ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΒΙΟΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΕΘΝΙΚΟ ΙΔΡΥΜΑ ΕΡΕΥΝΩΝ
ΙΝΣΤΙΤΟΥΤΟ ΒΙΟΛΟΓΙΑΣ, ΦΑΡΜΑΚΕΥΤΙΚΗΣ ΧΗΜΕΙΑΣ ΚΑΙ ΒΙΟΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΙΔΡΥΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΒΙΟΕΠΙΧΕΙΡΕΙΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΜΕΛΕΤΗ ΑΝΤΑΓΩΝΙΣΜΟΥ ΣΤΟ ΧΩΡΟ ΤΟΥ ΛΟΓΙΣΜΙΚΟΥ ΑΝΑΛΥΣΗΣ ΒΙΟΛΟΓΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕΓΑΛΟΥ ΟΓΚΟΥ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΕΡΕΥΝΗΤΗΣ Β΄ ΕΙΕ, ΓΕΩΡΓΙΑΔΗΣ ΠΑΝΑΓΙΩΤΗΣ

ΜΑΡΓΑΡΙΤΗΣ ΕΥΣΤΑΘΙΟΣ
Α.Μ : 00021

ΑΘΗΝΑ 2018

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

UNIVERSITY OF THESSALY
SCHOOL OF HEALTH SCIENCES

DEPARTMENT OF BIOCHEMISTRY AND BIOTECHNOLOGY

NATIONAL HELLENIC RESEARCH FOUNDATION
INSTITUTE OF BIOLOGY, MEDICINAL CHEMISTRY & BIOTECHNOLOGY

INTERISTITUTIONAL PROGRAM OF POSTGRADUATE STUDIES
IN
BIOENTREPRENEURSHIP

MASTER THESIS

STUDY OF THE COMPETITION IN THE FIELD OF SOFTWARE ANALYSIS OF BIG BIOLOGICAL DATA

SUPERVISOR: RESEARCHER 2ND DEGREE (B΄), GEORGIADIS PANAGIOTIS

MARGARITIS EFSTATHIOS

R.N. : 00021

ATHENS 2018

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο σπουδών

για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στο

ΒΙΟΕΠΙΧΕΙΡΕΙΝ

που απονέμει το Τμήμα Βιοχημείας και Βιοτεχνολογίας του Πανεπιστημίου Θεσσαλίας, σε συνεργασία με την εταιρεία HybridStat

Εγκρίθηκε την …….. από την τριμελή εξεταστική επιτροπή:

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	**ΒΑΘΜΙΔΑ**	**ΥΠΟΓΡΑΦΗ**
(ΓΕΩΡΓΙΑΔΗΣ)		
(ΠΛΕΤΣΑ)		
(ΣΑΡΑΦΙΔΟΥ)		

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Ευχαριστίες

Η συγγραφή αυτής της διπλωματικής εργασίας θα ήταν αδύνατη χωρίς τη βοήθεια και τη στήριξη ορισμένων ανθρώπων. Πρωτίστως θα ήθελα να ευχαριστήσω τους γονείς μου για τη συνεχή, ουσιαστική και ανιδιοτελή υποστήριξη που μου προσέφεραν καθ' όλη τη διάρκεια παρακολούθησης του μεταπτυχιακού προγράμματος <<ΒΙΟΕΠΙΧΕΙΡΕΙΝ>>. Στη συνέχεια, θα ήθελα να ευχαριστήσω τον Δρ. Παναγιώτη Μούλο, HybridStat Predictive Analytics, για τις εποικοδομητικές συμβουλές και διευκρινήσεις του χωρίς τις οποίες θα ήταν αδύνατη η ολοκλήρωση της εργασίας αυτής. Ακόμη, τον Δρ. Παναγιώτη Γεωργιάδη, Ερευνητή Β΄ στο Εθνικό Ίδρυμα Ερευνών (ΕΙΕ), για το ενδιαφέρον και την καθοδήγηση του και βέβαια την Δρ Βασιλική Πλέτσα, Ερευνήτρια Γ' στο Εθνικό Ίδρυμα Ερευνών (ΕΙΕ), για τη συνεχή και ουσιαστική βοήθεια της τόσο σε ότι αφορά τη διπλωματική αυτή όσο και σαν σύμβουλος σπουδών του μεταπτυχιακού προγράμματος.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

ΠΕΡΙΛΗΨΗ

Η ορθή και ενδελεχής ανάλυση των δεδομένων μεγάλου όγκου (Big Data) αποτελεί μία από τις μεγαλύτερες προκλήσεις του αιώνα που διανύουμε. Η βέλτιστη ανάλυση των Big Data που προέρχονται από βιολογικές διεργασίες μπορεί να οδηγήσει μεταξύ άλλων σε σημαντικές μειώσεις του κόστους πρόγνωσης, διάγνωσης και θεραπείας μιας ασθένειας αλλά και στην επιτάχυνση της λήψης εξατομικευμένων αποφάσεων για κάθε ασθενή του πληθυσμού μιας χώρας. Για αυτό το λόγο, πολλοί εξειδικευμένοι επιστήμονες ανά τον κόσμο καταβάλλουν συνεχείς προσπάθειες με σκοπό τη δημιουργία πλατφορμών λογισμικού για την ανάλυση των δεδομένων που προκύπτουν από την εφαρμογή μεθόδων αλληλούχισης με τη μεγαλύτερη δυνατή ακρίβεια. Ειδικότερα, οι τεχνικές που εφαρμόζονται στο χώρο της υγείας σήμερα ανήκουν κατά κύριο λόγο στη δεύτερη και τρίτη γενιά των μεθόδων αλληλούχισης και λιγότερο στην πρώτη γενιά που ονομάζεται αλλιώς Sanger αλληλούχιση. Ωστόσο, η έλλειψη σε εξειδικευμένο και κατάλληλα καταρτισμένο ανθρώπινο δυναμικό σε συνδυασμό με τους περιορισμένους υπολογιστικούς πόρους βάζουν τροχοπέδη στις προσπάθειες τόσο των επιστημόνων όσο και των εταιρειών που δραστηριοποιούνται στο χώρο πρόγνωσης, διάγνωσης και θεραπείας των ασθενειών.

ΛΕΞΕΙΣ-ΚΛΕΙΔΙΑ: Δεδομένα (Big Data), μέθοδοι αλληλούχισης, λογισμικά δεύτερης γενιάς αλληλούχισης (RNA-Seq), σύγκριση πλατφορμών, περιγραφή <<ιδανικού>> RNA-Seq λογισμικού

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

ABSTRACT

Among the biggest challenges of the 21th century is the optimal analysis of Big Data. The analysis of Big Data which is originated from biological process can lead to significant cost reductions in the field of prognosis, diagnosis and cure of a disease. For this reason, many scientists around the world are constantly striving to create software platforms in order to effectively analyse data resulting from sequencing methods. Particularly, the techniques, applied today in the health field, belong predominantly to next generation sequencing methods and less to the first generation which are otherwise called Sanger sequencing.

In this context, it is worth mentioning that the results of this diploma thesis are expected to be exploited by HybridSTAT, which is active in the United Kingdom and Greece in the area of predictive analysis. HybridSTAT is a company, established in 2014 according to its website that offers services through 2 applications it has already developed, covering biostatics, bioinformatics and high throughput bio-data analysis.

KEYWORDS: Big Data, sequencing methods, second generation sequencing software (RNA-Seq), comparison of software, description of "ideal RNA-Seq software"

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

ΣΚΟΠΟΣ

Η διπλωματική αυτή εργασία εκπονήθηκε με κύριο σκοπό τη σύγκριση, σε παγκόσμιο επίπεδο, λογισμικών των εταιρειών που δραστηριοποιούνται στο χώρο ανάλυσης βιολογικών δεδομένων μεγάλου όγκου. Πιο συγκεκριμένα, η σύγκριση αφορούσε την ανάλυση δεδομένων RNA-seq που προέρχονται από αλληλούχιση δεύτερης γενιάς ώστε να προταθεί το <<ιδανικό>> λογισμικό για ένα μη έμπειρο υπολογιστικά χρήστη. Τα αποτελέσματα της εργασίας θα αξιοποιηθούν περαιτέρω από την εταιρεία HybridStat Predictive Analytics για την ανάπτυξη ενός τέτοιου λογισμικού φιλικού προς το χρήστη (user-friendly).

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

1. ΕΙΣΑΓΩΓΗ

Στον τίτλο της παρούσας εργασίας ο όρος << βιολογικά δεδομένα μεγάλου όγκου>> (κατηγορία των Big Data) αναφέρεται σε δεδομένα προερχόμενα από τη χρήση τεχνολογιών Αλληλούχισης Νέας Γενιάς (Next Generation Sequencing, NGS). Επομένως, κρίνεται απαραίτητη η διενέργεια μιας σύντομης αναφοράς τόσο στα δεδομένα που προέρχονται από βιολογικές διεργασίες και ανήκουν στην κατηγορία των Big Data όσο και στις μεθόδους αλληλούχισης γενικότερα.

1.1 Βιολογικά Δεδομένα Μεγάλου Όγκου (Big Data)

Ο 21ος αιώνας σηματοδοτεί την έναρξη της επεξεργασίας των δεδομένων μεγάλου όγκου (Big Data) που υπάρχουν πλέον σε όλους σχεδόν τους τομείς της ανθρώπινης δραστηριότητας (Zhang, 2014). Πιο συγκεκριμένα, η πρόοδος που έχει σημειωθεί τις τελευταίες δεκαετίες σε όλους τους τομείς της τεχνολογίας των –omics- όπως στη γονιδιωματική (genomics) και στην προτεομική (proteomics) έχει ως αποτέλεσμα την παραγωγή μιας τεράστιας ποσότητας σχετικών με τη μοριακή βιολογία δεδομένων (Shan Li et al., 2014). Ωστόσο, τα δεδομένα μεγάλου όγκου που συναντώνται με μεγαλύτερη συχνότητα στο χώρο της βιοπληροφορικής κατατάσσονται σε πέντε ομάδες οι οποίες είναι οι εξής: Τα δεδομένα γονιδιακής έκφρασης (gene expression data), τα δεδομένα DNA, RNA και αλληλούχισης πρωτεϊνών (DNA, RNA and protein sequence), τα δεδομένα από την αλληλεπίδραση μεταξύ πρωτεϊνών (protein-protein interaction), τα δεδομένα βιοχημικών μονοπατιών (biochemical pathway data) και τέλος τα δεδομένα γονιδιακής οντολογίας (gene ontology) (Gharajeh, 2018) .

Αναλυτικότερα και σύμφωνα με την προηγούμενη αναφορά (Gharajeh, 2018), η πρώτη ομάδα αφορά την ανάλυση της έκφρασης γονιδίων με χρήση ενός προφίλ γονιδιακής έκφρασης που βασίζεται στις μικροσυστοιχίες. Η δεύτερη ομάδα αναφέρεται στη χρήση ποικίλων αναλυτικών μεθόδων με σκοπό την ταυτοποίηση των χαρακτηριστικών και της δομής των δεδομένων που προέρχονται από αλληλούχιση DNA (DNA-seq), από αλληλούχιση RNA (RNA-seq) καθώς και από αλληλούχιση πρωτεϊνών. Η τρίτη ομάδα εμπεριέχει την ανάλυση της αλληλεπίδρασης μεταξύ πρωτεϊνών από την οποία προκύπτουν πληροφορίες σχετικά με τη λειτουργία τους. Η τέταρτη ομάδα αποτελείται από δεδομένα βιοχημικών μονοπατιών που βοηθούν στην κατανόηση των ασθενειών σε μοριακό επίπεδο καθώς και στην εκτίμηση πιθανών στόχων των φαρμάκων. Η τελευταία ομάδα υποστηρίζει τη δομημένη ανάλυση του είδους των γονιδίων που συνδέονται με συγκεκριμένες βιολογικές διεργασίες, μοριακές λειτουργίες και κυτταρικά συστατικά.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

1.2 Ιστορική αναδρομή των μεθόδων αλληλούχισης

Η ιστορία της ανάπτυξης των μεθόδων αλληλούχισης ξεκινά το 1977 με τη δημοσίευση των τεχνικών αλληλούχισης πρώτης γενιάς και συνεχίζεται μέχρι σήμερα οπότε και εφαρμόζονται οι τεχνικές αλληλούχισης τρίτης γενιάς. Αναλυτικότερα, η **Αλληλούχιση Πρώτης Γενιάς** περιλαμβάνει 2 μεθόδους: Την Maxam και Gilbert Αλληλούχιση (Allan M. Maxam and Walter Gilbert., 1977) η οποία έγινε γνωστή για πρώτη φορά στο ευρύ κοινό το Φεβρουάριο του 1977 και την Sanger Αλληλούχιση (Sanger F. et al., 1977) η οποία δημοσιεύτηκε το Δεκέμβριο του ίδιου έτους. Η Sanger Αλληλούχιση είναι μία μέθοδος στην οποία επισημασμένα με βαφή δεοξυνουκλεοτίδια (dNTPS) αναμειγνύονται με τροποποιημένα διδεοξυνουκλεοτίδια. Στη συνέχεια λαμβάνει χώρα μία αντίδραση PCR και πραγματοποιείται επιμήκυνση μέχρις ότου κάποιοι κλώνοι (strands) να ενσωματώσουν ένα διδεοξυνουκλεοτίδιο. Ακολούθως, οι κλώνοι διαχωρίζονται πάνω σε μία γέλη και η τελευταία επισημασμένη βάση κάθε κλώνου ανιχνεύεται μέσω laser διέγερσης και ανάλυσης φασματικών εκπομπών (Goodwin S. et al., 2016).

Η Αλληλούχιση κατά Maxam και Gilbert περιλαμβάνει τη χημική τροποποίηση του DNA και επακόλουθη διάσπαση του σε συγκεκριμένες βάσεις (Liu Lin et al., 2012). Ωστόσο, η μέθοδος αυτή δεν χρησιμοποιήθηκε ευρέως καθώς παρουσίαζε 2 μειονεκτήματα (Gužvić, 2013). Το πρώτο ήταν ότι η εφαρμογή της απαιτούσε τη χρησιμοποίηση επικίνδυνων χημικών ουσιών όπως η υδραζίνη που είναι νευροτοξική για τον άνθρωπο. Το δεύτερο μειονέκτημα αφορούσε στο οξύ πρόβλημα που δημιουργούνταν εξαιτίας της απαίτησης χρήσης μεγάλης ποσότητας DNA σε συνδυασμό με την ανάγκη εκτέλεσης τεσσάρων ή ακόμη και οκτώ αντιδράσεων. Για αυτούς τους λόγους και εξαιτίας του ότι εμφανίζει υψηλότερη αποτελεσματικότητα και χαμηλότερη ραδιενέργεια η μέθοδος αλληλούχισης Sanger επικράτησε για 30 και πλέον χρόνια (Mardis Elaine R., 2008) στις εμπορικές εφαρμογές αλληλούχισης ενώ η μέθοδος Maxam και Gilbert χρησιμοποιήθηκε περισσότερο είτε στην περίπτωση αλληλούχισης πολύ μικρών κομματιών είτε στο σχεδιασμό εκκινητών όταν η διαδικασία της αλληλούχισης ήταν ανεπαρκής (Gužvić, 2013).

