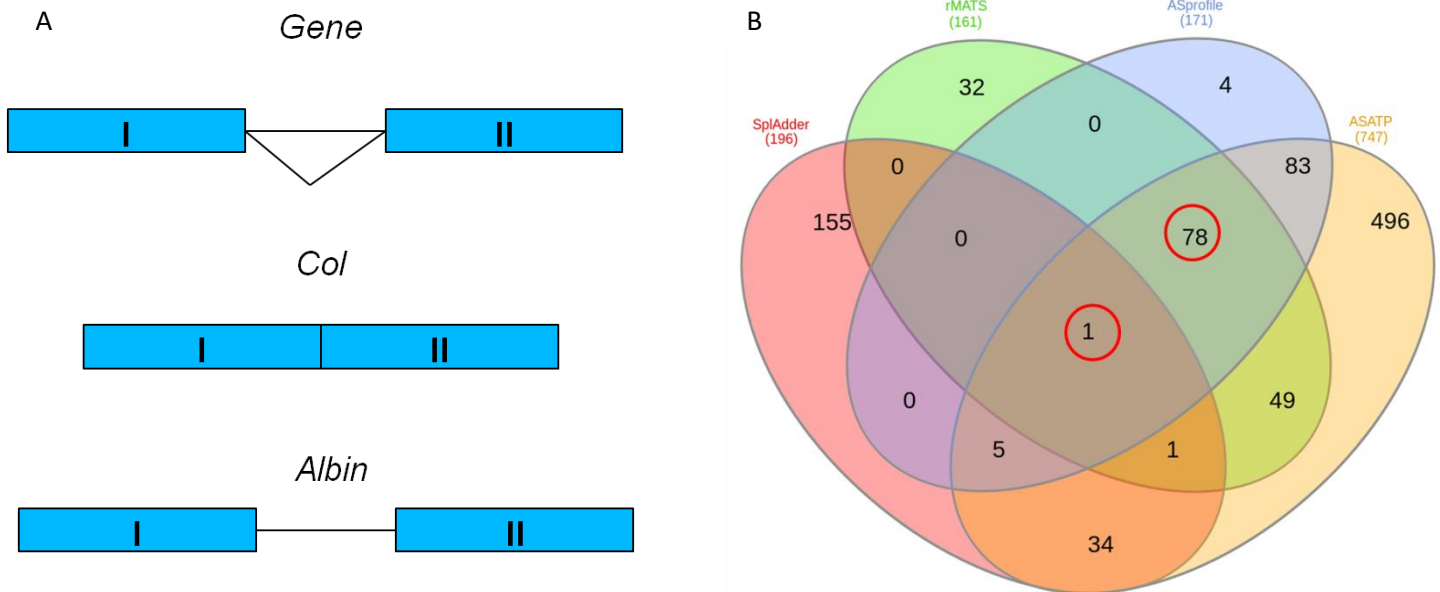


Εικόνα 3.8: Απεικόνιση των 2 σημαντικότερων κοινών γεγονότων εναλλακτικής 5' θέσης ματίσματος με το πρόγραμμα IGV (Integrative Genomics Viewer).

3.6 Κοινά γεγονότα διατήρησης ιντρονίου

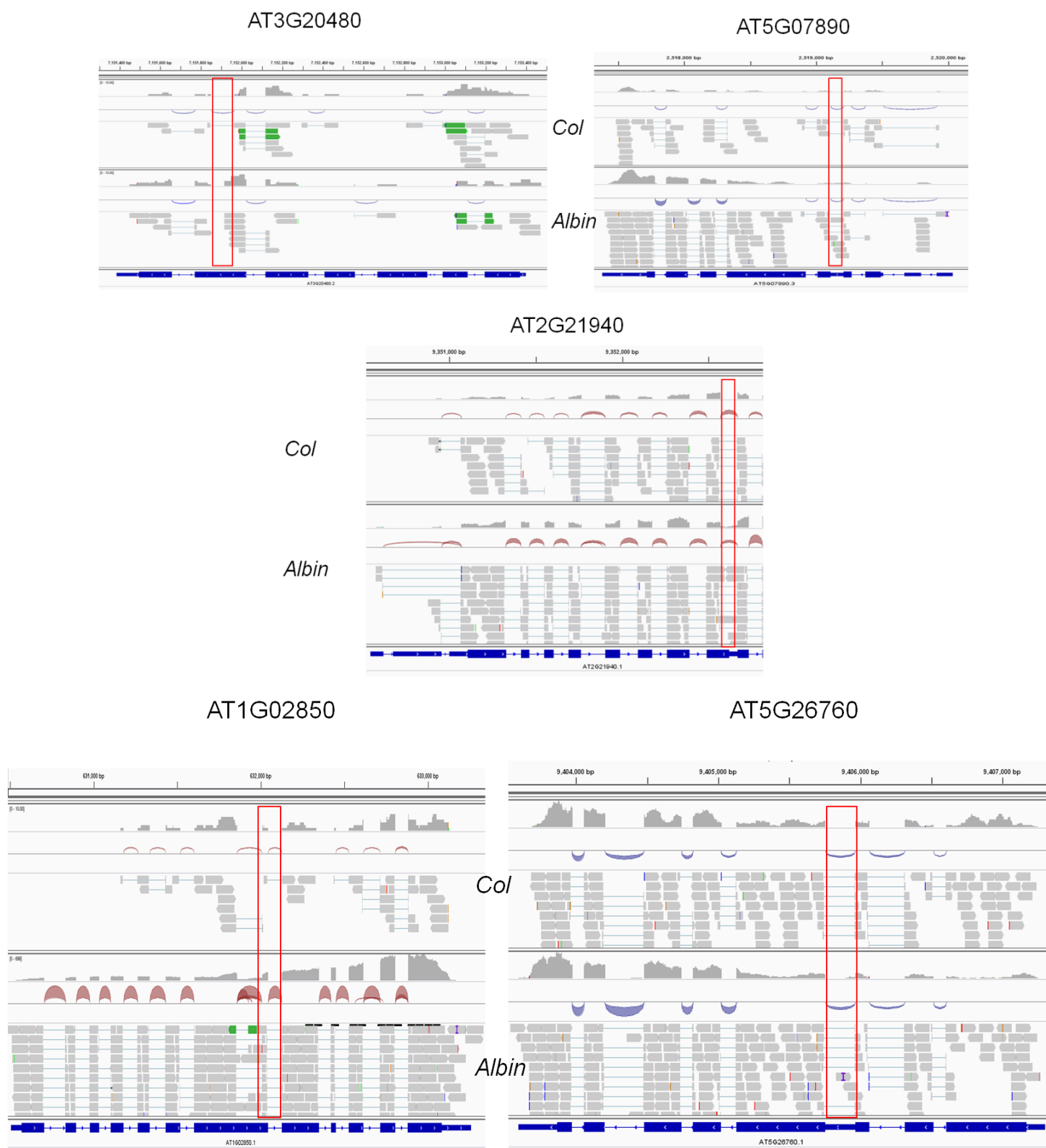
Στην τελευταία φάση της ανάλυσης επιχειρήθηκε η σύγκριση των αποτελεσμάτων των λογισμικών εντοπισμού γεγονότων εναλλακτικού ματίσματος για τα γεγονότα διατήρησης ιντρονίου ειδικά στο *lefko* κάνοντας χρήση τόσο των αναγνωριστικών γονιδίων (AGI) όσο και των συντεταγμένων (χρωμόσωμα-θέση έναρξης ιντρονίου-θέση λήξης ιντρονίου). Πιο συγκεκριμένα, με δεδομένο ότι το JuneBASE εντόπισε περίπου 34 φορές περισσότερα γεγονότα ειδικά στο μεταλλαγμένο στέλεχος σε σχέση με τα γεγονότα που εμφανίζονται ειδικά στο στέλεχος αγρίου τύπου, χρησιμοποιήθηκαν μόνο τα αποτελέσματα των 4 υπόλοιπων προγραμμάτων, στα οποία ο αντίστοιχος λόγος βρίσκεται περίπου στην τιμή 3.

Το διάγραμμα Venn (Εικόνα 3.9) αποκάλυψε την ύπαρξη ενός μόνο γεγονότος μεταξύ των SplAdder-rMATS-ASprofile-ASATP και επιπλέον 78 μεταξύ των rMATS-ASprofile-ASATP. Η λίστα των 79 αυτών εντοπισμένων γεγονότων διατήρησης ιντρονίου καθώς και οι αντίστοιχες τιμές του επιπέδου συμπερίληψης από τα προγράμματα rMATS και ASprofile περιέχονται στο Παράρτημα 3.



Εικόνα 3.9: Α. Σχηματική αναπαράσταση ενός γεγονότος διατήρησης ιντρονίου. Β. Απεικόνιση σε διάγραμμα Venn των αλληλεπικαλυπτόμενων γεγονότων διατήρησης ιντρονίου (IR) των προγραμμάτων SplAdder, rMATS, ASprofile και ASATP με βάση τα αναγνωριστικά γονιδίων (AGI) και τις συντεταγμένες των γεγονότων (χρωμόσωμα-έναρξη ιντρονίου-τέλος ιντρονίου) ειδικά στο *lefko*.

5 παραδείγματα απεικονίσεων των γεγονότων διατήρησης ιντρονίου πάνω στο γονιδίωμα του *A. thaliana* με το πρόγραμμα IGV παρουσιάζονται στην Εικόνα 3.10. Τα 5 συγκεκριμένα γεγονότα επιλέχθηκαν λόγω των υψηλών τιμών PSI και της ικανοποιητικής κάλυψης διαβασμάτων.



Εικόνα 3.10: Απεικόνιση 5 κοινών γεγονότων διατήρησης ιντρονίου με το πρόγραμμα IGV (Integrative Genomics Viewer).