Η **Αλληλούχιση Δεύτερης ή Νέας Γενιάς (Next Generation Sequencing, NGS)** αναφέρεται στην τεχνολογία της μαζικής παράλληλης αλληλούχισης υψηλής απόδοσης η οποία χρησιμοποιείται σε πολλές εφαρμογές συμπεριλαμβανομένων της RNA αλληλούχισης (RNA-Seq), της αλληλούχισης ολόκληρου του γονιδιώματος (Whole Genome Sequencing, WGS), ολόκληρου του εξοσώματος (Whole Exome Sequencing, WES) (Byron S.A. et al., 2016) καθώς και στην αλληλούχιση ανοσοκατακρήμνισης της χρωματίνης (ChIP-Seq) (Raza K. and Sabahuddin A., 2016) όπως επίσης στην επιγενετική αλλά και σε τεστ ανίχνευσης πολλαπλών γονιδίων (multi-gene tests) (Pereira M. Araújo et al., 2017).

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Πιο συγκεκριμένα, το 1996 θεωρήθηκε ως σημείο αναφοράς για την έναρξη εμφάνισης της αλληλούχισης νέας γενιάς αφού δημοσιεύτηκε για πρώτη φορά η τεχνική της Πυροφωσφορικής Αλληλούχισης (Pyrosequencing) (Ronaghi M. et al., 1996). Η μέθοδος αυτή διέφερε από τις αντίστοιχες της 1ης γενιάς επειδή δεν χρησιμοποιούνταν ούτε ραδιενεργά ή σημασμένα με φθόριο νουκλεοτίδια ούτε υπήρχε η ανάγκη για ηλεκτροφόρηση. Αντιθέτως, η ανωτέρω μέθοδος βασίστηκε στη δράση 2 ενζύμων (ATP σουλφορυλάση και λουσιφεράση). Η ATP σουλφορυλάση μετέτρεπε το πυροφωσφορικό που βρισκόταν στα νουκλεοτίδια σε ένα μόριο ATP το οποίο στη συνέχεια χρησιμοποιούταν από υπόστρωμα λουσιφεράσης. Η διαδικασία αυτή είχε ως αποτέλεσμα την απελευθέρωση φωτός μέσω σήματος σε αναλογία με την ποσότητα των ενσωματωμένων νουκλεοτιδίων και τον προσδιορισμό της αλληλουχίας σύμφωνα με τη σειριακή προσθήκη των νουκλεοτιδίων.

Η τεχνική αυτή βελτιώθηκε στα επόμενα χρόνια και δόθηκε άδεια χρήσης το 2005 όταν και κυκλοφόρησε για πρώτη φορά ως η 1η μέθοδος αλληλούχισης δεύτερης γενιάς από την εταιρεία βιοτεχνολογίας <<454 Life Sciences>> η οποία στη συνέχεια εξαγοράστηκε από την εταιρεία Roche (Pereira M. Araújo et al., 2017). Από τότε αναπτύχθηκαν άλλες 3 τεχνολογίες αλληλούχισης αυτής της γενιάς από τις εταιρείες Illumina, Applied Biosystems και Thermo Fischer Scientific με τις κυριότερες μεταξύ τους διαφορές να εντοπίζονται στον τρόπο προσθήκης και ανίχνευσης των νουκλεοτιδίων.

Ωστόσο, οι τεχνολογίες αυτής της γενιάς παρουσιάζουν τα ακόλουθα κοινά χαρακτηριστικά (Kulski K. Jerzy., 2016). Πρώτον, η ανάλυση αλληλουχίας των τυχαίων τμημάτων γενομικού DNA (genomic (fg) DNA) ή αντίστροφου συμπληρωματικού προερχόμενου από RNA, DNA (cDNA) πραγματοποιείται χωρίς να απαιτείται η κλωνοποίηση σε κύτταρο-ξενιστή. Αντιθέτως, μικρές αλληλουχίες συνδέτες (linkers sequences) η/και προσαρμοστές (adapters) συνδέονται με το fgDNA ή το cDNA για την κατασκευή πρότυπων βιβλιοθηκών. Δεύτερον, η ενσωμάτωση των νουκλεοτιδίων εντοπίζεται αφενός μέσω ανίχνευσης φωταύγεια και αφετέρου μέσω αλλαγών που παρατηρούνται στο ηλεκτρικό φορτίο κατά τη διαδικασία της αλληλούχισης. Τρίτον, είναι δυνατή η ανάγνωση των άκρων των αλληλουχιών (fragment ends) είτε μέσω ατομικής μέτρησης των νουκλεοτιδικών βάσεων που βρίσκονται στα άκρα (single reads) είτε μέσω μέτρησης των βάσεων αυτών κατά ζεύγη (paired end reads). Τέλος, αξίζει να σημειωθεί ότι οι μέθοδοι της δεύτερης γενιάς παράγουν πολλά εκατομμύρια ολιγονουκλεοτιδίων σύντομης ανάγνωσης (short reads) ταυτοχρόνως σε πολύ λιγότερο χρόνο και με πολύ μικρότερο κόστος σε σχέση με τις μεθόδους της πρώτης γενιάς.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Η **Αλληλούχιση Τρίτης Γενιάς** (Third Generation Sequencing, **TGS**) αφορά στην άμεση ανάγνωση μεγαλομοριακού DNA χωρίς να απαιτείται το κόψιμο/σπάσιμο του, ο πολλαπλασιασμός του μέσω αλυσιδωτούς αντίδρασης πολυμεράσης (PCR) και η χρήση ενζυμικού συστήματος αντιγραφής για την ταυτοποίηση της αλληλουχίας (Gut, 2013) .
Οι τεχνικές αλληλούχισης τρίτης γενιάς μπορούν να ταξινομηθούν σε 3 διαφορετικές κατηγορίες. Στην πρώτη κατηγορία συγκαταλέγονται οι τεχνικές αλληλούχισης μέσω σύνθεσης (Sequencing by Synthesis, SBS) στις οποίες μόρια DNA πολυμεράσης συνθέτουν ένα μόριο DNA. Στη δεύτερη κατηγορία ανήκουν οι τεχνικές nanopore-sequencing στις οποίες μόρια DNA διέρχονται από έναν πόρο μεγέθους κάποιων νανομέτρων ή τοποθετούνται κοντά σε αυτόν και αυτό έχει ως αποτέλεσμα τον εντοπισμό των νουκλεοτιδικών βάσεων. Η τρίτη κατηγορία περιλαμβάνει τις τεχνικές άμεσης απεικόνισης μορίων DNA χρησιμοποιώντας προηγμένες τεχνικές μικροσκοπίας όπως είναι η ηλεκτρονική μικροσκοπία σάρωσης-μετάδοσης (Scanning Transmission Electron Microscopy, STEM) (Schadt E. Eric et al., 2010).

Ειδικότερα, η πρώτη τεχνική κυκλοφόρησε το 2009 από την εταιρεία Helicos Sequencer και στη συνέχεια από την Pacific Biosciences και ανήκει στην πρώτη από τις προαναφερθείσες τεχνικές (Heather M. James and Chain Benjamin., 2016). Στη συνέχεια η εταιρεία Oxford Nanopore Technologies (Bayley Hagan., 2015) ανέπτυξε την πλατφόρμα MinION η οποία κατατάσσεται στη δεύτερη κατηγορία ενώ όσον αφορά την τρίτη κατηγορία (Bell C. David et al., 2012) δεν έχει προκύψει ακόμη κάποια εμπορική πλατφόρμα αλληλούχισης αν και υπάρχουν εταιρείες που δραστηριοποιούνται στο χώρο της ηλεκτρονικής μικροσκοπίας όπως είναι η ZS Genetics.

Καθεμιά από τις ανωτέρω τεχνικές παρέχει καινοτόμες προσεγγίσεις σχετικά με την αλληλούχιση του DNA και όπως είναι φυσικό παρουσιάζει τόσο πλεονεκτήματα αλλά και μειονεκτήματα. Το κοινό χαρακτηριστικό όμως όλων των τεχνικών της γενιάς αυτής συνίσταται στο ότι όλες συντελούν στη διευκόλυνση της συναρμολόγησης των σύνθετων περιοχών του γονιδιώματος όπου εμφανίζονται συγχωνεύσεις, προσθήκες και διαγραφές γονιδίων καθώς και επαναλαμβανόμενες περιοχές (Pereira M. Araújo et al., 2017).

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2. ΚΥΡΙΟ ΜΕΡΟΣ ΕΡΓΑΣΙΑΣ

Η αλληλούχιση RNA είναι μία μέθοδος δεκαπέντε περίπου ετών με πλεονεκτήματα, μειονεκτήματα, εφαρμογές, προβλήματα και προκλήσεις. Τα κύρια πλεονεκτήματα χρήσης αυτής της μεθόδου είναι η απλή γενική ροή εργασιών, η υψηλή ταχύτητα δημιουργίας εύκολα διαχειρίσιμων αποτελεσμάτων και η διαθεσιμότητα υποστήριξης της από διάφορες πλατφόρμες και λογισμικά βιοπληροφορικής.

Η διαδικασία RNA-seq αλληλούχισης δεύτερης γενιάς που εφαρμόζεται στη συγκεκριμένη διπλωματική αφορά 6 δείγματα, σε 2 διαφορετικές βιολογικές συνθήκες, του εξεταζόμενου οργανισμού που είναι το ποντίκι mus musculus. Τα τρία δείγματα εξ αυτών, τα οποία εφεξής θα αναφέρονται ως Wild Type δείγματα, αποτελούν τμήμα του φαινότυπου του mus musculus ενώ από τα υπόλοιπα τρία δείγματα έχει αφαιρεθεί με τεχνητό τρόπο το γονίδιο Smyd3 και στο εξής θα αναφέρονται ως Smyd3KO (Knock Out) δείγματα.

Στη συνέχεια παρουσιάζονται με χρήση κειμένου και εικόνων/στιγμιότυπων τα λογισμικά ανάλυσης δεδομένων RNA-seq αλληλούχισης δεύτερης γενιάς που προέκυψαν ύστερα από έρευνα που έγινε στο διαδίκτυο. Το αντικείμενο της έρευνας αφορά 18 συνολικά χαρακτηριστικά τα οποία ταξινομούνται σε 3 κατηγορίες ως ακολούθως: α) Τυπικά χαρακτηριστικά συστήματος, β) Δυνατότητες ανάλυσης δεδομένων, γ) Δυνατότητες οπτικοποίησης δεδομένων.

Στην πρώτη κατηγορία κατατάσσονται τα τυπικά χαρακτηριστικά του συστήματος που είναι τα εξής:

1) **Required Space:** Αναφέρεται στο χώρο που καταλαμβάνει το κάθε λογισμικό στο σκληρό δίσκο του υπολογιστή δοκιμής.

2) **System Architecture**: Αναφέρεται στο είδος της άδειας χρήσης που παρέχει το καθένα λογισμικό. Η άδεια χρήσης μπορεί να υποστηρίζει την εκτέλεση του λογισμικού είτε σε έναν υπολογιστή (Desktop) είτε μέσω διαδικτύου σε οποιονδήποτε υπολογιστή (Cloud) ή να υποστηρίζει και τα 2 είδη ταυτόχρονα.

3) **Operating System**: Αναφέρεται στο λειτουργικό σύστημα του υπολογιστή το οποίο μπορεί να είναι είτε Windows είτε Macintosh (Mac) ή Linux.

Στη δεύτερη κατηγορία που αφορά τις δυνατότητες ανάλυσης δεδομένων συμπεριλαμβάνονται τα ακόλουθα χαρακτηριστικά:

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

1) **FASTQ ή/ και BAM αρχεία**: Είναι το είδος των αρχείων που υποστηρίζει το εκάστοτε λογισμικό. Πρόκειται για αρχεία κειμένου σε δυαδική μορφή που χρησιμοποιούνται για την αποθήκευση βιολογικών αλληλουχιών όπως είναι η αλληλουχία νουκλεοτιδίων. Η διαφορά μεταξύ τους συνίσταται στο ότι τα BAM αρχεία καταλαμβάνουν περισσότερο χώρο στο σκληρό δίσκο ενός υπολογιστή και προκύπτουν ύστερα από την ευθυγράμμιση (alignment) του γονιδιώματος αναφοράς ενός οργανισμού το οποίο αποθηκεύεται σε αρχεία FASTQ.

2) **Organisms**: Συγκαταλέγονται τα γονιδιώματα των οργανισμών τα οποία υποστηρίζουν τα λογισμικά των εταιρειών.

3) **Quality Control**: Περιλαμβάνονται οι διαδικασίες ποιοτικού ελέγχου των ακατέργαστων (raw) δεδομένων. Ειδικότερα, η ανάλυση του λογισμικού των εκάστοτε εταιρειών, που πραγματοποιείται με παράθεση εικόνων/διαγραμμάτων στη συνέχεια του κεφαλαίου, οδηγεί στο συμπέρασμα ότι ορισμένα λογισμικά υποστηρίζουν ένα ή περισσότερα χαρακτηριστικά γνωρίσματα του ποιοτικού ελέγχου τα οποία συνοπτικά είναι τα εξής:

- *Η μέση κατανομή μήκους σε ζεύγη βάσεων (bp; Length distribution) των αλληλουχηθέντων ολιγονουκλεοτιδίων*

  Πρόκειται για διαγράμματα στα οποία ο οριζόντιος άξονας αποτελείται από το μήκος αλληλούχισης που υπολογίζεται σε ζεύγη νουκλεοτιδικών βάσεων ενώ ο κάθετος από τον αριθμό των αλληλουχίσεων συγκεκριμένου μήκους κανονικοποιημένο (normalized) ως προς το σύνολο των αλληλουχίσεων.

- *Η περιεκτικότητα των νουκλεοτιδικών βάσεων σε Γουανίνη- Κυτοσίνη (GC-content)*

  Σε αυτού του είδους τα διαγράμματα ο οριζόντιος άξονας περιλαμβάνει τη σχετική συνεισφορά σε GC-content καθεμιάς αλληλούχισης εκφρασμένη σε ποσοστό επί τοις εκατό ενώ ο κάθετος εκφράζει τον αριθμό των αλληλουχίσεων που εμφανίζουν συγκεκριμένο ποσοστό σε GC-content κανονικοποιημένο ως προς τον συνολικό αριθμό των αλληλουχίσεων.

- *Η κατανομή της ποιότητας των αλληλουχηθέντων ολιγονουκλεοτιδίων*

  Είναι η κατανομή της ακρίβειας της αντιγραφής κατά τη διάρκεια της αλληλούχισης. Ειδικότερα, όταν ένα αντιγραφέν ολιγονουκλεοτίδιο έχει PHRED score 30 στον οριζόντιο άξονα αυτό σημαίνει ότι η πιθανότητα αντιγραφής μίας λανθασμένης νουκλεοτιδικής βάσης είναι P= 1/1000 και το PHRED score = -

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

$10\log_{10}P$. Για να είναι αποδεκτή μία κατανομή θα πρέπει το PHRED score να είναι μεγαλύτερο από 30 και ακόμη καλύτερα μεγαλύτερο από 40 επειδή σε αυτή την περίπτωση η πιθανότητα P= 1/10000 και η ακρίβεια φτάνει στο 99.9999 %.