Το επόμενο βήμα μετά την εύρεση και σύγκριση των γεγονότων εναλλακτικού ματίσματος στα γονίδια του *A. thaliana* απουσία της πρωτεΐνης LEFKOTHEA ήταν η πρόβλεψη των λειτουργιών, στις οποίες εμπλέκεται η πρωτεΐνη αυτή, μέσα από την ανάλυση εμπλουτισμού γονιδίων χρησιμοποιώντας τα εργαλεία DAVID. Ειδικότερα, δόθηκε έμφαση μόνο στα γονίδια που εμφάνισαν κάποιο γεγονός διατήρησης

ιντρονίου, καθώς, όπως προαναφέρθηκε, ο τύπος αυτός εναλλακτικού ματίσματος είναι ο συνηθέστερος στα φυτικά ευκαρυωτικά κύτταρα. Από τη λίστα των 79 γεγονότων που απομονώθηκαν με την παραπάνω διαδικασία, επιλέχθηκαν αυτά με τιμές PSI με βάση το rMATS μεγαλύτερες του 0.1. Μ' αυτόν τον τρόπο προέκυψαν 22 γονίδια στα οποία πραγματοποιήθηκε ανάλυση εμπλουτισμού.

Η αξιολόγηση και η ερμηνεία των αποτελεσμάτων της ανάλυσης εμπλουτισμού με τα εργαλεία DAVID βασίστηκε κυρίως σε 3 παραμέτρους: στην τιμή p (p -value), στην διορθωμένη τιμή p με τη στατιστική μέθοδο Benjamini-Hochberg (Benjamini and Hochberg, 1995) και στον δείκτη εμπλουτισμού (fold enrichment). Ειδικότερα, η τιμή p συνιστά την πιθανότητα η υπερεμφάνιση ενός παραγόμενου βιολογικού όρου στην εισαγόμενη λίστα γονιδίων να οφείλεται σε τυχαίο γεγονός και υπολογίζεται στη βάση δεδομένων DAVID με μία τροποποιημένη έκδοση του τεστ στατιστικής σημασίας του Fisher (EASE Score) (Huang *et al.*, 2009). Στην παρούσα εργασία η μέγιστη αποδεκτή τιμή p ορίστηκε στο 5% ή 0,05. Ωστόσο, επειδή κατά την ανάλυση εμπλουτισμού πραγματοποιούνται πολλαπλές συγκρίσεις λόγω του μεγάλου αριθμού των στατιστικών τεστ, κρίνεται αναγκαία η χρήση στατιστικής διόρθωσης των τιμών p μειώνοντας τον αριθμό των ψευδώς θετικών αποτελεσμάτων. Για την επίτευξη του παραπάνω στόχου χρησιμοποιήθηκε η μέθοδος των Benjamini-Hochberg. Τέλος, ο δείκτης εμπλουτισμού αποτελεί τον λόγο του αριθμού των γονιδίων της εισαγόμενης λίστας που εμφάνισαν έναν βιολογικό όρο προς τον αντίστοιχο αναμενόμενο αριθμό για όλα τα γνωστά γονίδια του οργανισμού ενδιαφέροντος.

Η ανάλυση εμπλουτισμού ανέδειξε τους όρους “εναλλακτικό μάτισμα”, “χλωροπλάστης” και “ισομορφή ματίσματος” ως στατιστικά σημαντικούς με βάση τα γονίδια της εισαγόμενης λίστας (Πίνακας 3.3). Πέρα από τις τιμές p και Benjamini-Hochberg, οι τιμές των δεικτών εμπλουτισμού των 3 συγκεκριμένων όρων είναι ιδιαίτερα αυξημένες ενισχύοντας τις προβλέψεις της ανάλυσης.

Πίνακας 3.3: Τα αποτελέσματα της ανάλυσης εμπλουτισμού με τα εργαλεία DAVID για τα 22 γονίδια, στα οποία εμφανίστηκαν τα σημαντικότερα γεγονότα διατήρησης ιντρονίου ειδικά στο δείγμα *lefko*.

Category	Term	Count	PValue	Benjamini	Fold Enrichment
UP_KEYWORDS	Alternative splicing	13	4.21E-07	2.31E-05	5.28
GOTERM_CC_DIRECT	GO:0009507~chloroplast	10	0.0016	0.045470961	3.11
UP_SEQ_FEATURE	splice variant	4	0.0049	0.080725545	8.07

4. Συμπεράσματα και Συζήτηση

Η διαδικασία της ωρίμανσης των πρώιμων μεταγράφων του πυρήνα αλλά και των μιτροχονδρίων και των χλωροπλάστων κρίνεται απαραίτητη για την ομαλή ανάπτυξη και την εύρυθμη λειτουργία των ευκαρυωτικών κυττάρων. Όπως είναι γνωστό, το μάτισμα των πρόδρομων mRNAs πραγματοποιείται με τη συνεργιστική δράση ριβονουκλεοπρωτεϊνών μαζί με ένα πλήθος πρωτεϊνών. Μία πρωτεΐνη που ανήκει στην παραπάνω κατηγορία και συναντάται στον οργανισμό *Arabidopsis thaliana* είναι η LEFKOTHEA. Με βάση πειραματικά δεδομένα και βιοπληροφορικές αναλύσεις, έχει αποδειχθεί ότι η LEFKOTHEA συμμετέχει ταυτόχρονα σε 2 απομακρυσμένους μηχανισμούς επεξεργασίας των πρώιμων μεταγράφων: στον μηχανισμό ματίσματος των ιντρονίων τύπου II στους χλωροπλάστες καθώς και σε έναν μηχανισμό ματίσματος μεταγράφων του γονιδιώματος του πυρήνα που σχετίζονται με χλωροπλαστικές λειτουργίες. Οι 2 αυτοί μηχανισμοί φαίνεται να μοιράζονται κοινά χαρακτηριστικά, γεγονός που ενδέχεται να οφείλεται στην εξελικτική τους σύνδεση. Κατά αυτόν τον τρόπο, η LEFKOTHEA εμφανίζει διπλή υποκυτταρική τοποθέτηση τόσο στον πυρήνα όσο και στον χλωροπλάστη συμβάλλοντας στην επικοινωνία μεταξύ των δύο αυτών κυτταρικών διαμερισμάτων (Δάρας, 2011).

Ένα από τα ερωτήματα που προέκυψαν από τα παραπάνω ευρήματα και αποτέλεσε το αντικείμενο έρευνας της μεταπτυχιακής αυτής εργασίας αφορά την πρόβλεψη των μεταγράφων-στόχων της πρωτεΐνης LEFKOTHEA στον πυρήνα και τον χλωροπλάστη, ώστε να διαλευκανθούν περαιτέρω ο ρόλος της στην ανάπτυξη του φυτού σε εμβρυϊκό στάδιο και οι κυτταρικές λειτουργίες, τις οποίες ελέγχει. Για την επίτευξη του σκοπού αυτού, έγινε χρήση των αποτελεσμάτων που παράγουν οι ομικές τεχνολογίες, και ειδικότερα η αλληλούχιση RNA. Πιο αναλυτικά, αλληλουχήθηκαν τα mRNAs κυττάρων του στελέχους *A. thaliana* που φέρει τη θνησιγόνο σε εμβρυϊκή ηλικία μετάλλαξη *lefko1* και κυττάρων του στελέχους αγρίου τύπου (*Col*) και, στη συνέχεια, απομονώθηκαν τα γεγονότα εναλλακτικού ματίσματος που βρέθηκαν ειδικά στο μεταλλαγμένο στέλεχος και ειδικά στο στέλεχος αγρίου τύπου αφαιρώντας τα κοινά με τη βοήθεια 5 λογισμικών εντοπισμού γεγονότων εναλλακτικού ματίσματος.

Συνολικά τα αποτελέσματα και των 5 προγραμμάτων ανέδειξαν την ύπαρξη πολλαπλάσιων γεγονότων εναλλακτικού ματίσματος ειδικά στο *lefko* σε σχέση με το στέλεχος αγρίου τύπου επιβεβαιώνοντας την καταλυτική δράση της LEFKOTHEA στη ομαλή διεκπεραίωση της διαδικασίας του ματίσματος. Με στόχο την αξιολόγηση της απόδοσης του κάθε προγράμματος ξεχωριστά για καθέναν από τους 4 κύριους τύπους εναλλακτικού ματίσματος (Πίνακας 4.1) υπολογίστηκαν και συγκρίθηκαν

μεταξύ τους οι αναλογίες των γεγονότων που εντοπίστηκαν ειδικά στο *Col* προς τα αντίστοιχα που εντοπίστηκαν ειδικά στο *lefko*. Με βάση, λοιπόν, την εφαρμογή του προαναφερθέντος κριτηρίου στα δεδομένα αλληλούχησης RNA της ερευνητικής αυτής εργασίας προτείνεται:

- Για τα γεγονότα διατήρησης ιντρονίου η χρήση των SplAdder-rMATS-ASprofile-ASATP
- Για τα γεγονότα παράλειψης εξωνίου η χρήση των rMATS-JuncBASE-ASprofile-ASATP
- Για τα γεγονότα 3' εναλλακτικής θέσης ματίσματος η χρήση και των 5 λογισμικών
- Για τα γεγονότα 5' εναλλακτικής θέσης ματίσματος η χρήση των rMATS-JuncBASE-ASprofile-ASATP

Πίνακας 4.1: Αξιολόγηση των 5 λογισμικών εντοπισμού γεγονότων εναλλακτικού ματίσματος με βάση την σύγκριση των αναλογιών των αριθμών των γεγονότων που εμφανίζονται ειδικά στο *Col* προς τους αντίστοιχους αριθμούς ειδικά στο *lefko*.