- *Η κάλυψη κάθε νουκλεοτιδικής βάσης (Base coverage)*

  Πρόκειται για τον αριθμό των ανεξάρτητων αλληλουχίσεων κάθε βάσης μιας περιοχής. Όσο μεγαλύτερος είναι ο αριθμός τόσο μεγαλύτερη είναι η κάλυψη (Deep depth). Πιο συγκεκριμένα, στα διαγράμματα αυτού του χαρακτηριστικού γνωρίσματος ο οριζόντιος άξονας αποτελείται από τις θέσεις των βάσεων ενώ ο κάθετος άξονας περιλαμβάνει τον αριθμό των ανεξάρτητων αλληλουχίσεων κάθε βάσης κανονικοποιημένο ως προς το συνολικό αριθμό των αλληλουχίσεων.

- *Η συνεισφορά των νουκλεοτιδίων (Nucleotide contribution)*

  Σε αυτήν την περίπτωση η διαγραμματική απεικόνιση περιλαμβάνει την εκατοστιαία αναλογία της κάλυψης των τεσσάρων νουκλεοτιδικών βάσεων (A, C, G,T) για κάθε βάση μιας περιοχής αλληλούχισης. Σε μια ιδανική κατάσταση το αναμενόμενο είναι η ύπαρξη ελάχιστης ή ακόμα και μηδενικής διαφοράς μεταξύ των βάσεων οπότε οι γραμμές των διαγραμμάτων θα πρέπει να είναι παράλληλες μεταξύ τους. Με άλλα λόγια, οι σχετικές ποσότητες κάθε βάσης θα πρέπει να αντικατοπτρίζουν τη συνολική ποσότητα των βάσεων στο εξεταζόμενο γονιδίωμα. Η μεγάλη διακύμανση των γραμμών για ορισμένες θέσεις κατά μήκος του οριζόντιου άξονα υποδηλώνει ότι μια υπερ-αναπαριστώμενη αλληλουχία μολύνει το σύνολο των προς αλληλούχιση περιοχών. Ωστόσο αν αυτό παρατηρείται στο άκρο 5 ή 3 πρόκειται πιθανότατα για προσαρμογείς (adapters) των γονιδίων.

- *Η ανάλυση των εμπλουτισμένων πενταμερών (Enriched 5mers)*

  Η ανάλυση των πενταμερών εξετάζει τον εμπλουτισμό των πεντα-νουκλεοτιδίων. Ο εμπλουτισμός των πενταμερών υπολογίζεται ως ο λόγος με αριθμητή τις παρατηρούμενες συχνότητες των πενταμερών και παρονομαστή τις αναμενόμενες. Η αναμενόμενη συχνότητα υπολογίζεται από την πιθανότητα εμφάνισης των εμπειρικών νουκλεοτιδίων που αποτελούν το πενταμερές. Για παράδειγμα η πιθανότητα εμφάνισης του πενταμερούς CCCCC δεδομένου ότι οι κυτοσίνες έχουν παρατηρηθεί στο 20% των εξεταζόμενων αλληλουχιών, είναι 0.2^5. Ακόμη, αξίζει να σημειωθεί ότι τα πενταμερή που περιέχουν διφορούμενες βάσεις (οτιδήποτε διαφορετικό από A/T/C/G) αγνοούνται.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Η ανωτέρω ανάλυση υπολογίζει την κάλυψη και τον εμπλουτισμό σε εκατοστιαία αναλογία κάθε πενταμερούς για κάθε θέση νουκλεοτιδικής βάσης. Η απεικόνιση της ανάλυσης πραγματοποιείται με διαγράμματα στα οποία παρουσιάζονται τα 5 πρώτα σε ποσότητα εμπλουτισμού πενταμερή. Επιπλέον, η απεικόνιση αυτή αποκαλύπτει την πιθανότητα ύπαρξης προκατάληψης (bias) σε ορισμένες θέσεις κατά μήκος του οριζόντιου άξονα x η οποία μπορεί να προέρχεται από διάφορες πηγές, όπως είναι οι μη αφαιρούμενες αλληλουχίες προσαρμογέων (non-trimmed adapter sequences) και οι πολυ-αδενυλιωμένες ουρές (poly A tails) των αλληλουχιών.

4) **Differential Expression**: Ορίζεται ως η δυνατότητα εύρεσης της διαφορικής έκφρασης γονιδίων. Πρόκειται για τη δυνατότητα ανάλυσης της διαφορικής έκφρασης γονιδίων μεταξύ 2 διαφορετικών βιολογικών συνθηκών (π.χ. Smyd3KO vs Wild Type δείγματα) μετά από στατιστική ανάλυση.

5) **External Annotation Databases**: Είναι η σύνδεση των αποτελεσμάτων της διαφορικής γονιδιακής έκφρασης με εξωτερικές βάσεις δεδομένων.

6) **Gene Ontology Enrichment Analysis**: Αφορά στη δυνατότητα ανάλυσης εμπλουτισμού οντολογιών που περιλαμβάνουν τα διαφορικά εκφρασμένα γονίδια.

7) **Biochemical Pathway Analysis**: Αναφέρεται στη δυνατότητα ανάλυσης εμπλουτισμού βιοχημικών μονοπατιών που απαρτίζουν τα διαφορικά εκφρασμένα γονίδια.

8) **Workflows Creations**: Ονομάζεται η δυνατότητα δημιουργίας ροών/διαγραμμάτων εργασίας.

9) **Clustering Analysis**: Η διαδικασία της αλληλούχισης νέας γενιάς περιλαμβάνει την επεξεργασία μεγάλου όγκου μετρήσεων. Ο μεγάλος αριθμός γονιδίων όπως και η πολυπλοκότητα των NGS μεθόδων καθώς και τα βιολογικά δίκτυα αποτελούν παράγοντες που δυσχεραίνουν σημαντικά την κατανόηση και την ερμηνεία της προκύπτουσας μάζας δεδομένων. Ένα πρώτο βήμα για την αντιμετώπιση αυτής της πρόκλησης είναι η χρήση τεχνικών συσταδοποίησης (clustering methods) οι οποίες είναι απαραίτητες στη διαδικασία εξόρυξης δεδομένων (data mining). Η ανάλυση κατά συστάδες χρησιμοποιείται για το διαχωρισμό ενός δοθέντος συνόλου δεδομένων σε ομάδες βάσει συγκεκριμένων χαρακτηριστικών έτσι ώστε κάθε μέλος του γκρουπ (γονίδια, δείγματα) να έχει παρόμοια χαρακτηριστικά μεταξύ των μελών του και διαφορετικά από τα μέλη των άλλων γκρουπ.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

10) **Custom Reports**: Εμπεριέχει τη δυνατότητα έκδοσης αναφορών επιλεγμένων γονιδίων. Αναλυτικότερα, πρόκειται για την επιλογή μικρού αριθμού γονιδίων και τη δημιουργία παραπομπής σε βάσεις δεδομένων για τα συγκεκριμένα γονίδια.

Στην τρίτη κατηγορία που αφορά στις δυνατότητες οπτικοποίησης δεδομένων περιλαμβάνονται τα επόμενα 5 χαρακτηριστικά:

1) **Heat maps**: Αποτελεί τη δυνατότητα αναπαράστασης των δεδομένων κατά συστάδες με τη μορφή χαρτών θερμότητας στους οποίους οι τιμές δεδομένων αναπαρίστανται με χρώματα. Σε έναν χάρτη θερμότητας, κάθε γραμμή στον οριζόντιο άξονα αντιστοιχεί σε ένα χαρακτηριστικό (γονίδιο, χρωμόσωμα) ενώ κάθε στήλη στον κάθετο άξονα σε ένα από τα εξεταζόμενα κάθε φορά δείγματα. Το χρώμα που χρησιμοποιείται για κάθε σημείο του heat map αντικατοπτρίζει το επίπεδο έκφρασης του χαρακτηριστικού για κάθε δείγμα.

2α) **Volcano plots:** Πρόκειται για γραφήματα απεικόνισης της γραφικής παράστασης $-10\log_{10}P$ vs fold change (διαφορική έκφραση) που χρησιμοποιείται για την ανάλυση δεδομένων υψηλής απόδοσης και συνιστά επισκόπηση των γονιδίων-στόχων του εκάστοτε ερευνητή.

2β) **MA plots**: Είναι γραφήματα τα οποία απεικονίζουν τη διαφορά της γονιδιακής έκφρασης μεταξύ 2 δειγμάτων. Ο οριζόντιος άξονας παριστάνει το μέσο όρο έκφρασης ενός γονιδίου σε 2 διαφορετικές βιολογικές συνθήκες ενώ ο κάθετος άξονας τη διαφορά μεταξύ των δειγμάτων στα επίπεδα γονιδιακής έκφρασης. Στην ουσία, πρόκειται για μια οριζόντια γραμμή στην οποία τα διαφορετικά εκφρασμένα γονίδια βρίσκονται πάνω ή κάτω από την μηδενική τιμή του μέσου όρου.

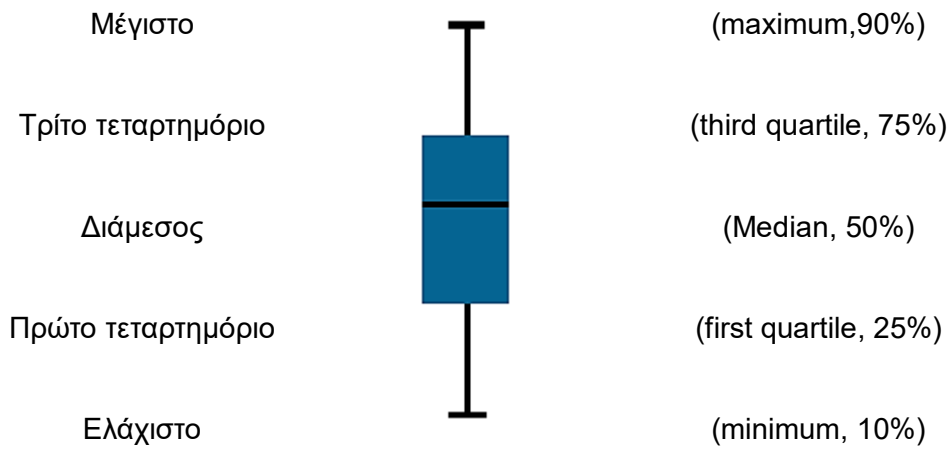
3) **Genome Browser**: Αναφέρεται στη δυνατότητα εκτέλεσης προγραμμάτων περιήγησης γονιδιωμάτων. Στην ουσία είναι προγράμματα απεικόνισης γονιδιωματικών δεδομένων τα οποία προσφέρουν ταυτόχρονα την επισκόπηση, τη σύγκριση καθώς και την ανάλυση των πληροφοριών που προκύπτουν από αυτά.

4α) **PCA plots**: Πρόκειται για γραφικές παραστάσεις οι οποίες προκύπτουν ύστερα από μια στατιστική διαδικασία (dimension reduction) για τη μετατροπή ενός μεγάλου συνόλου παρατηρήσεων/ μεταβλητών σε ένα μικρό σύνολο παρατηρήσεων/μεταβλητών μη γραμμικά συσχετισμένων οι οποίες ονομάζονται κύριες συνιστώσες.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

4β) **MDS plots**: Είναι γραφήματα πολυδιάστατης κλιμάκωσης τα οποία συνιστούν ένα μέσο απεικόνισης του επιπέδου ομοιότητας των ξεχωριστών περιπτώσεων ενός συνόλου δεδομένων. Ουσιαστικά τα γραφήματα είναι παρόμοια με αυτά της ανάλυσης κύριων συνιστωσών με τη διαφορά ότι χρησιμοποιούνται απόλυτες τιμές στη μέτρηση της απόστασης, όπως είναι η Ευκλείδεια απόσταση, αντί για τη διακύμανση των μεταβλητών στην οποία βασίζεται η ανάλυση κύριων συνιστωσών.

5) **Box plots**: Πρόκειται για διαγράμματα πέντε σημείων στα οποία ο οριζόντιος άξονας αναφέρεται στα ονόματα των δειγμάτων ενώ ο κάθετος στα επίπεδα γονιδιακής έκφρασης των δειγμάτων αυτών. Τα πέντε σημεία προέρχονται από την καμπύλη κατανομής των βιολογικών δεδομένων και είναι τα εξής: Το μέγιστο (maximum), το τρίτο τεταρτημόριο (third quartile), η διάμεσος (median), το πρώτο τεταρτημόριο (first quartile) και το ελάχιστο (minimum).Ένα τυπικό box plot διάγραμμα φαίνεται στο Σχήμα 2.1 που ακολουθεί.



Σχήμα 2.1: Η δομή ενός τυπικού box plot διαγράμματος

Σύμφωνα με τα αποτελέσματα της διαδικτυακής έρευνας επτά συνολικά λογισμικά πληρούν όλα ή σχεδόν όλα τα προαναφερθέντα χαρακτηριστικά και παρέχονται από εταιρείες εκ των οποίων πέντε έχουν την έδρα τους στις Ηνωμένες Πολιτείες Αμερικής, μία στη Γερμανία και μία στη Νέα Ζηλανδία. Η εμπορική ονομασία του λογισμικού, η επωνυμία της εταιρείας καθώς και η έδρα αυτής για καθένα από τα 7 λογισμικά φαίνονται στον Πίνακα 2.1 που ακολουθεί.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Πίνακας 2.1: Έδρες εταιρειών

Λογισμικό	Εταιρεία	Έδρα εταιρείας
1. CLC Genomics Workbench	CLC bio, Qiagen	ΓΕΡΜΑΝΙΑ
2. Full Lasergene Suite	DNASTAR	ΗΠΑ
3. Geneious	Biomatters	ΝΕΑ ΖΗΛΑΝΔΙΑ
4. Nexus Expression	Biodiscovery	ΗΠΑ
5. Partek Flow	Partek	ΗΠΑ
6. Rosalind	OnRamp Bioinformatics	ΗΠΑ
7. Strand NGS	Strand Scientific Intelligence	ΗΠΑ

Αξίζει να σημειωθεί ότι όσον αφορά στην πρώτη κατηγορία χαρακτηριστικών, οι πληροφορίες για τα λογισμικά αντλήθηκαν αποκλειστικά από τις ιστοσελίδες των εταιρειών ενώ τα στοιχεία που αφορούσαν στις άλλες δύο κατηγορίες συγκεντρώθηκαν από πέντε συνολικά πηγές. Αναλυτικότερα, οι πηγές αυτές περιλαμβάνουν : α) Το εγχειρίδιο (manual), β) την ιστοσελίδα (site), γ) την επικοινωνία μέσω email με τους υπεύθυνους των εταιρειών, δ) τη δυνατότητα online συνεδρίας παρουσίασης των χαρακτηριστικών και ε) τη δυνατότητα δοκιμής σε υπολογιστή ορισμένων από τα 18 χαρακτηριστικά που παρουσιάσθηκαν παραπάνω.

Επιπρόσθετα, τα ελάχιστα απαιτούμενα χαρακτηριστικά που πρέπει να διαθέτει ένας υπολογιστής προκειμένου να εκτελέσει οποιαδήποτε από τα επτά λογισμικά φαίνονται στον Πίνακα 2.2 ενώ τα χαρακτηριστικά του προσωπικού υπολογιστή δοκιμής από τον οποίο προέρχονται τα στιγμιότυπα (snapshots) των λογισμικών φαίνονται στον Πίνακα 2.3.

Πίνακας 2.2: Ελάχιστα Απαιτούμενα Χαρακτηριστικά

Λειτουργικό Σύστημα (Operating System)	Windows/Mac/Linux 64 bit
Μνήμη RAM	4 GB
Ανάλυση Οθόνης	1024 x 768 pixels

Πίνακας 2.3: Χαρακτηριστικά Προσωπικού Υπολογιστή Δοκιμής

Λειτουργικό Σύστημα	Windows 10 Home
Μνήμη RAM	8 GB
Ανάλυση Οθόνης	1366 x 768 pixels

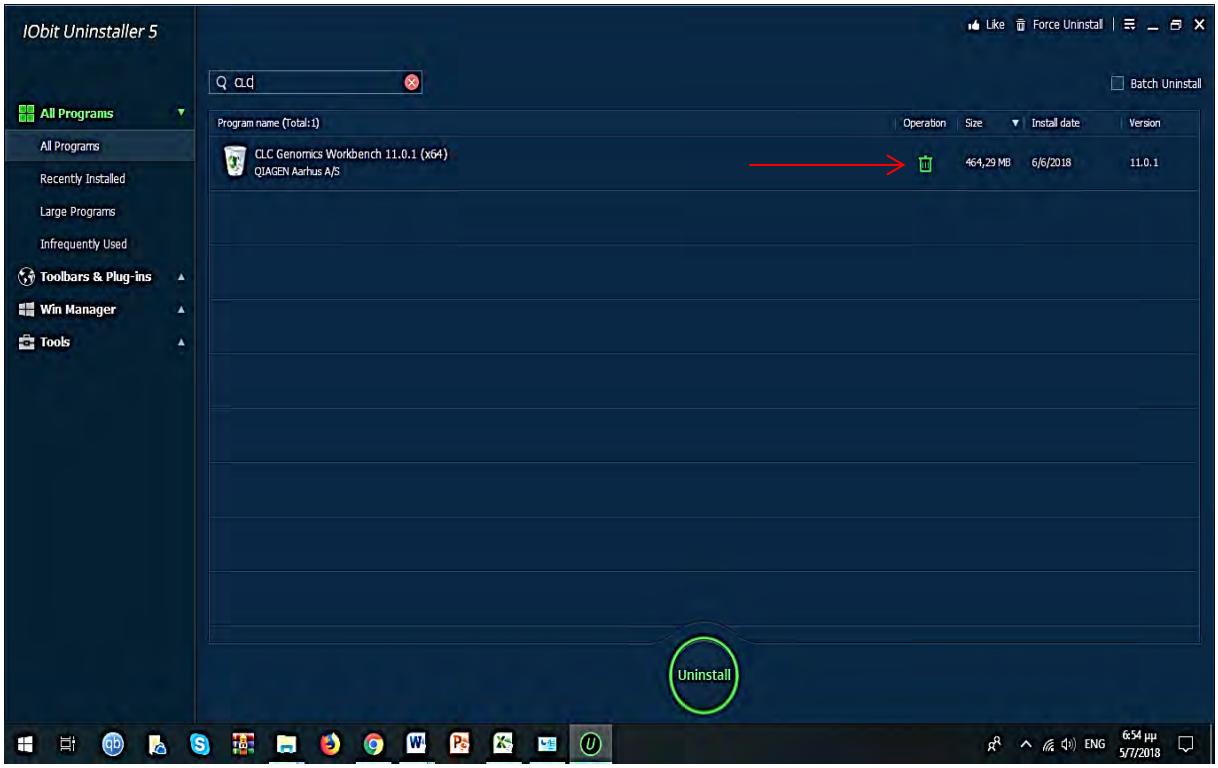
Ακόμη, αξίζει να αναφερθεί ότι τα αποτελέσματα της παρούσας διπλωματικής εργασίας αναμένεται να αξιοποιηθούν από την εταιρεία HybridSTAT η οποία δραστηριοποιείται στο Ηνωμένο Βασίλειο καθώς και στην Ελλάδα στο χώρο της προγνωστικής ανάλυσης. Η HybridSTAT ιδρύθηκε το 2014 σύμφωνα με την ιστοσελίδα της και προσφέρει υπηρεσίες,
μέσω 2 εφαρμογών που έχει ήδη αναπτύξει, οι οποίες αφορούν τους τομείς της βιοστατιστικής, της βιοπληροφορικής και της ανάλυσης βιολογικών δεδομένων υψηλής απόδοσης.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.1 CLC Genomics Workbench (ClC bio, Qiagen)

Πρώτη κατηγορία χαρακτηριστικών

Στην Εικόνα 2.1.1 που προέρχεται από τον υπολογιστή δοκιμής φαίνεται ότι ο χώρος που καταλαμβάνει το λογισμικό CLC Genomics Workbench είναι 465 MB περίπου. Η αρχιτεκτονική συστήματος είναι Desktop που σημαίνει ότι οποιοσδήποτε επιθυμεί να εκτελέσει το συγκεκριμένη λογισμικό θα πρέπει να την κατεβάσει σε έναν υπολογιστή. Με άλλα λόγια, το λογισμικό αυτό δεν είναι δυνατό να εκτελεστεί σε οποιονδήποτε υπολογιστή επειδή η άδεια χρήσης του προορίζεται για έναν κάθε φορά υπολογιστή σύμφωνα με τους όρους εγκατάστασης του· κάτι τέτοιο είναι εφικτό μόνο στην περίπτωση που η αρχιτεκτονική του συστήματος είναι Cloud.



Εικόνα 2.1.1: Χωρητικότητα του λογισμικού CLC Genomics Workbench

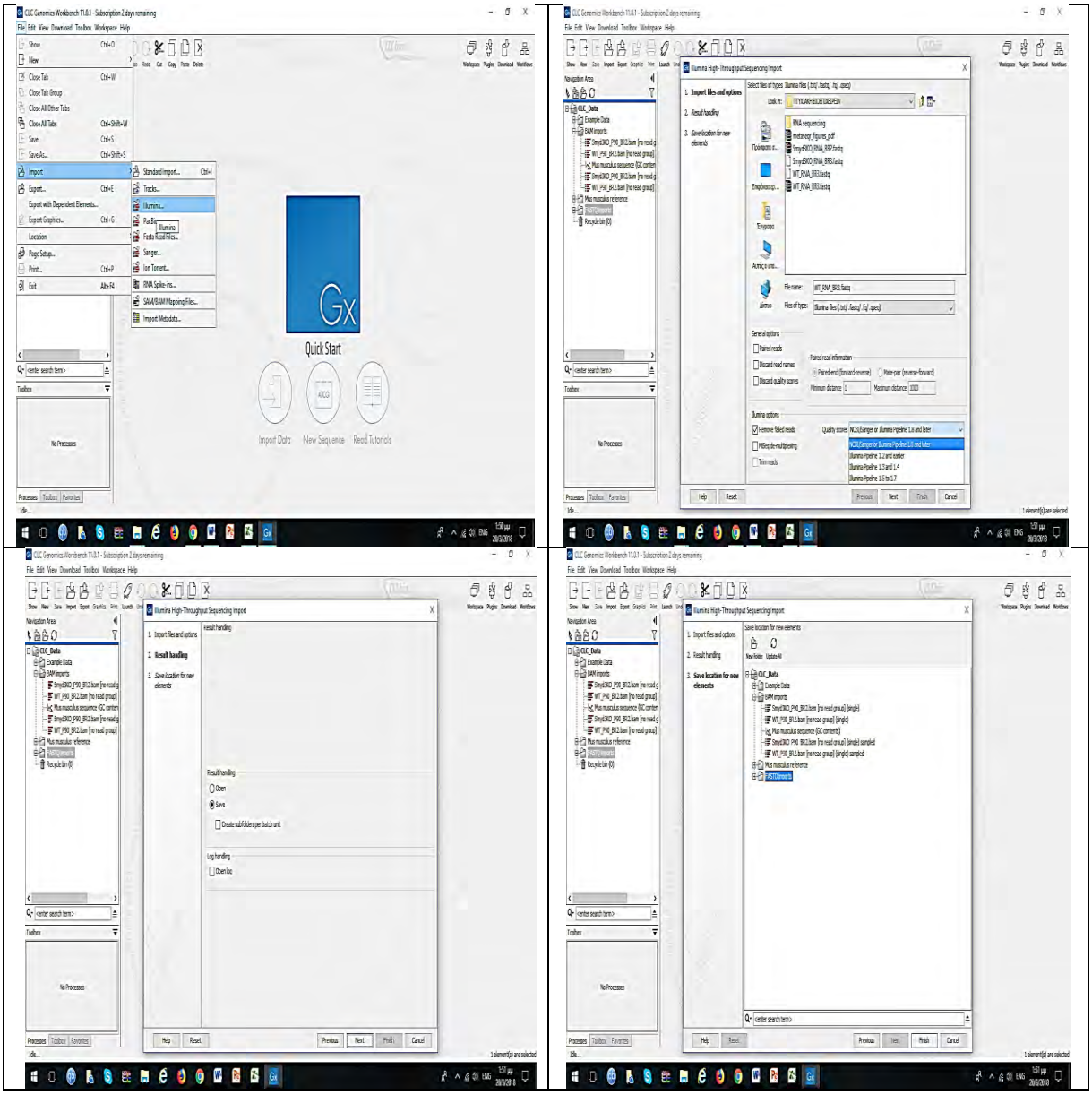
Όσον αφορά στο λειτουργικό σύστημα, το λογισμικό υποστηρίζει και τους 3 τύπους συστημάτων (Windows/Mac/Linux 64 bit) σύμφωνα με την ιστοσελίδα της γερμανικής εταιρείας.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Δεύτερη κατηγορία χαρακτηριστικών

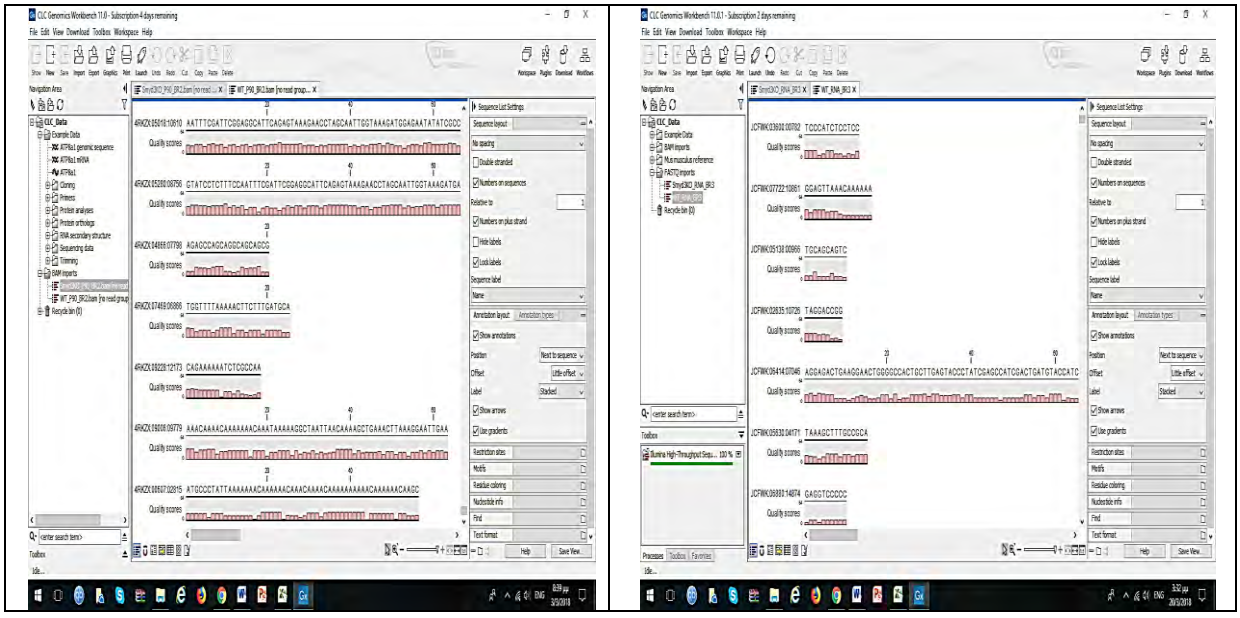
- FASTQ / BAM αρχεία

Στην Εικόνα 2.1.2 που ακολουθεί φαίνονται τα βήματα που απαιτούνται προκειμένου να γίνει η εισαγωγή κάποιου αρχείου FASTQ ή BAM στο λογισμικό. Τα FASTQ ή BAM αρχεία προέρχονται από τον εξεταζόμενο οργανισμό που είναι ο mus musculus. Στην Εικόνα 2.1.3 απεικονίζονται ενδεικτικά κάποια από τα αποτελέσματα που προέκυψαν από την παραπάνω διαδικασία.



Εικόνα 2.1.2: Εισαγωγή FASTQ αρχείου στο CLC Genomics Workbench

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.3: Ενδεικτικά αποτελέσματα μετά την εισαγωγή BAM αρχείου στην αριστερή στήλη και FASTQ αρχείου στη δεξιά στήλη

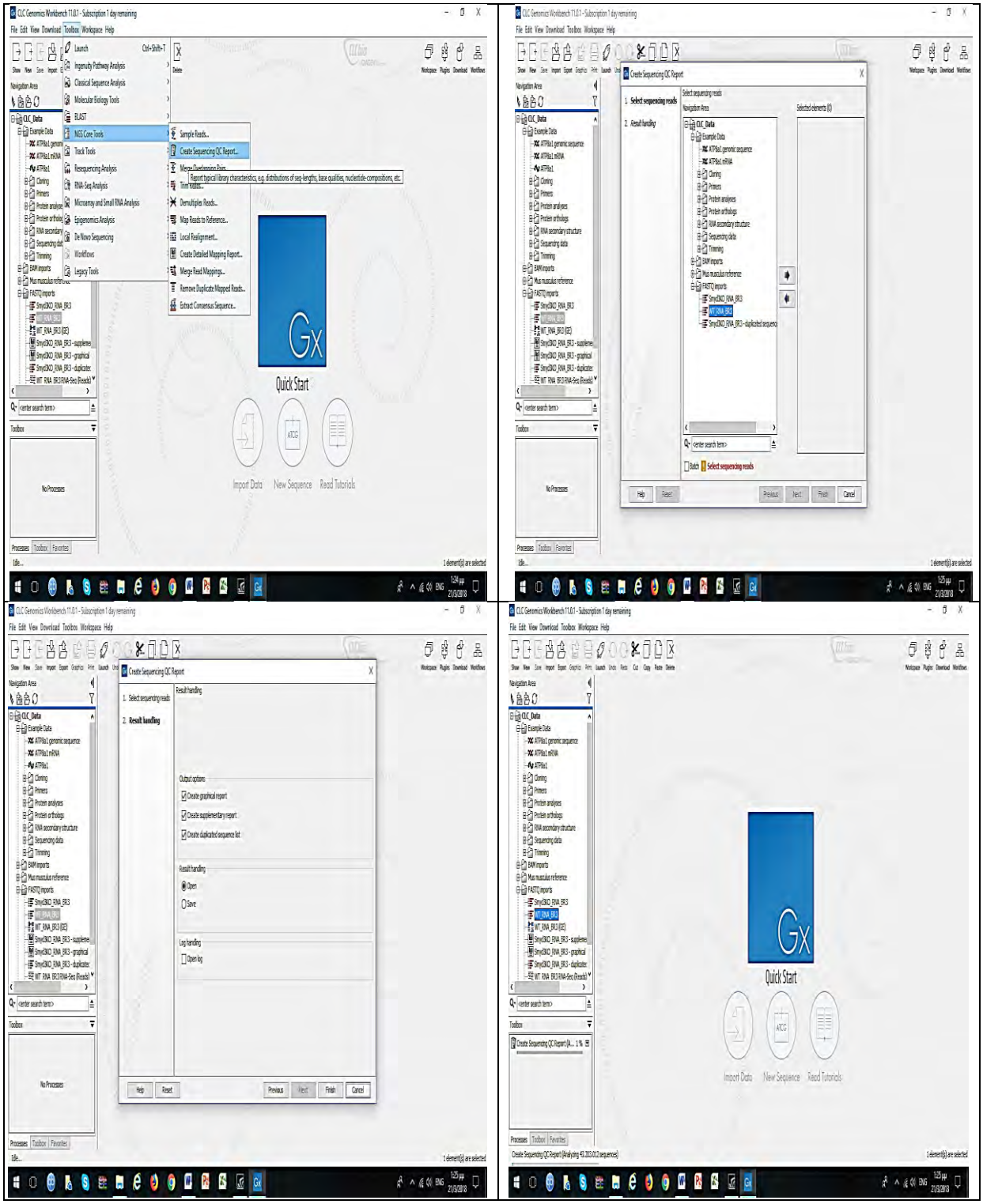
- Organisms

Τα γονιδιώματα όλων των προκαρυωτικών και ευκαρυωτικών οργανισμών συμπεριλαμβανομένων των θηλαστικών, φυτών και ζώων υποστηρίζονται στο συγκεκριμένο λογισμικό σύμφωνα με τους υπεύθυνους της εταιρείας ύστερα από επικοινωνία μαζί τους μέσω email.