	SplAdder	rMATS	JuncBASE	Asprofile	ASATP
IR	✓	✓	×	✓	✓
ES	×	✓	✓	✓	✓
A3SS	✓	✓	✓	✓	✓
A5SS	×	✓	✓	✓	✓

Εξαιτίας της συχνότερης εμφάνισης γεγονότων διατήρησης ιντρονίου στους φυτικούς οργανισμούς, η έρευνά μας εστιάστηκε στον συγκεκριμένο τύπο εναλλακτικού ματίσματος. Ειδικότερα, απομονώθηκαν ειδικά στο μεταλλαγμένο στέλεχος 1 κοινό γεγονός μεταξύ και των 4 προγραμμάτων που χρησιμοποιήθηκαν και 78 μεταξύ των προγραμμάτων rMATS-ASprofile-ASATP. Από τα 79 συνολικά γονίδια επιλέχθηκαν τα σημαντικότερα 22 (επίπεδο συμπερίληψης > 0.1), στα οποία πραγματοποιήθηκε αναλύση εμπλουτισμού. Η προσέγγιση αυτή οδήγησε στην πρόβλεψη ότι οι όροι εναλλακτικό μάτισμα (ως λέξη-κλειδί στη βάση δεδομένων UniProt), χλωροπλάστης (ως όρος γονιδιακής οντολογίας) και ισομορφή ματίσματος

(ως χαρακτηριστικό της αλληλουχίας στη UniProt) υπερεμφανίζονται στην επιλεγμένη λίστα μας, γεγονός που συμβαδίζει απόλυτα με τα ήδη γνωστά δεδομένα για τη λειτουργία της πρωτεΐνης LEFKOTHEA. Τα γονίδια που συσχετίστηκαν με καθέναν από τους 3 βιολογικούς αυτούς όρους παρουσιάζονται στον Πίνακα 4.2.

Πίνακας 4.2: Τα γονίδια που βρέθηκε πως σχετίζονται με τους όρους εναλλακτικό μάτισμα, χλωροπλάστης και ισομορφή ματίσματος.

Alternative splicing	Chloroplast	splice variant
40s ribosomal protein SA B(RPSAb)	40s ribosomal protein SA B(RPSAb)	40s ribosomal protein SA B(RPSAb)
Iron-sulfur cluster biosynthesis family protein(AT5G03900)	Iron-sulfur cluster biosynthesis family protein(AT5G03900)	Iron-sulfur cluster biosynthesis family protein(AT5G03900)
RING/U-box superfamily protein(AT5G66070)	RNA-binding protein 47C(RBP47C)	beta glucosidase 11(BGLU11)
RNA polymerase II subunit B1 CTD phosphatase RPAP2-like protein(AT5G26760)	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(MENG)	shikimate kinase 1(SK1)
RNA-binding protein 47C(RBP47C)	antitermination NusB domain-containing protein(AT4G26370)	
S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(MENG)	beta-carotene isomerase D27-like protein(D27)	
beta glucosidase 11(BGLU11)	branched-chain amino acid transaminase 2(BCAT-2)	
beta-carotene isomerase D27-like protein(D27)	exocyst subunit exo70 family protein A1(EXO70A1)	
branched-chain amino acid transaminase 2(BCAT-2)	shikimate kinase 1(SK1)	
exocyst subunit exo70 family protein A1(EXO70A1)	solute carrier family 35 protein (DUF914)(AT3G59320)	
protein affected trafficking 2(PAT2)		
shikimate kinase 1(SK1)		
tetraacyldisaccharide 4'-kinase family protein(LpxK)		

Το τελευταίο εγχείρημα στα πλαίσια της μεταπτυχιακής αυτής διατριβής ήταν η προσπάθεια ανεύρεσης ενός χαρακτηριστικού μοτίβου αλληλουχιών στις θέσεις αποκοπής των ιντρονίων από τον μηχανισμό που συμμετέχει η πρωτεΐνη LEFKOTHEA. Γι' αυτό το σκοπό συγκεντρώθηκαν οι αλληλουχίες στις 5' και 3' θέσεις ματίσματος των αλληλεπικαλυπτόμενων γεγονότων διατήρησης ιντρονίου που εντοπίστηκαν από τα 4 λογισμικά στο δείγμα *lefko* και απεικονίστηκαν με τη μορφή λογότυπων (sequence logos). Επίσης, πραγματοποιήθηκε σύγκριση των παραπάνω αλληλουχιών με τις αντίστοιχες αλληλουχίες όλων των ιντρονίων τύπου U2 του γονιδιωμάτος του *A. thaliana*, οι οποίες προήλθαν από την βάση δεδομένων ERISdb (Szcześniak *et al.*, 2013).

U2 introns



130171

rMATS-
Asprofile-
ASATP



79

Εικόνα 4.1: Απεικόνιση με τη μορφή λογοτύπων των αλληλουχιών στις 5' και 3' θέσεις ματίσματος των 79 κοινών γεγονότων διατήρησης ιντρονίου μεταξύ των rMATS-Asprofile-ASATP και σύγκριση με τις αντίστοιχες θέσεις ματίσματος στα ιντρόνια τύπου U2 του γονιδιώματος του *A. thaliana*.

Η εμφανής σύγκλιση των αλληλουχιών στις 5' και 3' θέσεις ματίσματος των ιντρονίων που αποκόπτονται με τη δράση της LEFKOTHEA σε σχέση με τις αντίστοιχες αλληλουχίες των ιντρονίων τύπου U2 (Εικόνα 4.1), αποτελεί μία ένδειξη ότι η πρωτεΐνη αυτή πιθανότατα δεν αναγνωρίζει κάποιο μη κανονικό τύπο ιντρονίων. Συνεπώς, ο κανόνας GT-AG φαίνεται πως ισχύει και στην περίπτωση αυτή.

Συμπερασματικά, η παρούσα ερευνητική εργασία συνιστά ένα αντιπροσωπευτικό παράδειγμα χρήσης της Βιολογίας Συστημάτων για την απάντηση πολύπλοκων βιολογικών ερωτημάτων, όπως η πρόβλεψη των mRNA-στόχων της πρωτεΐνης LEFKOTHEA. Παρολ' αυτά, θα πρέπει να επισημανθεί ότι τα αποτελέσματα τέτοιου είδους μελετών βασίζονται κυρίως σε *in silico* προβλέψεις και γι' αυτό το λόγο για την καθολική αποδοχή τους κρίνεται αναγκαία η πειραματική τους επιβεβαίωση.

5. Βιβλιογραφία

- Δάρας Γ. (2011), Συστήματα μοριακών μηχανισμών ανάπτυξης στα φυτά, Διδακτορική Διατριβή, Εργαστήριο Μοριακής Βιολογίας – Γεωπονικό Πανεπιστήμιο Αθηνών
- Benjamini Y., Hochberg Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the royal statistical society Series B (Methodological)*: 289-300
- 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* **53**(8): 474-485
- Berget S. M., Moore C., Sharp P. A. (1977), Spliced segments at the 5' terminus of adenovirus 2 late mRNA, *Proc Natl Acad Sci U.S.A.* **74**(8): 3171-3175
- Boulikas T. (1994), Putative nuclear localization signals (NLS) in protein transcription factors, *J Cell Biochem* **55**(1): 32-58
- Brooks A. N., Yang L., Duff M. O., Hansen K. D., Park J. W., Dudoit S., Brenner S.E., Graveley B.R. (2011), Conservation of an RNA regulatory map between Drosophila and mammals, *Genome Res* **21**(2): 193-202
- Carvalho R. F., Feijão C. V., Duque P. (2013), On the physiological significance of alternative splicing events in higher plants, *Protoplasma* **250**(3): 639-650
- Chow L. T., Gelinas R. E., Broker T. R., Roberts R. J. (1977), An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA, *Cell* **12**(1): 1-8
- Christopher D. A. and Hallick R. B. (1989), Euglena gracilis chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III, *Nucleic Acids Res* **17**(19): 7591-7608
- Clancy S. (2008), RNA splicing: introns, exons and spliceosome, *Nature Education* **1**(1):31
- Cock P. J. A., Fields C. J., Goto N., Heuer M. L., Rice P. M. (2010), The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, *Nucleic Acids Res* **38**(6): 1767-1771
- Crooks G. E., Hon G., Chandonia J. M., Brenner S. E. (2004), WebLogo: A sequence logo generator, *Genome Res* **14**(6):1188-1190
- Darnell J. E. Jr. (1978), Implications of RNA-RNA splicing in evolution of eukaryotic cells, *Science* **202**(4374): 1257-60
- de Longevialle A. F., Small I. D., Lurin C. (2010), Nuclearly encoded splicing factors implicated in RNA splicing in higher plant organelles, *Mol Plant* **3**(4):691-705
- Early P., Rogers J., Davis M., Calame K., Bond M., Wall R., Hood L. (1980), Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways, *Cell* **20**(2): 313-319
- Florea L., Song L., Salzberg S. L. (2013), Thousands of exon skipping events differentiate among splicing patterns in sixteen human tissues, *F1000Res* **2**: 188

- Francs-Small C. C., Kroeger T., Zmudjak M., Ostersetzer-Biran O., Rahimi N., Small I., Barkan A. (2012), A PORR domain protein required for rpl2 and ccmF(C) intron splicing and for the biogenesis of c-type cytochromes in Arabidopsis mitochondria, *Plant J* **69**(6): 996-1005
- Heberle H., Meirelles G. V., da Silva F. R., Telles G. P., Minghim R. (2015) InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams, *BMC Bioinformatics* **16**: 169
- Huang da W., Sherman B. T., Lempicki R. A. (2009), Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat Protoc* **4**(1): 44-57
- Huang da W., Sherman B. T., Lempicki R. A. (2009), Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res* **37**(1): 1-13
- Kahles A., Ong C.S., Zhong Y., Räscht G. (2016), SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data, *Bioinformatics* **32**(12): 1840-1847
- Kanehisa M., Goto S. (2000), KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**(1): 27-30
- Keren H., Lev-Maor G., Ast G. (2010), Alternative splicing and evolution: diversification, exon definition and function, *Nature Reviews Genetics* **11**: 345-355
- Kim E., Magen A., Ast G. (2007), Different levels of alternative splicing among eukaryotes, *Nucleic Acids Res* **35**(1): 125-131
- Kosugi S., Hasebe M., Matsumura N., Takashima H., Miyamoto-Sato E., Tomita M., Yanagawa H. (2009), Six classes of nuclear localization signals specific to different binding grooves of importin alpha, *J Biol Chem* **284**(1): 478-485
- Krause K., Krupinska K. (2009), Nuclear regulators with a second home in organelles, *Trends Plant Sci.* **14**(4): 194-199
- Kroeger T. S., Watkins K. P., Friso G., van Wijk K. J., Barkan A. (2009), A plant-specific RNA-binding domain revealed through analysis of chloroplast group II intron splicing, *Proc Natl Acad Sci U S A* **106**(11): 4537-4542
- la Cour T., Kierner L., Mølgaard A., Gupta R., Skriver K., Brunak S. (2004), Analysis and prediction of leucine-rich nuclear export signals, *Protein Eng Des Sel* **17**(6): 527-536
- Langmead B., Trapnell C., Pop M., Salzberg S. L. (2009), Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol* **10**(3):R25
- Li H. *et al.* (2009), The Sequence Alignment/Map format and SAMtools, *Bioinformatics* **25**(16): 2078-2079
- Mackenzie S. A. (2005), Plant organellar protein targeting: a traffic plan still under construction, *Trends Cell Biol* **15**(10): 548-554

- Natsoulis G., Hilger F., Fink G.R. (1986), The HTS1 gene encodes both the cytoplasmic and mitochondrial histidine tRNA synthetases of *S. cerevisiae*, *Cell* **46**(2): 235-243
- Nishimura D. (2004), BioCarta, *Biotech Software & Internet Report* **2**(3): 117-120
- Patel A.A., Steitz J.A. (2003), Splicing double: Insights from the second spliceosome, *Nat Rev Mol Cell Biol* **4**: 960–970
- Peeters N., Small I. (2001), Dual targeting to mitochondria and chloroplasts, *Biochim Biophys Acta* **1541**(1-2): 54-63
- Quinlan A. R., Hall I. M. (2010), BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics* **26**(6): 841-842
- Robinson J. T. *et al.* (2011), Integrative Genomics Viewer, *Nature Biotechnology* **29**: 24-26
- Schafer S., Miao K., Benson C. C., Heinig M., Cook S. A. and Hubner N. (2015), Alternative splicing signatures in RNA-seq data: percent spliced in (PSI), *Curr Protoc Hum Genet* **87**: 11.16.1-11.16.14
- Schmitz-Linneweber C., Barkan A. (2007), RNA splicing and RNA editing in chloroplasts, *Topics in Current Genetics* **19**: 213-248
- Schneider T. D., Stephens R. M. (1990), Sequence Logos: A New Way to Display Consensus Sequences, *Nucleic Acids Res* **18**(20):6097-6100
- Shen S., Park J.W., Lu Z.X., Lin L., Henry M.D., Wu Y.N., Zhou Q., Xing Y. (2014), rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data, *Proc Natl Acad Sci USA* **111**(51): E5593-5601
- Szcześniak M. W., Kabza M., Pokrzywa R., Gudyś A., Makołowska I. (2013), ERISdb: a database of plant splice sites and splicing signals, *Plant Cell Physiol* **54**(2): e10
- Thorvaldsdottir H. *et al.* (2013), Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Briefings in Bioinformatics* **14**: 178-192
- Trapnell *et al.* (2010), Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, *Nat Biotechnology* **28**: 511-515
- Trapnell C., Lior Pachter L., Salzberg S. L. (2009), TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics* **25**(9): 1105-1111
- Tzafrir I., Pena-Muralla R., Dickerman A., Berg M., Rogers R., Hutchens S., Sweeney T. C., McElver J., Aux G., Patton D., Meinke D. (2004), Identification of genes required for embryo development in Arabidopsis, *Plant Physiol* **135**(3): 1206-1220
- Valadkhan S. (2007), The spliceosome: a ribozyme at heart?, *Biol Chem* **388**(7): 693-697
- Ziding Zhang's Laboratory for Protein Bioinformatics, College of Biological Sciences, China Agricultural University (2016), Alternative Splicing Analysis Tool Package (ASATP), <https://github.com/Z-G-L/ASATP>

6. Παράρτημα 1

Ακολουθία εντολών σε τερματικό linux για την ανάλυση με το πρόγραμμα SplAdder

```
python ../spladder/python/spladder.py -a ../reference/TAIR10_GFF3_genes.gff -b A
```

```
lbin.sort.bam -o Albin -T y > log_albin
```

```
python ../spladder/python/spladder.py -a ../reference/TAIR10_GFF3_genes.gff -b C
```

```
ol.sort.bam -o Col -T y > log_col
```

#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Albin

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Albin/merge_graphs_alt_3prime_C3.confirmed.txt  
Col/merge_graphs_alt_3prime_C3.confirmed.txt > alt_3prime_C3.confirmed_only_Albin.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' alt_3prime_C3.confirmed_only_Albin.txt >  
alt_3prime_C3.confirmed_only_Albin-gene-Chr-start-end.txt
```

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Albin/merge_graphs_alt_5prime_C3.confirmed.txt  
Col/merge_graphs_alt_5prime_C3.confirmed.txt > alt_5prime_C3.confirmed_only_Albin.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' alt_5prime_C3.confirmed_only_Albin.txt >  
alt_5prime_C3.confirmed_only_Albin-gene-Chr-start-end.txt
```

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Albin/merge_graphs_intron_retention_C3.confirmed.txt  
Col/merge_graphs_intron_retention_C3.confirmed.txt > intron_retention_confirmed_only_albin.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' intron_retention_confirmed_only_albin.txt >  
intron_retention_confirmed_only_albin-genes-Chr-start-end.txt
```

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Albin/ merge_graphs_exon_skip_C3.confirmed.txt Col/  
merge_graphs_exon_skip_C3.confirmed.txt > exon_skip_confirmed_only_Albin.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' exon_skip_confirmed_only_Albin.txt > exon_skip_confirmed_only_Albin-gene-  
Chr-start-end.txt
```

#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Col

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Col/merge_graphs_alt_3prime_C3.confirmed.txt  
Albin/merge_graphs_alt_3prime_C3.confirmed.txt > alt_3prime_C3.confirmed_only_Col.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' alt_3prime_C3.confirmed_only_Col.txt > alt_3prime_C3.confirmed_only_Col-  
gene-Chr-start-end.txt
```

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Col/merge_graphs_alt_5prime_C3.confirmed.txt  
Albin/merge_graphs_alt_5prime_C3.confirmed.txt > alt_5prime_C3.confirmed_only_Col.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' alt_5prime_C3.confirmed_only_Col.txt > alt_5prime_C3.confirmed_only_Col-  
gene-Chr-start-end.txt
```

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Col/merge_graphs_intron_retention_C3.confirmed.txt  
Albin/merge_graphs_intron_retention_C3.confirmed.txt > intron_retention_confirmed_only_Col.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' intron_retention_confirmed_only_Col.txt >  
intron_retention_confirmed_only_Col-genes-Chr-start-end.txt
```

```
perl ~/thanasis/lefkothea_bam/spladder/comparison.pl Col/ merge_graphs_exon_skip_C3.confirmed.txt Albin/  
merge_graphs_exon_skip_C3.confirmed.txt > exon_skip_confirmed_only_Col.txt
```

```
awk '{print $4"\t"$1"\t"$7"\t"$8}' exon_skip_confirmed_only_Col.txt > exon_skip_confirmed_only_Col-gene-Chr-start-end.txt
```

Ακολουθία εντολών σε τερματικό linux για την ανάλυση με το πρόγραμμα rMATS

```
python ../../mats/rMATS.3.2.5/RNASeq-MATS.py -b1 ../Albin.sort.bam -b2 ../Col.sort.bam -gtf ../../reference/TAIR10_GFF3_genes.gtf -t paired -len 100 -o output
```

#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Albin

```
awk '{if ($22 == 0) print}' A3SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_A3SS.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{if ($22 == 0) print}' A5SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_A5SS.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{if ($22 == 0) print}' RI.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_RI.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{if ($22 == 1) print}' SE.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_SE.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{print $2"\t"$4"\t"$6+1"\t"$7}' only_Albin_A3SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_A3SS-gene-chr-start-end.txt
```

```
awk '{print $2"\t"$4"\t"$6+1"\t"$7}' only_Albin_A5SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_A5SS-gene-chr-start-end.txt
```

```
awk '{print $2"\t"$4"\t"$9+1"\t"$10}' only_Albin_RI.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_RI-gene-chr-start-end.txt
```

```
awk '{print $2"\t"$4"\t"$6+1"\t"$7}' only_Albin_SE.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Albin_SE-gene-chr-start-end.txt
```

#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Col

```
awk '{if ($21 == 0) print}' A3SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_A3SS.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{if ($21 == 0) print}' A5SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_A5SS.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{if ($21 == 0) print}' RI.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_RI.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{if ($21 == 1) print}' SE.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_SE.MATS.ReadsOnTargetAndJunctionCounts.txt
```

```
awk '{print $2"\t"$4"\t"$6+1"\t"$7}' only_Col_A3SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_A3SS-gene-chr-start-end.txt
```

```
awk '{print $2"\t"$4"\t"$6+1"\t"$7}' only_Col_A5SS.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_A5SS-gene-chr-start-end.txt
```

```
awk '{print $2"\t"$4"\t"$9+1"\t"$10}' only_Col_RI.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_RI-gene-chr-start-end.txt
```