- Quality Control

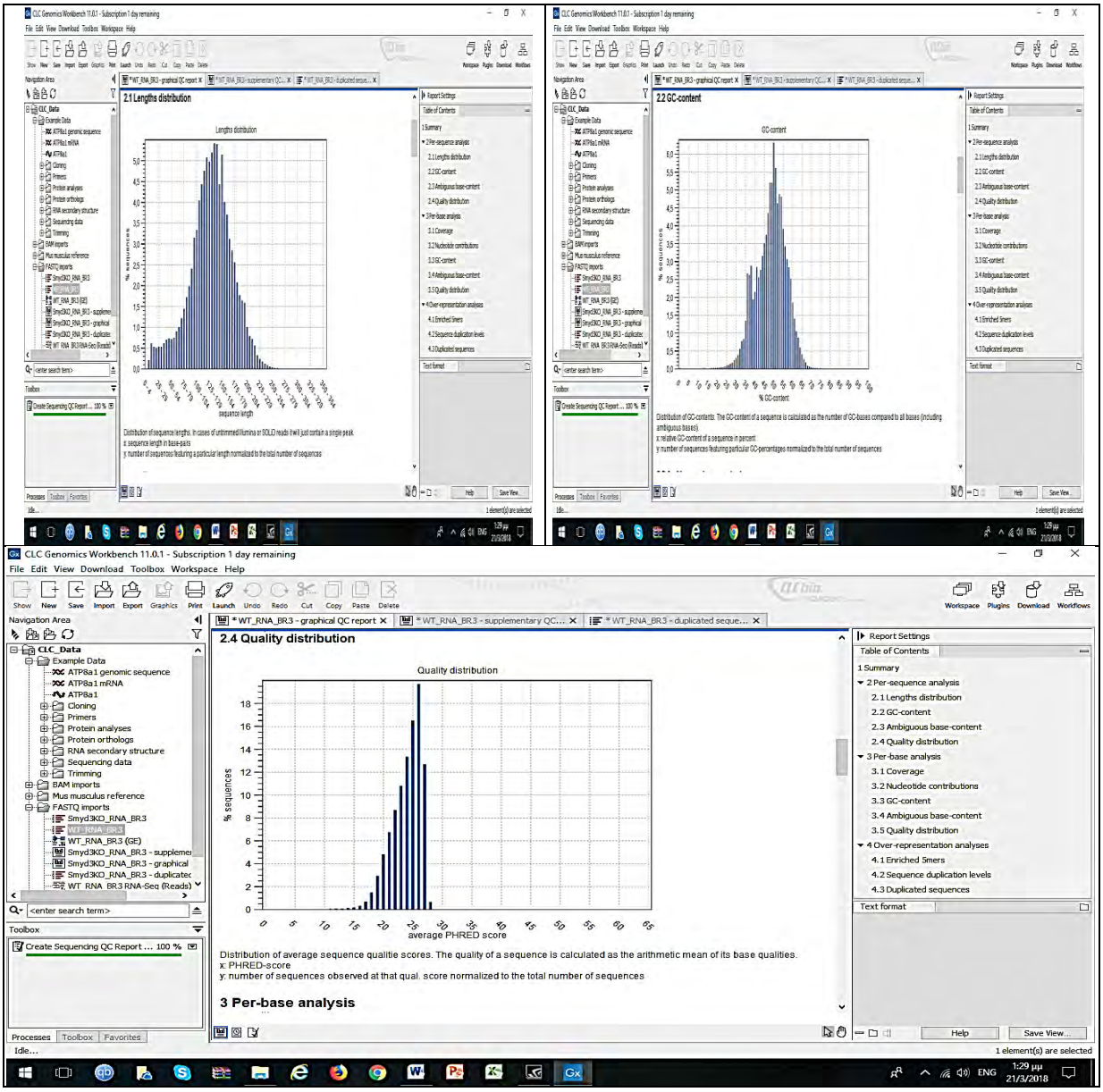
Τα βήματα εκτέλεσης του ποιοτικού ελέγχου που προσφέρει το συγκεκριμένο λογισμικό παρουσιάζονται στην Εικόνα 2.1.4. Στις Εικόνες 2.1.5, 2.1.6 και 2.1.7 παρουσιάζονται τα αποτελέσματα του ποιοτικού ελέγχου.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.4: Τα 4 βήματα εκτέλεσης του ποιοτικού ελέγχου

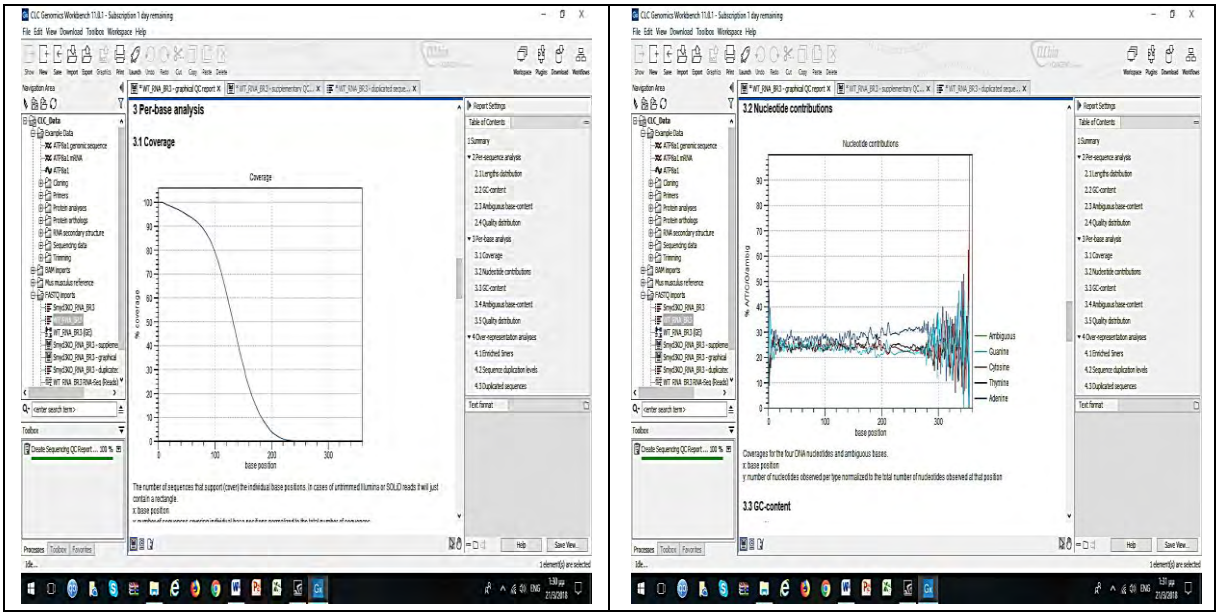
Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.5: Ποιοτικός έλεγχος που αφορά στην κατανομή μήκους, στην περιεκτικότητα των νουκλεοτιδικών βάσεων σε Γουανίνη-Κυτοσίνη και στην κατανομή της ποιότητας των αλληλουχηθέντων ολιγονουκλεοτιδίων

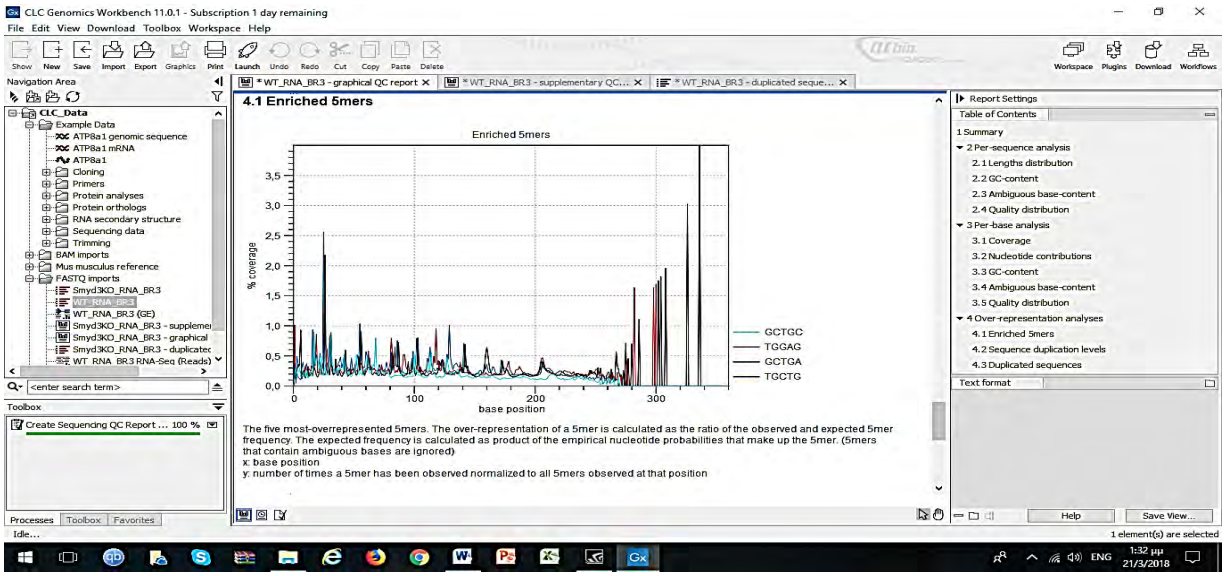
Από το πρώτο διάγραμμα της Εικόνας 2.1.5 προκύπτει ότι το μέσο μήκος των αλληλουχιών ανέρχεται στις 150-154 βάσεις. Το δεύτερο διάγραμμα της ίδιας εικόνας δείχνει ότι η κατανομή της περιεκτικότητας των νουκλεοτιδικών βάσεων σε Γουανίνη-Κυτοσίνη αποκλίνει ελαφρώς από την κανονική κατανομή. Στο τρίτο διάγραμμα φαίνεται ότι η υψηλότερη κορυφή αντιστοιχεί σε PHRED σκορ ίσο με 26. Η τιμή αυτή υπολείπεται του φυσιολογικού ορίου/σκορ που είναι το 40 και αυτό πιθανότατα οφείλεται σε σφάλματα που σχετίζονται με τη διαδικασία αλληλούχισης των δειγμάτων στο εργαστήριο.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.6: Ποιοτικός έλεγχος της κάλυψης των βάσεων και της συνεισφοράς των νουκλεοτιδίων

Από το διάγραμμα της αριστερής στήλης της Εικόνας 2.1.6 φαίνεται ότι το ποσοστό των ανεξάρτητων αλληλουχίσεων για κάθε θέση βάσης μειώνεται κατά 20% μετά τις πρώτες 100 βάσεις, φτάνει στο 50% στις 140 περίπου βάσεις ενώ μηδενίζεται μετά τις 220 περίπου βάσεις. Το δεύτερο διάγραμμα δείχνει την ύπαρξη ελάχιστης διαφοράς μεταξύ των τεσσάρων νουκλεοτιδικών βάσεων η οποία εντοπίζεται κυρίως στην υψηλότερη τροχιά της αδενίνης σε σχέση με τις τροχιές των υπόλοιπων βάσεων που βρίσκονται σχεδόν στο ίδιο επίπεδο.



Εικόνα 2.1.7: Ποιοτικός έλεγχος των εμπλουτισμένων πενταμερών (Enriched 5mers)

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Από το διάγραμμα της Εικόνας 2.1.7 διακρίνονται τέσσερα περισσότερο εμπλουτισμένα πενταμερή που είναι τα GCTGC, TGGAG, GCTGA, TGCTG. Οι καμπύλες αυτών των πενταμερών φαίνονται παρόμοιες και το κοινό τους σημείο είναι η ύπαρξη φασματοσκοπικού θορύβου μετά τις 300 περίπου βάσεις. Το φαινόμενο αυτό γίνεται αντιληπτό από τη μεγάλης συχνότητα εμφάνισης των κορυφών των οποίων το ύψος ανέρχεται σε υψηλότερα επίπεδα στον κάθετο άξονα σε σχέση με αυτό των προηγούμενων κορυφών.

- Differential Expression

Το χαρακτηριστικό αυτό υπάρχει στο λογισμικό της Qiagen σύμφωνα με ενημέρωση των επιστημονικών υπευθύνων της εταιρείας μέσω email. Ωστόσο δεν κατέστη εφικτή η ολοκλήρωση της όλης διαδικασίας λόγω ανεπάρκειας σε μνήμη του υπολογιστή δοκιμής και για αυτό το λόγο δεν παρατίθενται μία ή περισσότερες εικόνες.

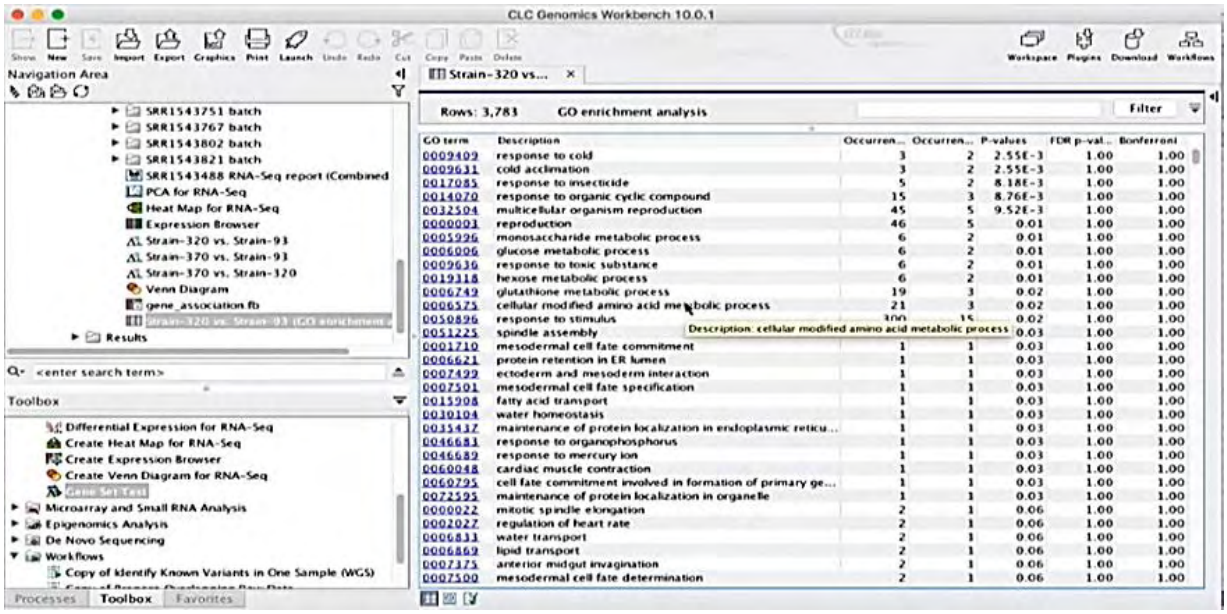
- External Annotation Databases

Οι εξωτερικές βάσεις δεδομένων με τις οποίες συνδέεται το συγκεκριμένο λογισμικό είναι η Ensembl και η RefSeq σύμφωνα με την απάντηση που δόθηκε μέσω email από εκπρόσωπο της Qiagen.