```
awk '{print $2"\t"$4"\t"$6+1"\t"$7}' only_Col_SE.MATS.ReadsOnTargetAndJunctionCounts.txt > only_Col_SE-gene-chr-start-end.txt
```

Ακολουθία εντολών σε τερματικό linux για την ανάλυση με το πρόγραμμα JuncBASE

#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Albin και ειδικά στο Col

```
python ../../juncbase/juncBASE/build_DB_FromGTF.py -g ../../reference/TAIR10_GFF3_genes1.gtf -d db1 --
initialize

python ../../juncbase/juncBASE/build_DB_FromGTF.py -g ../../reference/TAIR10_GFF3_genes1.gtf -d db1

sudo python ~/juncbase/juncBASE/preProcess_getASEventReadCounts.py -s ../Albin.sort.bam -n Albin -o input

sudo python ~/juncbase/juncBASE/preProcess_getASEventReadCounts.py -s ../Col.sort.bam -n Col -o input

sudo python ~/juncbase/juncBASE/preProcess_getASEventReadCounts_by_chr_step2.py -i input/

sudo python ~/juncbase/juncBASE/preProcess_getASEventReadCounts_step3.py -i input/ -n Albin -t
input/tmp_preProcess_getASEventReadCounts_step2.bed --min_overhang 6

sudo python ~/juncbase/juncBASE/preProcess_getASEventReadCounts_step3.py -i input -n Col -t
input/tmp_preProcess_getASEventReadCounts_step2.bed -l 100 --min_overhang 6

sudo python ~/juncbase/juncBASE/createPseudoSample.py -i input -s Albin

python ~/juncBASE/run_getASEventReadCounts_multiSample.py -s Albin,Col -i input/ -o results --txt_db1 db1 --
txt_db2 db1 --txt_db3 db1 --jcn_seq_len 188 -p 1 --sqlite_db_dir database/

python ~/juncBASE/createAS_CountTables.py -d results/ -o results_table --jcn_seq_len 188 -s Albin,Col

perl ~/thanasis/perl_scripts/chrChr_to_Chrr.pl results_table_AS_exclusion_inclusion_counts.txt >
results_table_AS_exclusion_inclusion_counts_Chrr.txt

grep "intron_retention" results_table_AS_exclusion_inclusion_counts_Chrr.txt > RI.txt

grep "acceptor" results_table_AS_exclusion_inclusion_counts_Chrr.txt > A3SS.txt

grep "donor" results_table_AS_exclusion_inclusion_counts_Chrr.txt > A5SS.txt

grep "cassette" results_table_AS_exclusion_inclusion_counts_Chrr.txt > SE.txt

cat RI.csv | tr -s ';' '\t' | awk '{if ($9==0) print}' > only_Albin_RI.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_RI.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf only_Albin_RI.txt
> only_Albin_RI_annotated-gene-Chr-start-end.txt

cat RI.csv | tr -s ';' '\t' | awk '{if ($7==0) print}' > only_Col_RI.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_RI.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf only_Col_RI.txt >
only_Col_RI_annotated-gene-Chr-start-end.txt

cat SE.csv | tr -s ';' '\t' | awk '{if ($8==0) print}' > only_Albin_SE.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_for_SE.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf
only_Albin_SE.txt > only_Albin_SE_annotated-gene-Chr-start-end.txt

cat SE.csv | tr -s ';' '\t' | awk '{if ($6==0) print}' > only_Col_SE.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_for_SE.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf
only_Col_SE.txt > only_Col_SE_annotated-gene-Chr-start-end.txt

cat A3SS.csv | tr -s ';' '\t' | awk '{if ($9==0) print}' > only_Albin_A3SS.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_RI.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf
only_Albin_A3SS.txt > only_Albin_A3SS_annotated-gene-Chr-start-end.txt
```

```
cat A3SS.csv | tr -s ';' '\t' | awk '{if ($7==0) print}' > only_Col_A3SS.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_RI.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf
only_Col_A3SS.txt > only_Col_A3SS_annotated-gene-Chr-start-end.txt

cat A5SS.csv | tr -s ';' '\t' | awk '{if ($9==0) print}' > only_Albin_A5SS.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_A5SS.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf
only_Albin_A5SS.txt > only_Albin_A5SS_annotated-gene-Chr-start-end.txt

cat A5SS.csv | tr -s ';' '\t' | awk '{if ($7==0) print}' > only_Col_A5SS.txt

perl ~/thanasis/perl_scripts/AGI_from_gtf_A5SS.pl ~/thanasis/reference/TAIR10_GFF3_genes.gtf
only_Col_A5SS.txt > only_Col_A5SS_annotated-gene-Chr-start-end.txt
```

Ακολουθία εντολών σε τερματικό linux για την ανάλυση με το πρόγραμμα ASprofile

```
#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Albin και ειδικά στο Col

cufflinks Albin.sort.bam -o Albin_cufflinks/ -G ~/thanasis/reference/TAIR10_GFF3_genes.gtf

awk '{ if ($6 > 1) print}' transcripts.gtf > transcripts_greater_than_1.gtf

cufflinks Col.sort.bam -o Col_cufflinks/ -G ~/thanasis/reference/TAIR10_GFF3_genes.gtf

awk '{ if ($6 > 1) print}' transcripts.gtf > transcripts_greater_than_1.gtf

~/thanasis/class2/ASprofile.b-1.0.4/extract-as ../Albin_cufflinks/transcripts_greater_than_1.gtf
~/thanasis/reference/TAIR10_chr_all.fasta.headers > Albin_cufflinks_greater_than_1_AS-events.txt

~/thanasis/class2/ASprofile.b-1.0.4/extract-as ../Col_cufflinks/transcripts_greater_than_1.gtf
~/thanasis/reference/TAIR10_chr_all.fasta.headers > Col_cufflinks_greater_than_1_AS-events.txt

awk '{print $2"\t"$3"\t"$4"\t"$5"\t"$6}' Albin_cufflinks_greater_than_1_AS-events.txt > Albin_cufflinks_AS-
events-type-gene-Chr-start-end.txt

awk '{print $2"\t"$3"\t"$4"\t"$5"\t"$6}' Col_cufflinks_greater_than_1_AS-events.txt > Col_cufflinks_AS-events-
type-gene-Chr-start-end.txt

perl ~/thanasis/perl_scripts/compare.pl Albin_cufflinks_AS-events-type-gene-Chr-start-end.txt Col_cufflinks_AS-
events-type-gene-Chr-start-end.txt > only_Albin_cufflinks_AS-events-type-gene-Chr-start-end.txt

grep "IR" only_Albin_cufflinks_AS-events-type-gene-Chr-start-end.txt | awk '{print $2"\t"$3"\t"$4+1"\t"$5-1}' |
sort | uniq > only_Albin_cufflinks_IR-gene-Chr-start-end.txt

grep "AE" only_Albin_cufflinks_AS-events-type-gene-Chr-start-end.txt | awk '{print $2"\t"$3"\t"$4"\t"$5}' | sort |
uniq > only_Albin_cufflinks_AE-gene-Chr-start-end.txt

grep "SKIP" only_Albin_cufflinks_AS-events-type-gene-Chr-start-end.txt | awk '{print $2"\t"$3"\t"$4"\t"$5}' | sort
| uniq > only_Albin_cufflinks_SE-gene-Chr-start-end.txt

perl ~/thanasis/perl_scripts/compare.pl Col_cufflinks_AS-events-type-gene-Chr-start-end.txt Albin_cufflinks_AS-
events-type-gene-Chr-start-end.txt > only_Col_cufflinks_AS-events-type-gene-Chr-start-end.txt

grep "IR" only_Col_cufflinks_AS-events-type-gene-Chr-start-end.txt | awk '{print $2"\t"$3"\t"$4+1"\t"$5-1}' | sort |
uniq > only_Col_cufflinks_IR-gene-Chr-start-end.txt

grep "AE" only_Col_cufflinks_AS-events-type-gene-Chr-start-end.txt | awk '{print $2"\t"$3"\t"$4"\t"$5}' | sort |
uniq > only_Col_cufflinks_AE-gene-Chr-start-end.txt

grep "SKIP" only_Col_cufflinks_AS-events-type-gene-Chr-start-end.txt | awk '{print $2"\t"$3"\t"$4"\t"$5}' | sort |
uniq > only_Col_cufflinks_SE-gene-Chr-start-end.txt
```

Ακολουθία εντολών σε τερματικό linux για την ανάλυση με το πρόγραμμα ASATP

#Απομόνωση γεγονότων εναλλακτικού ματίσματος που εμφανίζονται ειδικά στο Albin και ειδικά στο Col

```
perl ~/thanasis/perl_scripts/compare.pl Albin-asatp/ASRecovist_out/AS_event.txt Col-  
asatp/ASRecovist_out/AS_event.txt > only_Albin_AS_event.txt
```

```
grep "A5SS" only_Albin_AS_event.txt > only_Albin_A5SS.txt
```

```
grep "A3SS" only_Albin_AS_event.txt > only_Albin_A3SS.txt
```

```
grep "IR" only_Albin_AS_event.txt > only_Albin_IR.txt
```

```
grep "CE" only_Albin_AS_event.txt > only_Albin_SE.txt
```

```
awk '{print $1"\t"$2"\t"$8}' only_Albin_IR.txt | awk -F'\t' '{print $1"\t"$2"\t"$3"\t"$4}' | sort | uniq >  
only_Albin_IR-gene-Chr-start-end.txt
```

```
awk '{print $1"\t"$2"\t"$9}' only_Albin_A3SS.txt | tr -s ',' '\t' | tr -s '-' '\t' | awk '{print $1"\t"$2"\t"$7"\t"$10}' | sort  
| uniq > only_Albin_A3SS-gene-Chr-start-end.txt
```

```
awk '{print $1"\t"$2"\t"$9}' only_Albin_A5SS.txt | tr -s ',' '\t' | tr -s '-' '\t' | awk '{print $1"\t"$2"\t"$3"\t"$6}' | sort |  
uniq > only_Albin_A5SS-gene-Chr-start-end.txt
```

```
awk '{print $1"\t"$2"\t"$9}' only_Albin_SE.txt | tr -s ',' '\t' | tr -s '-' '\t' | awk '{print $1"\t"$2"\t"$7"\t"$8}' | sort |  
uniq > only_Albin_SE-gene-Chr-start-end.txt
```

```
perl ~/thanasis/perl_scripts/compare.pl Col-asatp/ASRecovist_out/AS_event.txt Albin-  
asatp/ASRecovist_out/AS_event.txt > only_Col_AS_event.txt
```

```
grep "A5SS" only_Col_AS_event.txt > only_Col_A5SS.txt
```

```
grep "A3SS" only_Col_AS_event.txt > only_Col_A3SS.txt
```

```
grep "CE" only_Col_AS_event.txt > only_Col_SE.txt
```

```
grep "IR" only_Col_AS_event.txt > only_Col_IR.txt
```

```
awk '{print $1"\t"$2"\t"$8}' only_Col_IR.txt | awk -F'\t' '{print $1"\t"$2"\t"$3"\t"$4}' | sort | uniq > only_Col_IR-  
gene-Chr-start-end.txt
```

```
awk '{print $1"\t"$2"\t"$9}' only_Col_A3SS.txt | tr -s ',' '\t' | tr -s '-' '\t' | awk '{print $1"\t"$2"\t"$7"\t"$10}' | sort |  
uniq > only_Col_A3SS-gene-Chr-start-end.txt
```

```
awk '{print $1"\t"$2"\t"$9}' only_Col_A5SS.txt | tr -s ',' '\t' | tr -s '-' '\t' | awk '{print $1"\t"$2"\t"$3"\t"$6}' | sort |  
uniq > only_Col_A5SS-gene-Chr-start-end.txt
```

```
awk '{print $1"\t"$2"\t"$9}' only_Col_SE.txt | tr -s ',' '\t' | tr -s '-' '\t' | awk '{print $1"\t"$2"\t"$7"\t"$8}' | sort | uniq  
> only_Col_SE-gene-Chr-start-end.txt
```

Δημιουργία weblogos για τις 5' και 3' θέσεις ματίσματος των 79 γεγονότων διατήρησης ιντρονίου

```
awk '{print $2"\t"$4"\t"$5"\t"$9+1"\t"$10}' ../only_Albin_RI.MATS.ReadsOnTargetAndJunctionCounts.txt >  
only_Albin_RI-gene-chr-strand-start-end.txt
```

```
perl ~/thanasis/perl_scripts/find_strand_79_overlap.pl 79_overlap-RI-rmats-asprofile-asatp.txt only_Albin_RI-  
gene-Chr-strand-start-end > 79_overlap-RI-rmats-asprofile-asatp-with-strand
```

```
awk '{if ($3 == "+") print $2"\t"$4-5"\t"$4+5}' 79_overlap-RI-rmats-asprofile-asatp-with-strand > only_Albin_RI-  
positive-strand-Chr-start-minus-plus-5.txt
```

```
bedtools getfasta -fi ~/thanasis/reference/TAIR10_chr_for_galaxy_all.fasta -bed only_Albin_RI-positive-strand-  
Chr-start-minus-plus-5.txt -fo only_Albin_RI-positive-strand-Chr-start-minus-plus-5.fasta
```

```

awk 'if ($3 == "-") print $2"\t"$4-6"\t"$4+4}' 79_overlap-RI-rmats-asprofile-asatp-with-strand > only_Albin_RI-negative-strand-Chr-start-minus-plus-5.txt

bedtools getfasta -fi ~/thanasis/reference/TAIR10_chr_for_galaxy_all.fasta -bed only_Albin_RI-negative-strand-Chr-start-minus-plus-5.txt -fo only_Albin_RI-negative-strand-Chr-start-minus-plus-5.fasta

perl ~/thanasis/perl_scripts/reverse_negative_strand.pl only_Albin_RI-negative-strand-Chr-start-minus-plus-5.fasta > only_Albin_RI-negative-strand-Chr-end-minus-plus-5-reverse.fasta

awk 'if ($3 == "+") print $2"\t"$5-5"\t"$5+5}' 79_overlap-RI-rmats-asprofile-asatp-with-strand > only_Albin_RI-positive-strand-Chr-end-minus-plus-5.txt

bedtools getfasta -fi ~/thanasis/reference/TAIR10_chr_for_galaxy_all.fasta -bed only_Albin_RI-positive-strand-Chr-end-minus-plus-5.txt -fo only_Albin_RI-positive-strand-Chr-end-minus-plus-5.fasta

awk 'if ($3 == "-") print $2"\t"$5-6"\t"$5+4}' only79_overlap-RI-rmats-asprofile-asatp-with-strand > only_Albin_RI-negative-strand-Chr-end-minus-plus-5.txt

bedtools getfasta -fi ~/thanasis/reference/TAIR10_chr_for_galaxy_all.fasta -bed only_Albin_RI-negative-strand-Chr-end-minus-plus-5.txt -fo only_Albin_RI-negative-strand-Chr-end-minus-plus-5.fasta

perl ~/thanasis/perl_scripts/reverse_negative_strand.pl only_Albin_RI-negative-strand-Chr-end-minus-plus-5.fasta > only_Albin_RI-negative-strand-Chr-start-minus-plus-5-reverse.fasta

cat only_Albin_RI-positive-strand-Chr-start-minus-plus-5.fasta only_Albin_RI-negative-strand-Chr-start-minus-plus-5-reverse.fasta > only_Albin_RI-positive-and-negative-reverse-strand-Chr-start-minus-plus-5.fasta

cat only_Albin_RI-positive-strand-Chr-end-minus-plus-5.fasta only_Albin_RI-negative-strand-Chr-end-minus-plus-5-reverse.fasta > only_Albin_RI-positive-and-negative-reverse-strand-Chr-end-minus-plus-5.fasta

~/Downloads/weblogo/seqlogo -f only_Albin_RI-positive-and-negative-reverse-strand-Chr-start-minus-plus-5.fasta -F PNG -s -5 -c -S -n -w 17 > only_Albin_RI-start-weblogo.png

~/Downloads/weblogo/seqlogo -f only_Albin_RI-positive-and-negative-reverse-strand-Chr-end-minus-plus-5.fasta -F PNG -s -4 -c -S -n -w 17 > only_Albin_RI-end-weblogo.png

```

7.Παράρτημα 2

comparison.pl - perl script για την εύρεση μοναδικών γεγονότων εναλλακτικού ματίσματος μεταξύ των 2 δειγμάτων μετά την ανάλυση με το SplAdder:

```
#!/usr/bin/perl -w

@ARGV == 2 or die "Usage <mutant> <wildtype>\n";

open (FILE1,"$ARGV[0]");

open (FILE2,"$ARGV[1]");

$n = 0;

while ($f1 = <FILE1>)

{

@file1 = split ("\t", $f1);

#print "$file1[4]\n";

    while ($f2 = <FILE2>)

    {

        @file2 = split ("\t", $f2);

        if ( ($file2[0] eq $file1[0]) && ($file2[1] eq $file1[1]) && ($file2[3] eq $file1[3]) && ($file2[4] == $file1[4]) && ($file2[5] == $file1[5]) && ($file2[6] == $file1[6]) && ($file2[7] == $file1[7]) && ($file2[8] == $file1[8]) && ($file2[9] == $file1[9]) )

        {

            $n=1;

            last;

        }

    }

    if ($n != 1)

    {

        print "$f1";

    }

    $n = 0;

    seek (FILE2,0,0);

}

}
```

Perl scripts για την αντιστοίχιση των ονομάτων των γονιδίων του *Arabidopsis* (AGI) με τα γεγονότα εναλλακτικού ματίσματος όπως βρέθηκαν από το πρόγραμμα JuncBASE:

ATG_from_gtf_RL.pl

```
#!/usr/bin/perl -w

open (GTF, $ARGV[0]);

open (FILE, $ARGV[1]);

while ($file = <FILE>)

{

    if ($file =~ /(Chr.)\:[0-9]*\-[0-9]*/)

    {

        $chr = $1;

        $start = $2;

        $end = $3;

        $ref = $start - 1;

        #print "$chr\t$start";

        while ($gtf = <GTF>)

        {

            if (($gtf =~ /$chr/) && $gtf =~ /$ref/)

            {

                if ($gtf =~ /gene_id\s\".*\";\sgene_name\s\"(.*)\"/)

                {

                    print "$1\t$chr\t$start\t$end\n";

                }

            }

            last;

        }

    }

    seek (GTF,0,0);

}
```

ATG_from_gtf_A5SS.pl

```
#!/usr/bin/perl -w

open (GTF, $ARGV[0]);
open (FILE, $ARGV[1]);
while ($file = <FILE>)
{
    if ($file =~ /(Chr.)\;([0-9]*)\-([0-9]*)/g)
    {
        $chr = $1;
        $start = $2;
        $end = $3;
        $ref = $end + 1;
        #print "$chr\t$start";
        while ($gtf = <GTF>)
        {
            if (($gtf =~ /$chr/) && $gtf =~ /$ref/)
            {
                if ($gtf =~ /gene_id\s\".*\";\s?gene_name\s\"(.*)\"/)
                {
                    print "$1\t$chr\t$start\t$end\n";
                }
            }
            last;
        }
    }
    seek (GTF,0,0);
}
```

ATG_from_gtf_for_SE.pl

```
#!/usr/bin/perl -w
```

```

open (GTF, $ARGV[0]);

open (FILE, $ARGV[1]);

while ($file = <FILE>)
{
    if ($file =~ /(Chr.)\:[0-9]*\-[0-9]*)/)
    {
        $chr = $1;