- Gene Ontology Enrichment Analysis

Αν και γίνεται αναφορά στην ανάλυση των οντολογιών των διαφορικά εκφρασμένων γονιδίων στο manual που συνοδεύει το λογισμικό, εκτιμάται ότι δεν είναι αρκετά κατανοητή για έναν αρχάριο υπολογιστικά χρήστη ώστε να εκτελέσει με επιτυχία την όλη διαδικασία. Ωστόσο, η Εικόνα 2.1.8 παρουσιάζει ενδεικτικά το αποτέλεσμα αυτής της ανάλυσης όπως φαίνεται στο manual του CLC Genomics Workbench.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.8: Στιγμιότυπο ανάλυσης εμπλουτισμού οντολογιών από το manual της CLC Genomics Workbench

- Biochemical Pathway Analysis

Η ανάλυση εμπλουτισμού των βιοχημικών μονοπατιών δεν παρέχεται από τη βασική έκδοση του λογισμικού CLC Genomics Workbench αλλά απαιτείται η λήψη ενός άλλου λογισμικού που ονομάζεται Ingenuity Pathway Analysis (IPA).

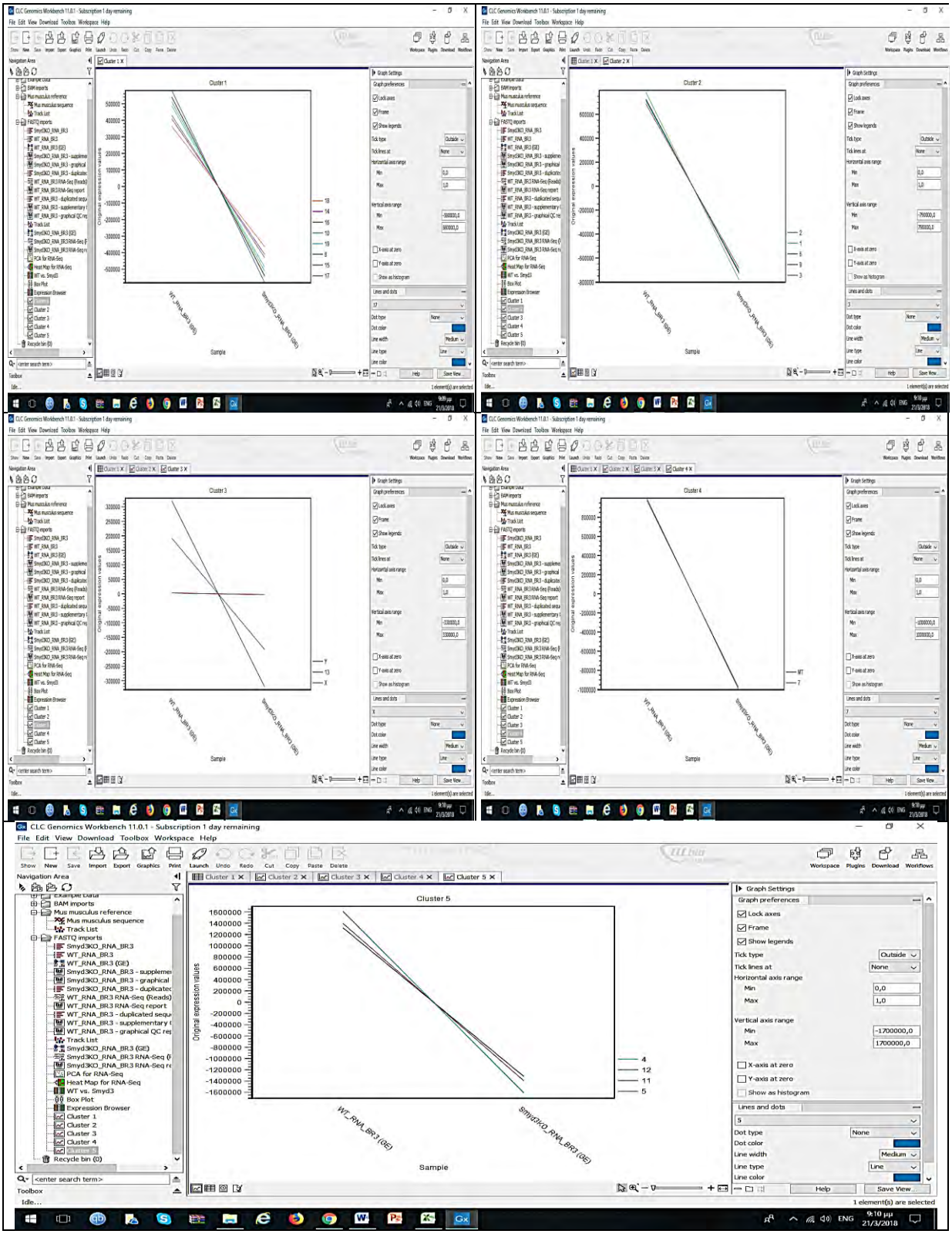
- Workflows Creations

Η δημιουργία διαγραμμάτων/ροών εργασιών προσφέρεται σύμφωνα με τους ιθύνοντες από το λογισμικό. Ωστόσο δεν κατέστη εφικτή η δημιουργία ενός τέτοιου διαγράμματος καθώς θεωρήθηκε ότι ήταν ιδιαίτερα δύσκολη και περίπλοκη η όλη διαδικασία για έναν μη έμπειρο χρήστη.

- Clustering Analysis

Η ανάλυση κατά συστάδες υποστηρίζεται σε εξαιρετικά ικανοποιητικό βαθμό από το λογισμικό της Qiagen. Πιο συγκεκριμένα, η Εικόνα 2.1.9 παρουσιάζει τα διαγράμματα που προέκυψαν από την ανάλυση κατά συστάδες με χρήση του αλγόριθμου K-means. Στα διαγράμματα αυτά απεικονίζεται η συσταδοποίηση των 21 χρωμοσωμάτων (19 αυτοσωμικά και 2 φυλετικά) του ποντικού mus musculus. Πιο συγκεκριμένα, η ανάλυση αυτή δείχνει τα γονίδια των χρωμοσωμάτων του 5^{ου} cluster να έχουν την μεγαλύτερη κλίση όσον αφορά τις τιμές έκφρασης τους σε σχέση με τα υπόλοιπα cluster ενώ μηδενική κλίση φαίνεται να έχουν τα εκφρασμένα γονίδια του φυλετικού χρωμοσώματος Y που είναι τοποθετημένα στο 3^{ο} cluster.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.9: Διαγράμματα από την ανάλυση κατά συστάδες του εξεταζόμενου οργανισμού

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

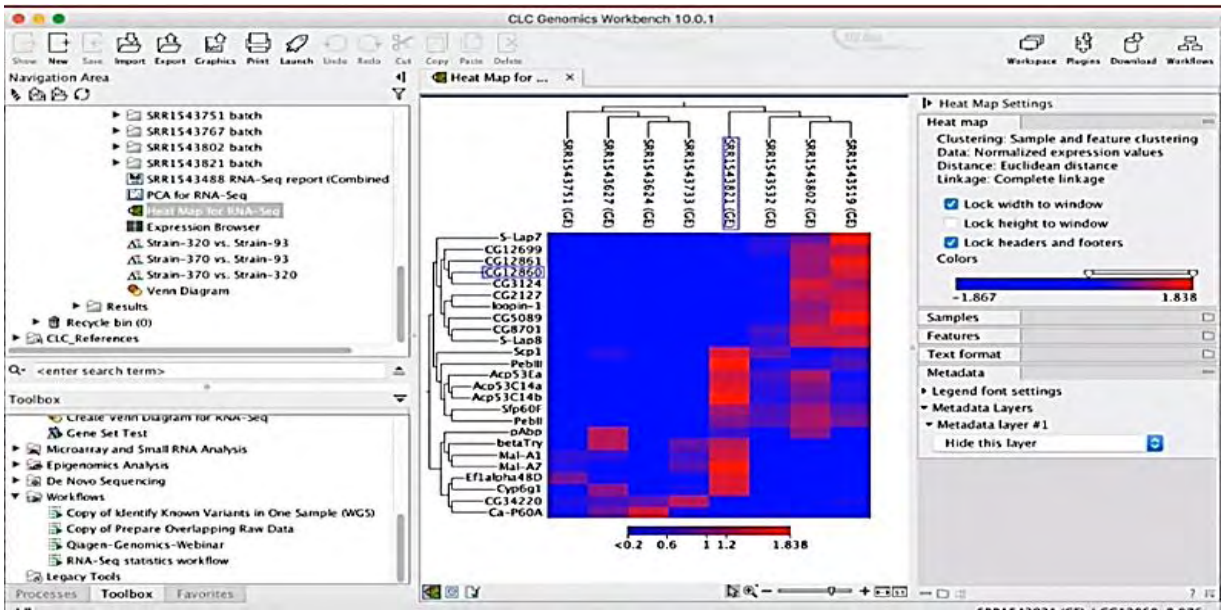
- Custom Reports

Δεν κατέστη δυνατή η παράθεση κάποιας εικόνας που να αφορά στη δυνατότητα του λογισμικού για δημιουργία αναφορών επιλεγμένων από το χρήστη γονιδίων.

Τρίτη κατηγορία χαρακτηριστικών

- Heat maps

Η ανάλυση με χρήση χαρτών θερμότητας είναι μια διαδικασία που προσφέρεται από το λογισμικό. Ένα στιγμιότυπο αποτελέσματος αυτής της ανάλυσης όπως δίνεται από το manual του λογισμικού φαίνεται στην Εικόνα 2.1.10.

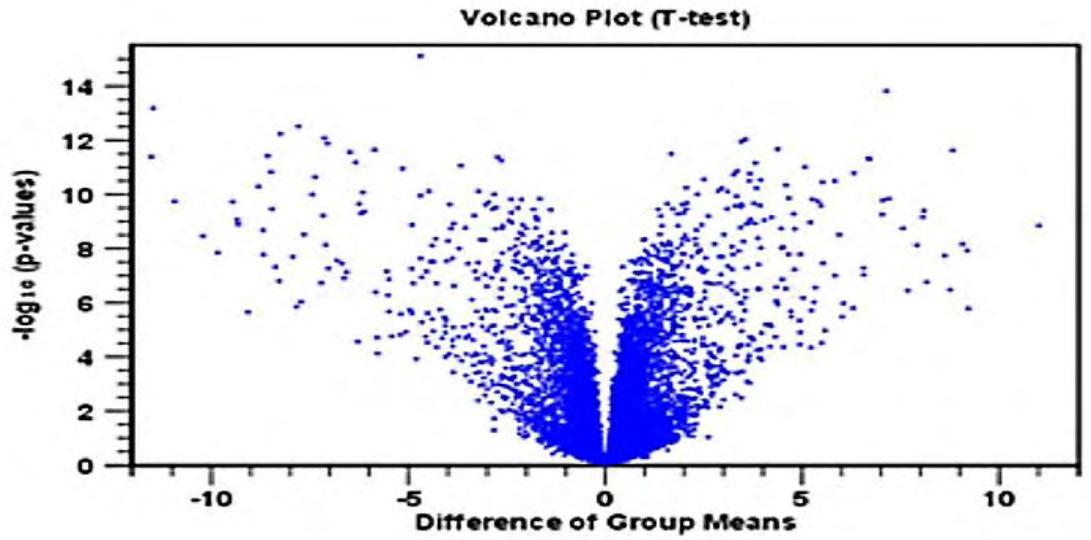


Εικόνα 2.1.10: Στιγμιότυπο ενός Heat map από το manual του λογισμικού

- Volcano plots

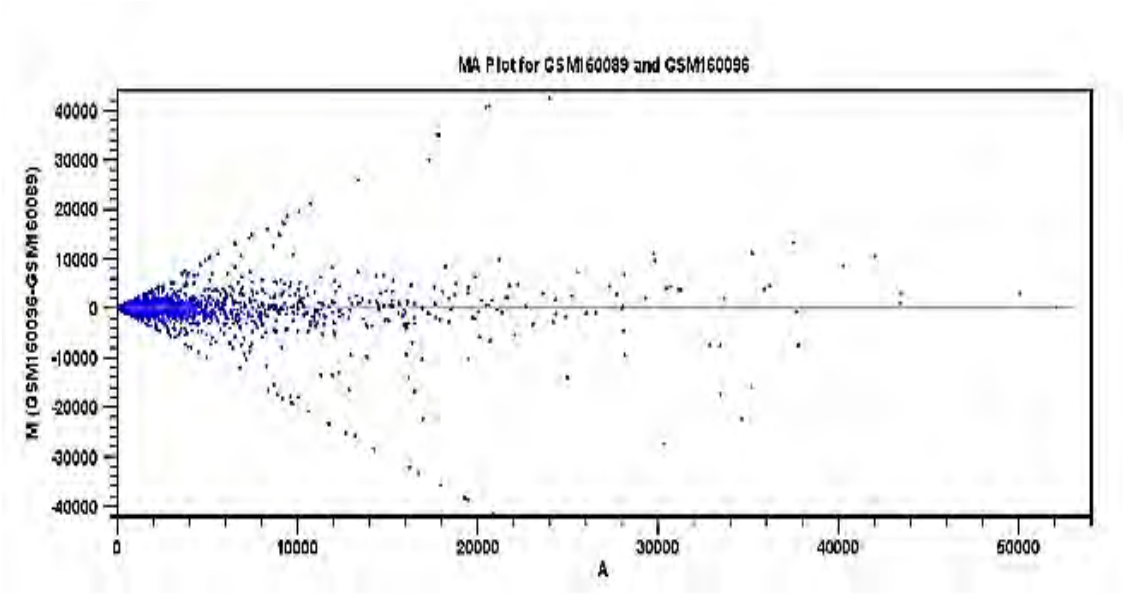
Στην Εικόνα 2.1.11 παρατίθεται μια Volcano γραφική παράσταση που προέρχεται από το εγχειρίδιο του λογισμικού.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.11: Volcano γραφική παράσταση από το manual του λογισμικού

- MA plots: Μία ενδεικτική MA γραφική παράσταση από το manual του λογισμικού απεικονίζεται στην Εικόνα 2.1.12.

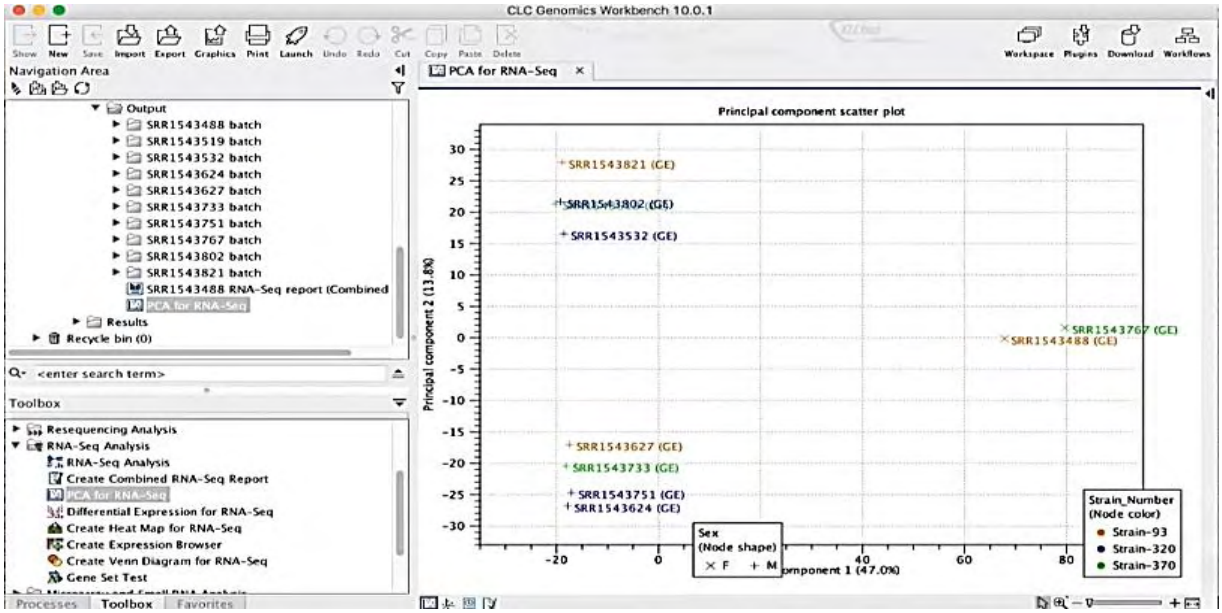


Εικόνα 2.1.12: MA γραφική παράσταση από το CLC Genomics Workbench

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Genome Browser: Σύμφωνα με τους ιθύνοντες της εταιρείας το λογισμικό δίνει τη δυνατότητα περιήγησης γονιδιωμάτων μέσω ενός υπολογιστικού εργαλείου παρόμοιου με το UCSC Genome Browser. Παρ' όλα αυτά δεν κατέστη δυνατός ο εντοπισμός κάποιας τέτοιας επιλογής κατά τη διάρκεια εξερεύνησης του λογισμικού και για αυτό το λόγο δεν παρατίθεται κάποια εικόνα ή στιγμιότυπο από αυτό το χαρακτηριστικό.

- PCA: Στην Εικόνα 2.1.13 παρουσιάζεται ένα ενδεικτικό γράφημα ανάλυσης κύριων συνιστωσών όπως παρατίθεται στο αντίστοιχο τμήμα του εγχειριδίου του λογισμικού.

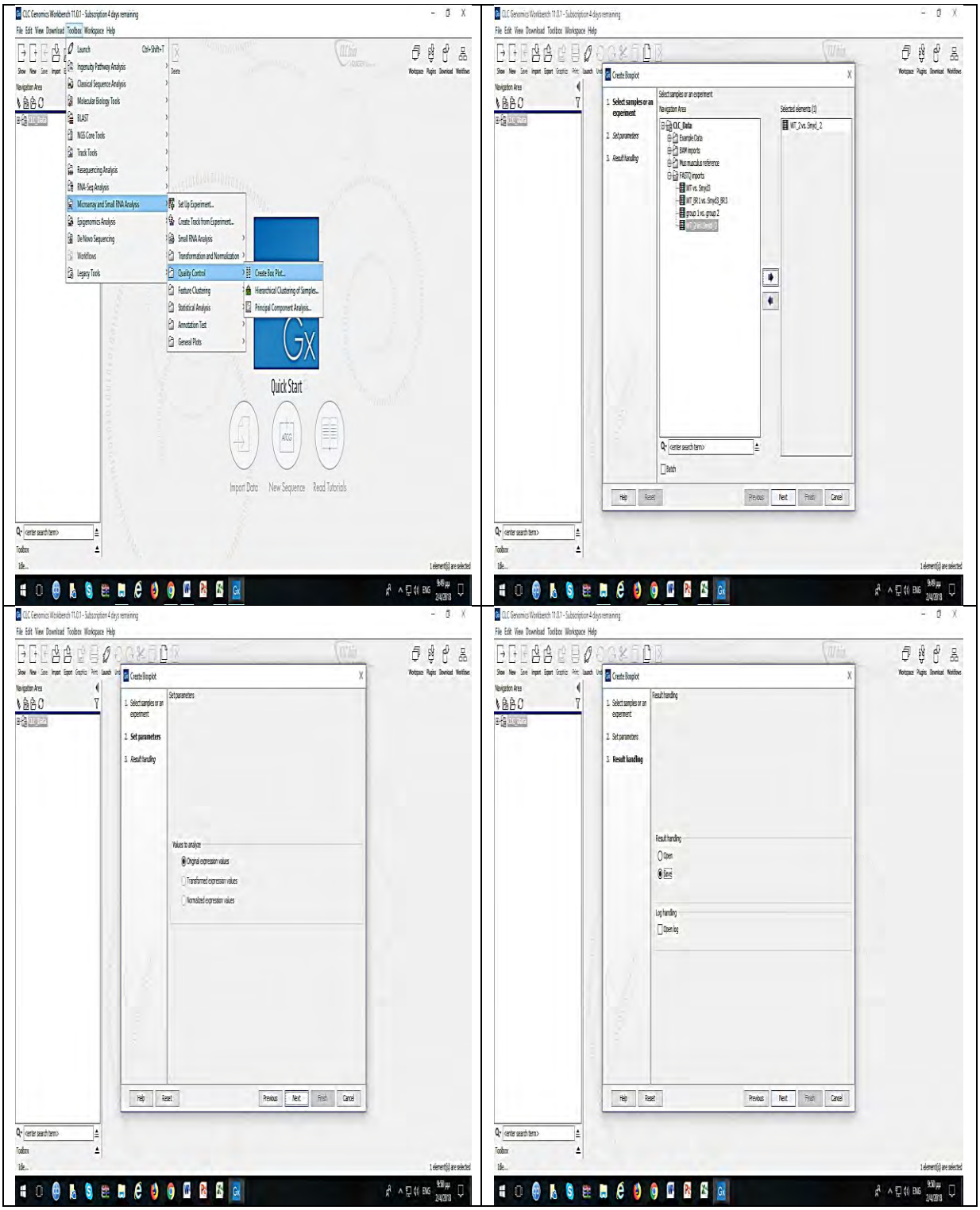


Εικόνα 2.1.13: PCA γράφημα από το manual του CLC Genomics Workbench

- MDS: Η κατασκευή γραφημάτων πολυδιάστατης κλιμάκωσης δεν παρέχεται τουλάχιστον μέχρι στιγμής από το λογισμικό σύμφωνα και με τους υπευθύνους της εταιρείας Qiagen.

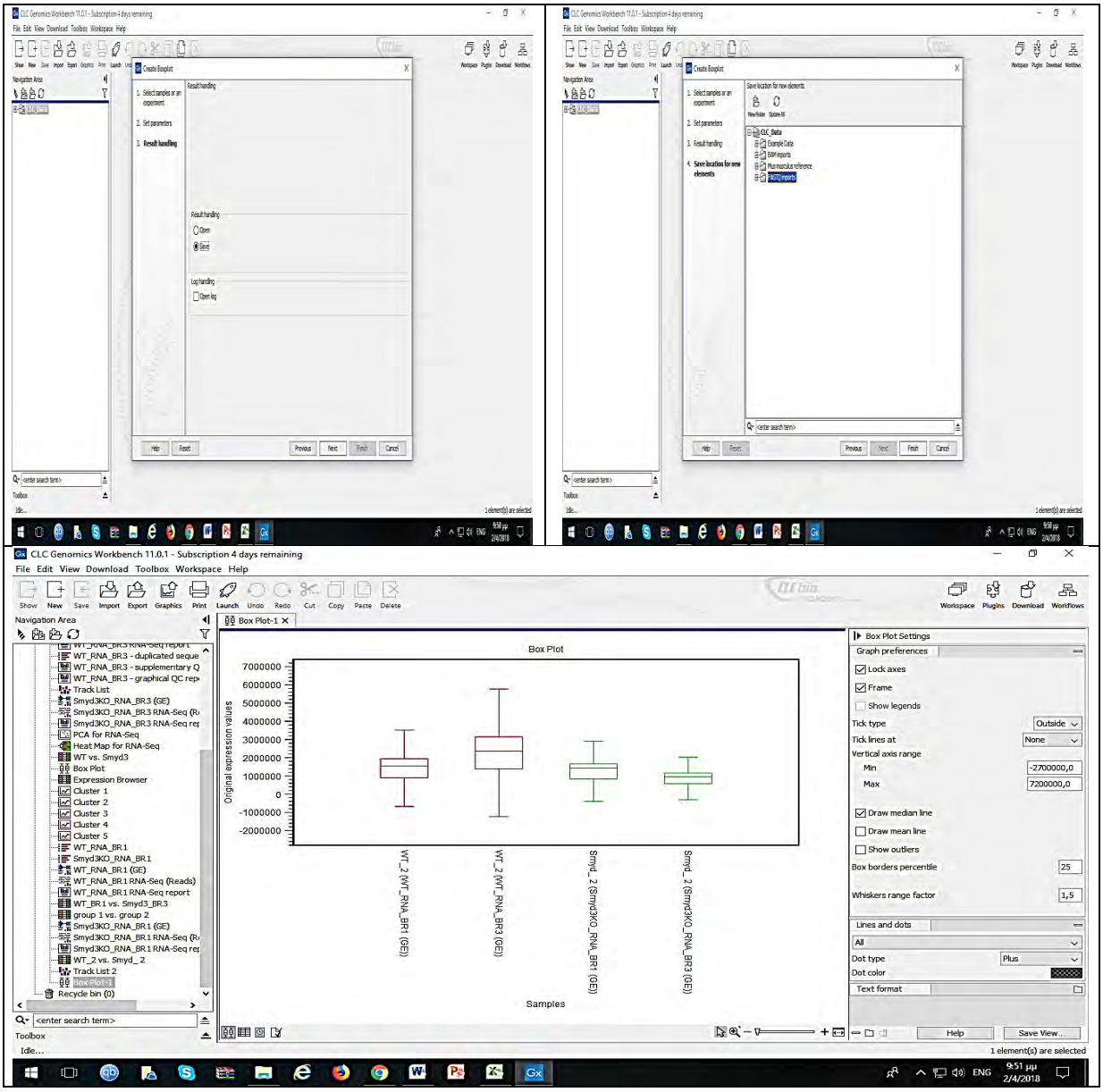
- Box plots: Η δυνατότητα δημιουργίας τέτοιου είδους διαγραμμάτων παρέχεται από το λογισμικό αυτό. Η διαδικασία και ένα παράδειγμα που αφορά 4 από τα δείγματα (2 WT και 2 Smyd3KO) του εξεταζόμενου ποντικού απεικονίζονται στις Εικόνες 2.1.14 και 2.1.15.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.1.14: Διαδικασία δημιουργίας ενός Box plot διαγράμματος

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

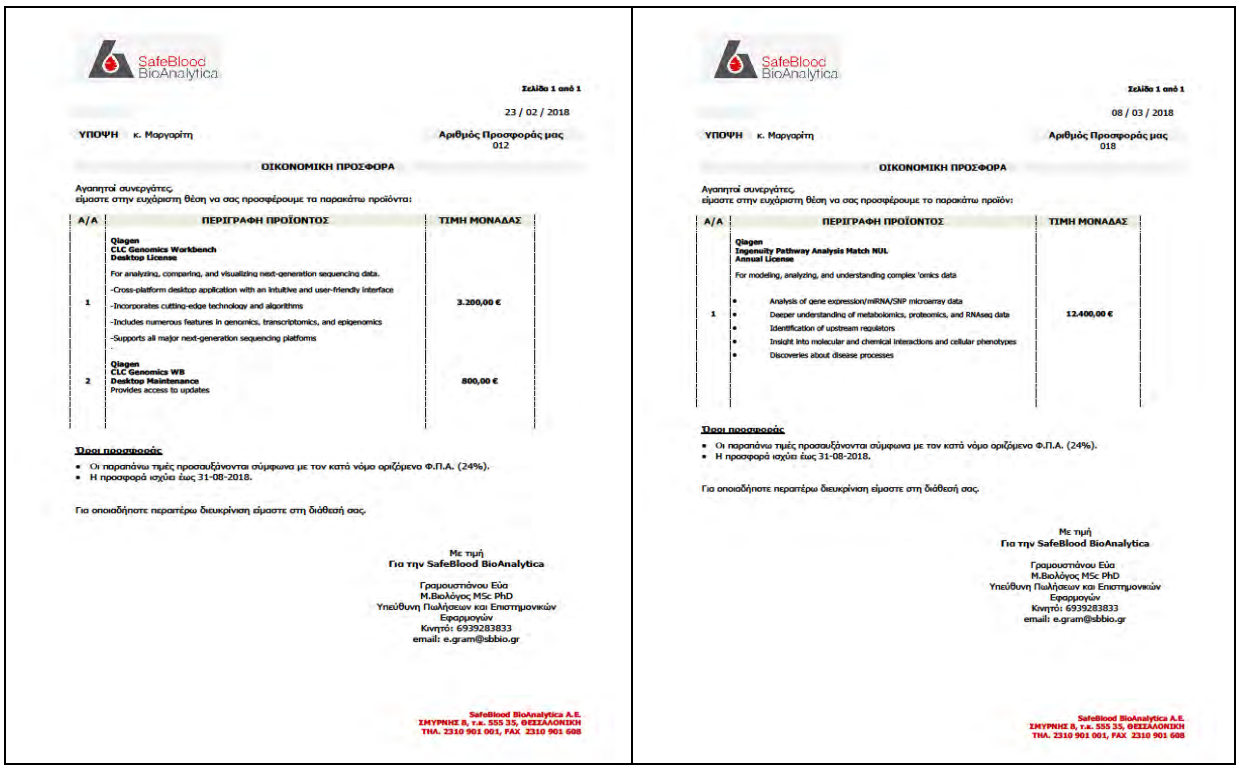


Εικόνα 2.1.15: Box plots διαγράμματα 4 δειγμάτων του mus musculus

Από τα διαγράμματα της Εικόνας 2.1.15 προκύπτει ότι στα Wild Type δείγματα οι τιμές της γονιδιακής έκφρασης που βρίσκονται στον κάθετο άξονα είναι μεγαλύτερες σε σχέση με τα δείγματα από τα οποία έχει αφαιρεθεί με τεχνητό τρόπο το γονίδιο Smyd3. Επίσης, παρατηρείται μεγαλύτερη απόσταση ανάμεσα στο ελάχιστο και το μέγιστο σημείο σε σχέση με τα Smyd3KO (Knock-Out) δείγματα ενώ ουσιαστική διαφορά εντοπίζεται και στη διάμεσο. Ακόμη, διαφορές στη δομή των διαγραμμάτων παρατηρούνται τόσο μεταξύ των WT δειγμάτων όσο και μεταξύ αυτών των Smyd3KO δειγμάτων.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Όσον αφορά στο μηνιαίο κόστος απόκτησης άδειας χρήσης για έναν ακαδημαϊκό χρήστη, αυτό ανέρχεται στα 413 € σύμφωνα με την Εικόνα 2.1.16 και προκύπτει ύστερα από επικοινωνία μέσω email με την εταιρεία SafeBlood BioAnalytica η οποία αποτελεί συνεργάτη της CLC bio, Qiagen στην Ελλάδα. Επίσης στην ίδια εικόνα διαφαίνεται το κόστος ετήσιας χρήσης του λογισμικού IPA το οποίο σε μηνιαία βάση ανέρχεται στο ποσό των 1281 €.