        $start = $2;

        $end = $3;

        $ref = $start;

        #print "$chr\t$start";

        while ($gtf = <GTF>)
        {
            if (($gtf =~ /$chr/) && $gtf =~ /$ref/)
            {
                if ($gtf =~ /gene_id\s\".*\";\sgene_name\s\"(.*)\"/)
                {
                    print "$1\t$chr\t$start\t$end\n";
                }
            }
            last;
        }
    }
}

seek (GTF,0,0);

```

compare.pl - perl script για την εύρεση μοναδικών γεγονότων εναλλακτικού ματίσματος μεταξύ των 2 δειγμάτων μετά την ανάλυση με το ASprofile και το ASATP:

```

#!/usr/bin/perl

open (MUT,$ARGV[0]);

open (WT,$ARGV[1]);

while ($file1 = <MUT>)
{

```

```

    $n=0;

    while ($file2 = <WT>)
    {

        if ($file1 eq $file2)
        {

            $n=1;

            last;

        }

    }

    if ( $n == 0 )
    {

        print $file1;

    }

    seek (WT,0,0);
}

```

find_strand_asprofile.pl - perl script για την εύρεση της κατεύθυνσης της αλυσίδας στα γεγονότα παράλειψης ιντρονίου μετά την ανάλυση με το ASprofile:

```

#!/usr/bin/perl -w

open (FILE1, "$ARGV[0]") or die "Cannot open AS file\n";

open (FILE2, "$ARGV[1]") or die "Cannot open GTF file\n";

while ($line = <FILE1>)
{

    chomp $line;

    @array = split("\t", $line);

    $ref = $array[2] - 1;

    while ( $gtf = <FILE2> )
    {

        if ( ($gtf =~ /$ref/) && ($gtf =~ /$array[0]/) )

```

```

        {
            if ( $gtf =~ /(\\+)/ )
            {
                print "$line\\t$1\\n";
                last;
            }
            elsif ( $gtf =~ /(\\-)/ )
            {
                print "$line\\t$1\\n";
                last;
            }
        }
    }
    seek (FILE2,0,0);
}

close (FILE1);
close (FILE2);

```

reverse_negative_strand.pl - perl script για την δημιουργία των ανάστροφων συμπληρωματικών αλληλουχιών για όσες αλληλουχίες έχουν κατεύθυνση 3' -> 5' (-):

```

#!/usr/bin/perl -w

open (FASTA, "$ARGV[0]") or die "Cannot open fasta file\\n";

while ($line = <FASTA>)
{
    chomp $line;
    if ($line !~ />/)
    {
        $rev = ($line =~ tr/ATCG/TAGC/);
        $rev = reverse ($line);
        print "$rev\\n";
    }
}

```

```

        else
        {
            print "$line\n";
        }
    }
}

close (FASTA);

```

find_strand_79_overlap.pl – perl script για την εύρεση της κατεύθυνσης της αλυσίδας στα 79 κοινά γεγονότα παράλειψης ιντρονίου

```

#!/usr/bin/perl -w

open (STRAND, $ARGV[0]);

open (FILE, $ARGV[1]);

while ($file = <FILE>)
{
    @f = split(" ", $file);

    while ($strand = <STRAND>)
    {
        @s = split(" ", $strand);

        if ( ($f[0] eq $s[0]) && ($f[1] eq $s[1]) && ($f[2] == $s[3]) && ($f[3] == $s[4]) )
        {
            print ("$strand");

            last;
        }
    }

    seek (STRAND,0,0);
}

close(STRAND);

close(FILE);

```

8. Παράρτημα 3

Λίστα των 9 κοινών γεγονότων παράλειψης εξωνίου

AGI	Γονίδιο	Χρωμόσωμα	Έναρξη εξωνίου	Τέλος εξωνίου	Επίπεδο συμπερίληψης (rMATS)	Επίπεδο συμπερίληψης (Asprofile)
AT5G20450	myosin(AT5G20450)	Chr5	6912004	6912135	0.632	0.271
AT5G06950	bZIP transcription factor family protein(AHBP-1B)	Chr5	2152061	2152122	0.787	0.666
AT1G61150	LisH and RanBPM domains containing protein(AT1G61150)	Chr1	22542958	22542981	0.849	0.77
AT1G61150	LisH and RanBPM domains containing protein(AT1G61150)	Chr1	22542863	22542981	0.849	0.73
AT1G14760	KNOX meinox protein(KNATM)	Chr1	5084826	5084951	0.931	0.938
AT1G55840	Sec14p-like phosphatidylinositol transfer family protein(AT1G55840)	Chr1	20874058	20874086	0.985	0.999
AT5G65050	AGAMOUS-like 31(AGL31)	Chr5	25985654	25985695	0.986	0.999
AT2G38880	nuclear factor Y, subunit B1(NF-YB1)	Chr2	16240045	16240050	0.989	0.999
AT2G33380	Caleosin-related family protein(RD20)	Chr2	14145270	14145395	0.989	0.967

*Με έντονη γραφή σημειώνονται τα κοινά γεγονότα των rMATS-JuncBASE-ASprofile-ASATP, ενώ με κανονική γραφή μόνο αυτά των rMATS-ASprofile-ASATP.

Λίστα των 18 κοινών γεγονότων εναλλακτικής 3' θέσης ματίσματος

AGI	Γονίδιο	Επίπεδο συμπερίληψης (ASprofile)	Επίπεδο συμπερίληψης (JuncBASE)
AT4G13100	RING/U-box superfamily protein(AT4G13100)	0.92	0.76
AT3G26670	magnesium transporter, putative (DUF803)(AT3G26670)	0.75	0.5
AT2G32640	Lycopene beta/epsilon cyclase protein(AT2G32640)	0.74	0.75
AT2G22610	Di-glucose binding protein with Kinesin motor domain-containing protein(AT2G22610)	0.71	0.4
AT5G01780	2-oxoglutarate-dependent dioxygenase family protein(AT5G01780)	0.617	0.75
AT4G02760	RNI-like superfamily protein(AT4G02760)	0.6	0.36
AT3G52170	DNA binding protein(AT3G52170)	0.35	0.33
AT1G30160	hypothetical protein (DUF295)(AT1G30160)	0.31	1
AT2G05520	glycine-rich protein 3(GRP-3)	0.29	0.09
AT1G78150	N-lysine methyltransferase(AT1G78150)	0.23	0.016
AT1G02850	beta glucosidase 11(BGLU11)	0.22	0.11
AT4G20380	LSD1 zinc finger family protein(LSD1)	0.205	0.083
AT4G39140	RING/U-box superfamily protein(AT4G39140)	0.121	0.266
AT4G29810	MAP kinase kinase 2(MKK2)	0.086	0.105
AT4G12790	P-loop containing nucleoside triphosphate hydrolases superfamily protein(AT4G12790)	0.078	0.083
AT5G19400	Telomerase activating protein Est1(SMG7)	0.068	0.063
AT3G17365	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(AT3G17365)	0.03	0.04
AT1G02840	RNA-binding (RRM/RBD/RNP motifs) family protein(SR34)	0.007	0.007

*Με έντονη γραφή σημειώνονται τα κοινά γεγονότα και των 5 προγραμμάτων, ενώ με κανονική γραφή μόνο αυτά των JuncBASE-ASprofile-ASATP.

Λίστα των 12 κοινών γεγονότων εναλλακτικής 5' θέσης ματίσματος

AGI	Γονίδιο	Επίπεδο συμπερίληψης (rMATS)	Επίπεδο συμπερίληψης (ASprofile)
AT2G41700	ATP-binding cassette A1(ABCA1)	0.384	0.071
AT3G61160	Protein kinase superfamily protein(AT3G61160)	0.331	0.505
AT3G57470	Insulinase (Peptidase family M16) family protein(AT3G57470)	0.206	0.05
AT3G07740	ADA2 2A(ADA2A)	0.105	0.198
AT2G40935	PLAC8 family protein(AT2G40935)	0.099	0.204
AT5G49970	pyridoxin (pyrodoxamine) 5'-phosphate oxidase(PPOX)	0.09	0.009
AT4G09980	Methyltransferase MT-A70 family protein(EMB1691)	0.087	0.059
AT1G28100	hypothetical protein(AT1G28100)	0.074	0.133
AT5G57060	60S ribosomal L18a-like protein(AT5G57060)	0.068	0.058
AT4G19160	transglutaminase family protein(AT4G19160)	0.037	0.052
AT4G19040	ENHANCED DISEASE RESISTANCE 2(EDR2)	0.0162	0.05
AT4G31530	NAD(P)-binding Rossmann-fold superfamily protein(AT4G31530)	0.01	0.014

*Με έντονη γραφή σημειώνονται τα κοινά γεγονότα των προγραμμάτων rMATS-JuncBASE-ASprofile-ASATP, ενώ με κανονική γραφή μόνο αυτά των rMATS-ASprofile-ASATP.

Λίστα των 79 κοινών γεγονότων διατήρησης ιντρονίου

AGI	Γονίδιο	Χρωμόσωμα	Έναρξη ιντρονίου	Τέλος ιντρονίου	Επίπεδο συμπερίληψης (rMATS)	Επίπεδο συμπερίληψης (ASprofile)
AT3G20480	tetraacyldisaccharide 4 ⁺ -kinase family protein(LpxK)	Chr3	7151851	7151963	1	0.719
AT5G07890	myosin heavy chain-like protein(AT5G07890)	Chr5	2519113	2519210	0.501	0.271
AT3G55480	protein affected trafficking 2(PAT2)	Chr3	20570721	20570792	0.367	0.186
AT2G21940	shikimate kinase 1(SK1)	Chr2	9352570	9352668	0.333	0.211
AT3G59320	solute carrier family 35 protein (DUF914)(AT3G59320)	Chr3	21926136	21926247	0.319	0.344
AT1G02850	beta glucosidase 11(BGLU11)	Chr1	631859	632008	0.286	0.133
AT1G03055	beta-carotene isomerase D27-like protein(D27)	Chr1	710802	710954	0.282	0.174
AT5G10320	ATP synthase subunit B(AT5G10320)	Chr5	3244605	3244673	0.282	0.552
AT5G66070	RING/U-box superfamily protein(AT5G66070)	Chr5	26422519	26422590	0.278	0.253
AT3G26100	Regulator of chromosome condensation (RCC1) family protein(AT3G26100)	Chr3	9538282	9538376	0.263	0.431
AT5G58200	Calcineurin-like metallo-phosphoesterase superfamily protein(AT5G58200)	Chr5	23549904	23550031	0.207	0.164
AT1G10070	branched-chain amino acid transaminase 2(BCAT-2)	Chr1	3288321	3288430	0.191	0.134
AT5G03900	Iron-sulfur cluster biosynthesis family protein(AT5G03900)	Chr5	1050994	1051225	0.176	0.120
AT4G26370	antitermination NusB domain-containing protein(AT4G26370)	Chr4	13334942	13335078	0.173	0.047
AT2G37050	Leucine-rich repeat protein kinase family protein(AT2G37050)	Chr2	15572538	15572612	0.165	0.104
AT5G22820	ARM repeat superfamily protein(AT5G22820)	Chr5	7624042	7624136	0.161	0.222
AT5G26760	RNA polymerase II subunit B1 CTD phosphatase RPAP2-like protein(AT5G26760)	Chr5	9405755	9405964	0.138	0.275
AT5G43310	COP1-interacting protein-like protein(AT5G43310)	Chr5	17383590	17383685	0.127	0.224
AT5G03540	exocyst subunit exo70 family protein A1(EXO70A1)	Chr5	889984	890103	0.114	0.035
AT3G04770	40s ribosomal protein SA B(RPSAb)	Chr3	1309856	1309932	0.111	0.058
AT1G47490	RNA-binding protein 47C(RBP47C)	Chr1	17426097	17426209	0.109	0.100
AT1G23360	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(MENG)	Chr1	8296343	8296418	0.102	0.141
AT5G64960	cyclin dependent kinase group C2(CDKC2)	Chr5	25956088	25956185	0.1	0.054
AT5G04510	3'-phosphoinositide-dependent protein kinase 1(PDK1)	Chr5	1289174	1289289	0.098	0.040
AT5G01230	S-adenosyl-L-methionine-dependent methyltransferases superfamily protein(AT5G01230)	Chr5	94002	94085	0.09	0.039
AT5G05540	small RNA degrading nuclease 2(SDN2)	Chr5	1638251	1638336	0.082	0.092
AT1G33050	hypothetical protein(AT1G33050)	Chr1	11968508	11968588	0.078	0.181
AT5G09230	sirtuin 2(SRT2)	Chr5	2873272	2873343	0.067	0.086
AT1G29170	SCAR family protein(WAVE2)	Chr1	10190647	10190746	0.064	0.049
AT5G65440	transmembrane protein(AT5G65440)	Chr5	26156501	26156583	0.064	0.183
AT4G25650	ACD1-like protein(ACD1-LIKE)	Chr4	13082091	13082159	0.063	0.060
AT4G02260	RELA/SPOT homolog 1(RSH1)	Chr4	990663	990748	0.063	0.145
AT4G30993	Calcineurin-like metallo-phosphoesterase superfamily protein(AT4G30993)	Chr4	15099411	15099503	0.062	0.054
AT4G33000	calcineurin B-like protein 10(CBL10)	Chr4	15926041	15926140	0.062	0.084
AT1G77080	K-box region and MADS-box transcription factor family protein(MAF1)	Chr1	28958664	28958744	0.061	0.113
AT2G36680	Modifier of rudimentary (Mod(r)) protein(AT2G36680)	Chr2	15369367	15369448	0.049	0.061
AT2G48100	Exonuclease family protein(AT2G48100)	Chr2	19672381	19672449	0.044	0.032
AT5G48440	FAD-dependent oxidoreductase family protein(AT5G48440)	Chr5	19630199	19630294	0.041	0.048
AT4G36090	oxidoreductase, 2OG-Fe(II) oxygenase family protein(AT4G36090)	Chr4	17078950	17079045	0.041	0.059

AT1G54360	TBP-ASSOCIATED FACTOR 6B(TAF6B)	Chr1	20290876	20290943	0.038	0.049
AT5G15610	Proteasome component (PCI) domain protein(AT5G15610)	Chr5	5081588	5081674	0.037	0.024
AT4G24290	MAC/Perforin domain-containing protein(AT4G24290)	Chr4	12596499	12596808	0.036	0.022
AT1G33270	Acyl transferase/acyl hydrolase/lysophospholipase superfamily protein(AT1G33270)	Chr1	12068655	12068828	0.036	0.039
AT1G11310	Seven transmembrane MLO family protein(MLO2)	Chr1	3801461	3801549	0.034	0.032
AT3G22425	imidazoleglycerol-phosphate dehydratase(IGPD)	Chr3	7952238	7952346	0.032	0.045
AT2G23080	Protein kinase superfamily protein(CKA3)	Chr2	9829007	9829077	0.032	0.009
AT5G58970	uncoupling protein 2(UCP2)	Chr5	23809037	23809538	0.031	0.033
AT5G14250	Proteasome component (PCI) domain protein(COP13)	Chr5	4599939	4600030	0.029	0.027
AT4G38225	glycerol kinase(AT4G38225)	Chr4	17928197	17928284	0.029	0.019
AT3G12100	Cation efflux family protein(AT3G12100)	Chr3	3855029	3855116	0.026	0.044
AT5G01670	NAD(P)-linked oxidoreductase superfamily protein(AT5G01670)	Chr5	252522	252602	0.026	0.134
AT1G69250	Nuclear transport factor 2 (NTF2) family protein with RNA binding (RRM-RBD-RNP motifs) domain-containing protein(AT1G69250)	Chr1	26034712	26034809	0.025	0.020
AT2G44470	beta glucosidase 29(BGLU29)	Chr2	18356821	18356916	0.024	0.031
AT1G07420	sterol 4-alpha-methyl-oxidase 2-1(SMO2-1)	Chr1	2278725	2278971	0.024	0.010
AT1G60850	DNA-directed RNA polymerase family protein(ATRPAC42)	Chr1	22398565	22398677	0.023	0.032
AT5G62520	similar to RCD one 5(SRO5)	Chr5	25098862	25098944	0.023	0.017
AT5G15550	Transducin/WD40 repeat-like superfamily protein(AT5G15550)	Chr5	5059267	5059350	0.022	0.015
AT3G14420	Aldolase-type TIM barrel family protein(GOX1)	Chr3	4821941	4822032	0.021	0.033
AT4G37550	Acetamidase/Formamidase family protein(AT4G37550)	Chr4	17645229	17645341	0.02	0.010
AT2G01350	quinolinate phosphoribosyltransferase(QPT)	Chr2	166857	166934	0.019	0.016
AT4G17150	alpha/beta-Hydrolases superfamily protein(AT4G17150)	Chr4	9640924	9641005	0.019	0.023
AT1G45160	Protein kinase superfamily protein(AT1G45160)	Chr1	17084115	17084213	0.018	0.031
AT3G56130	biotin/lipoyl attachment domain-containing protein(AT3G56130)	Chr3	20827694	20827900	0.017	0.015
AT3G18140	Transducin/WD40 repeat-like superfamily protein(LST8-1)	Chr3	6213149	6213228	0.017	0.013
AT3G06960	pigment defective 320(PDE320)	Chr3	2195024	2195337	0.017	0.012
AT5G63440	hypothetical protein (DUF167)(AT5G63440)	Chr5	25407739	25407866	0.015	0.018
AT3G19460	Reticulon family protein(AT3G19460)	Chr3	6748499	6748600	0.014	0.006
AT5G66040	sulfurtransferase protein 16(STR16)	Chr5	26410806	26410868	0.012	0.008
AT4G31860	Protein phosphatase 2C family protein(AT4G31860)	Chr4	15407035	15407119	0.012	0.016
AT2G37340	arginine/serine-rich zinc knuckle-containing protein 33(RS2233)	Chr2	15671421	15671560	0.01	0.009
AT3G52180	dual specificity protein phosphatase (DsPTP1) family protein(SEX4)	Chr3	19350347	19350612	0.009	0.004
AT5G02020	E3 ubiquitin-protein ligase RLIIM-like protein(SIS)	Chr5	386722	387040	0.008	0.012
AT5G16390	chloroplastic acetylcoenzyme A carboxylase 1(CAC1)	Chr5	5361338	5361565	0.007	0.006
AT1G07705	NOT2 / NOT3 / NOT5 family(AT1G07705)	Chr1	2385279	2385569	0.007	0.013
AT5G64130	cAMP-regulated phosphoprotein 19-related protein(AT5G64130)	Chr5	25665030	25665104	0.007	0.005
AT4G29360	O-Glycosyl hydrolases family 17 protein(AT4G29360)	Chr4	14451852	14451942	0.006	0.028
AT2G46280	TGF-beta receptor interacting protein 1(TRIP-1)	Chr2	19004107	19004199	0.004	0.036
AT5G60600	4-hydroxy-3-methylbut-2-enyl diphosphate synthase(HDS)	Chr5	24362882	24362966	0.003	0.005
AT1G63770	Peptidase M1 family protein(AT1G63770)	Chr1	23658110	23658215	0.001	0.021

*Με έντονη γραφή σημειώνεται το κοινό γεγονός μεταξύ των προγραμμάτων SplAdder-rMATS-ASprofile-ASATP, ενώ με κανονική γραφή μόνο αυτά των rMATS-ASprofile-ASATP