SafeBlood BioAnalytica

Σελίδα 1 από 1

23 / 02 / 2018

ΥΠΟΨΗ κ. Μαργαρίτη

Αριθμός Προσφοράς μας
012

ΟΙΚΟΝΟΜΙΚΗ ΠΡΟΣΦΟΡΑ

Αγαπητοί συνεργάτες,
είμαστε στην ευχάριστη θέση να σας προσφέρουμε τα παρακάτω προϊόντα:

Α/Α	ΠΕΡΙΓΡΑΦΗ ΠΡΟΪΟΝΤΟΣ	ΤΙΜΗ ΜΟΝΑΔΑΣ
1	**Qiagen** **CLC Genomics Workbench** **Desktop License** For analyzing, comparing, and visualizing next-generation sequencing data. -Cross-platform desktop application with an intuitive and user-friendly interface -Incorporates cutting-edge technology and algorithms -Includes numerous features in genomics, transcriptomics, and epigenomics -Supports all major next-generation sequencing platforms	3.200,00 €
2	**Qiagen** **CLC Genomics WB** **Desktop Maintenance** Provides access to updates	800,00 €

Όροι προσφοράς
- Οι παραπάνω τιμές προσαυξάνονται σύμφωνα με τον κατά νόμο οριζόμενο Φ.Π.Α. (24%).
- Η προσφορά ισχύει έως 31-08-2018.

Για οποιαδήποτε περαιτέρω διευκρίνιση είμαστε στη διάθεσή σας.

Με τιμή
Για την SafeBlood BioAnalytica

Γραμουστιάνου Εύα
M.Βιολόγος MSc PhD
Υπεύθυνη Πωλήσεων και Επιστημονικών Εφαρμογών
Κινητό: 6939283833
email: e.gram@sbbio.gr

SafeBlood BioAnalytica A.E.
ΣΜΥΡΝΗΣ 8, τ.κ. 555 35, ΘΕΣΣΑΛΟΝΙΚΗ
ΤΗΛ. 2310 901 001, FAX 2310 901 608

SafeBlood BioAnalytica

Σελίδα 1 από 1

08 / 03 / 2018

ΥΠΟΨΗ κ. Μαργαρίτη

Αριθμός Προσφοράς μας
018

ΟΙΚΟΝΟΜΙΚΗ ΠΡΟΣΦΟΡΑ

Αγαπητοί συνεργάτες,
είμαστε στην ευχάριστη θέση να σας προσφέρουμε τα παρακάτω προϊόν:

Α/Α	ΠΕΡΙΓΡΑΦΗ ΠΡΟΪΟΝΤΟΣ	ΤΙΜΗ ΜΟΝΑΔΑΣ
1	**Qiagen** **Ingenuity Pathway Analysis Match NUL** **Annual License** For modeling, analyzing, and understanding complex 'omics data • Analysis of gene expression/miRNA/SNP microarray data • Deeper understanding of metabolomics, proteomics, and RNAseq data • Identification of upstream regulators • Insight into molecular and chemical interactions and cellular phenotypes • Discoveries about disease processes	12.400,00 €

Όροι προσφοράς
- Οι παραπάνω τιμές προσαυξάνονται σύμφωνα με τον κατά νόμο οριζόμενο Φ.Π.Α. (24%).
- Η προσφορά ισχύει έως 31-08-2018.

Για οποιαδήποτε περαιτέρω διευκρίνιση είμαστε στη διάθεσή σας.

Με τιμή
Για την SafeBlood BioAnalytica

Γραμουστιάνου Εύα
M.Βιολόγος MSc PhD
Υπεύθυνη Πωλήσεων και Επιστημονικών Εφαρμογών
Κινητό: 6939283833
email: e.gram@sbbio.gr

SafeBlood BioAnalytica A.E.
ΣΜΥΡΝΗΣ 8, τ.κ. 555 35, ΘΕΣΣΑΛΟΝΙΚΗ
ΤΗΛ. 2310 901 001, FAX 2310 901 608

(α) **(β)**

Εικόνα 2.1.16: α) Το ετήσιο κόστος άδειας χρήσης του λογισμικού CLC Genomics Workbench

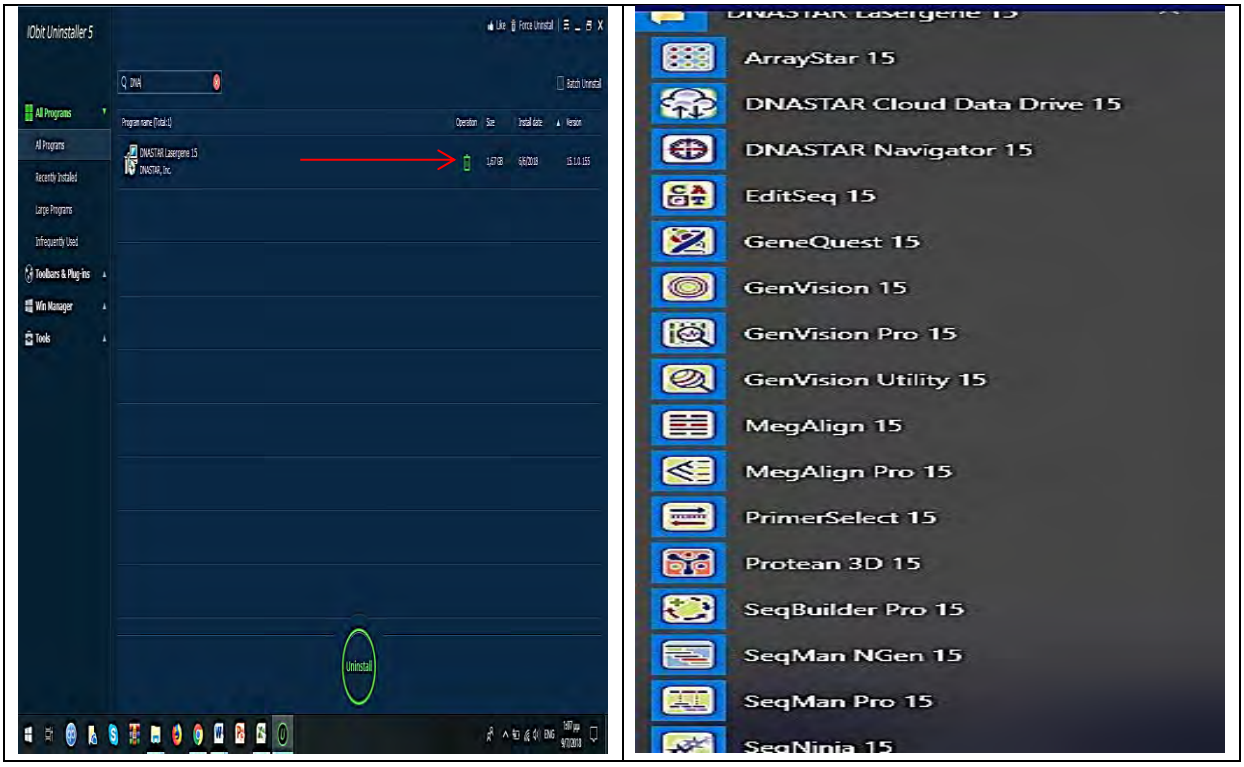
β) Το ετήσιο κόστος άδειας χρήσης του λογισμικού Ingenuity Pathway Analysis

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.2 Full Lasergene Suite (DNASTAR)

Πρώτη κατηγορία χαρακτηριστικών

Στην Εικόνα 2.2.1 που προέρχεται από τον υπολογιστή δοκιμής φαίνεται ότι ο χώρος που καταλαμβάνει το λογισμικό Full Lasergene Suite της εταιρείας DNASTAR είναι 1,67 GB περίπου. Το λογισμικό αυτό καταλαμβάνει σημαντικά μεγαλύτερο χώρο σε σχέση με το προηγούμενο και αυτό οφείλεται στο ότι προσφέρονται συνολικά 16 διαφορετικά προγράμματα που σχετίζονται με την αλληλούχιση άμεσα ή έμμεσα. Τα προγράμματα απεικονίζονται στη δεξιά στήλη της Εικόνας 2.2.1. Το λογισμικό αυτό υποστηρίζει και την Desktop αλλά και την Cloud αρχιτεκτονική συστήματος. Τέλος, όσον αφορά στο λειτουργικό σύστημα το λογισμικό υποστηρίζει μόνο σε 2 (Windows, Mac) από τους 3 τύπους των λειτουργικών συστημάτων σύμφωνα με την ιστοσελίδα της αμερικάνικης εταιρείας.



α) **β)**

Εικόνα 2.2.1: **α)** Χωρητικότητα του λογισμικού Full Lasergene Suite

β) Απεικόνιση των 16 προγραμμάτων της εταιρείας DNASTAR

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Δεύτερη κατηγορία χαρακτηριστικών

- FASTQ/BAM αρχεία

Στις Εικόνες 2.2.2, 2.2.3 και 2.2.4 που ακολουθούν φαίνονται τα βήματα που απαιτούνται προκειμένου να γίνει η εισαγωγή κάποιου αρχείου FASTQ από τον υπολογιστή στο πρόγραμμα SeqMan NGen το οποίο αποτελεί ένα από τα 16 συνολικά προγράμματα. Το αρχείο αυτό περιέχει δεδομένα που σχετίζονται με τον οργανισμό της Listeria monocytogenes. Η ίδια διαδικασία ακολουθείται σε γενικές γραμμές και για την εισαγωγή ενός BAM αρχείου. Η Listeria monocytogenes επιλέχθηκε ως παράδειγμα το οποίο προσφέρει η ίδια η εταιρεία ώστε να δοκιμάσει ο χρήστης τις δυνατότητες του λογισμικού Full Lasergene Suite. Ωστόσο δεν κατέστη εφικτή η επανάληψη της όλης διαδικασίας για τον οργανισμό mus musculus για λόγους που δεν μπορούν να προσδιοριστούν επακριβώς.

- Organisms

Δεν γίνεται σαφής περιγραφή των γονιδιωμάτων των οργανισμών με εξαίρεση τα γονιδιώματα του ανθρώπου και της αρκούδας στα οποία εφαρμόζεται η διαδικασία της αλληλούχισης δεύτερης γενιάς. Παρόλα αυτά, σύμφωνα με τα όσα υποστηρίζουν οι επιστημονικοί υπεύθυνοι της εταιρείας υπάρχουν διαθέσιμα πακέτα πρότυπου γονιδιώματος για κάποιους κοινούς οργανισμούς αλλά δίνεται ταυτόχρονα και η δυνατότητα στο χρήστη να προμηθεύσει τα δικά του αρχεία για άλλους οργανισμούς.

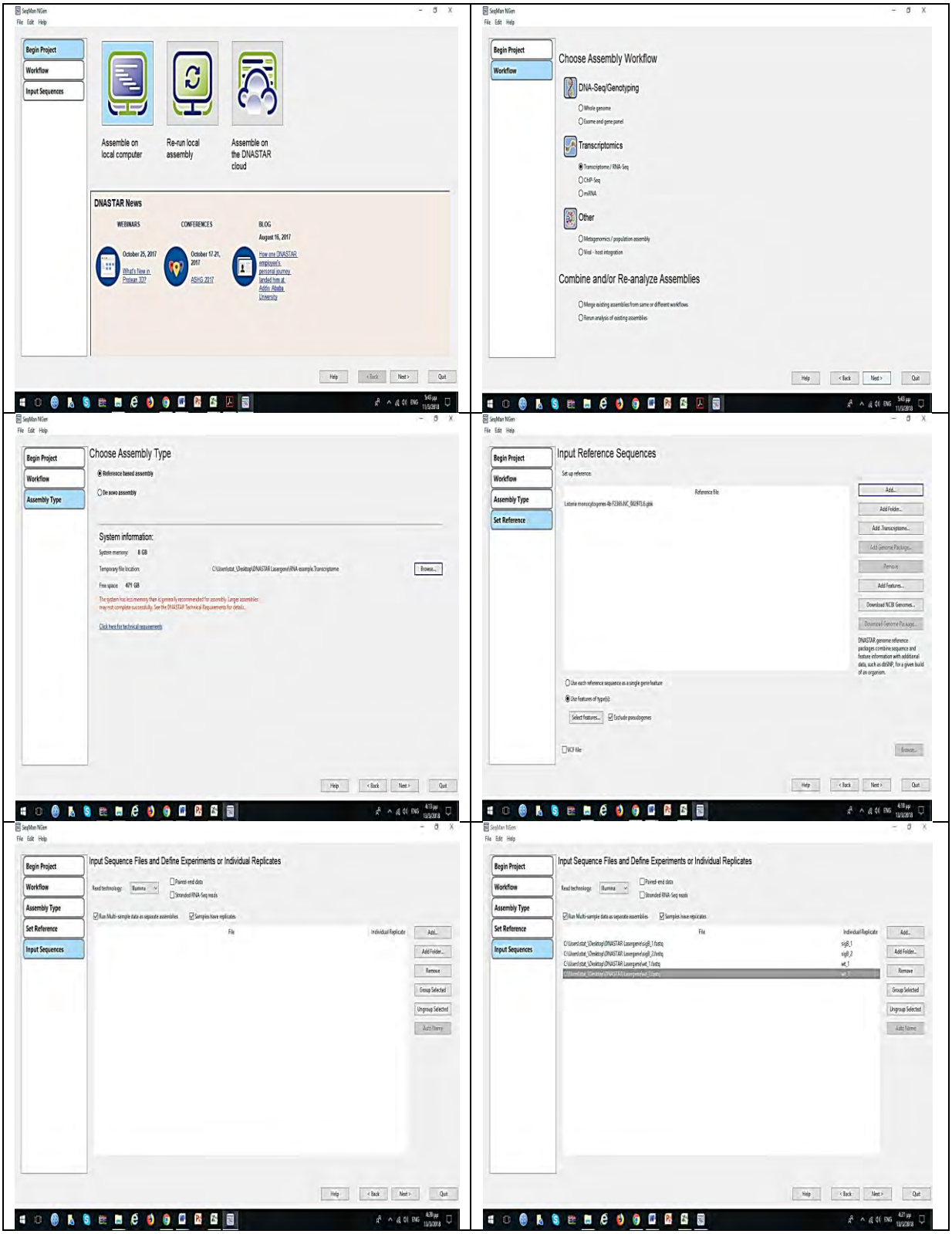
- Quality Control

Δεν κατέστη δυνατός ο εντοπισμός χαρακτηριστικών ποιοτικού ελέγχου σε αυτό το λογισμικό.

- Differential Expression

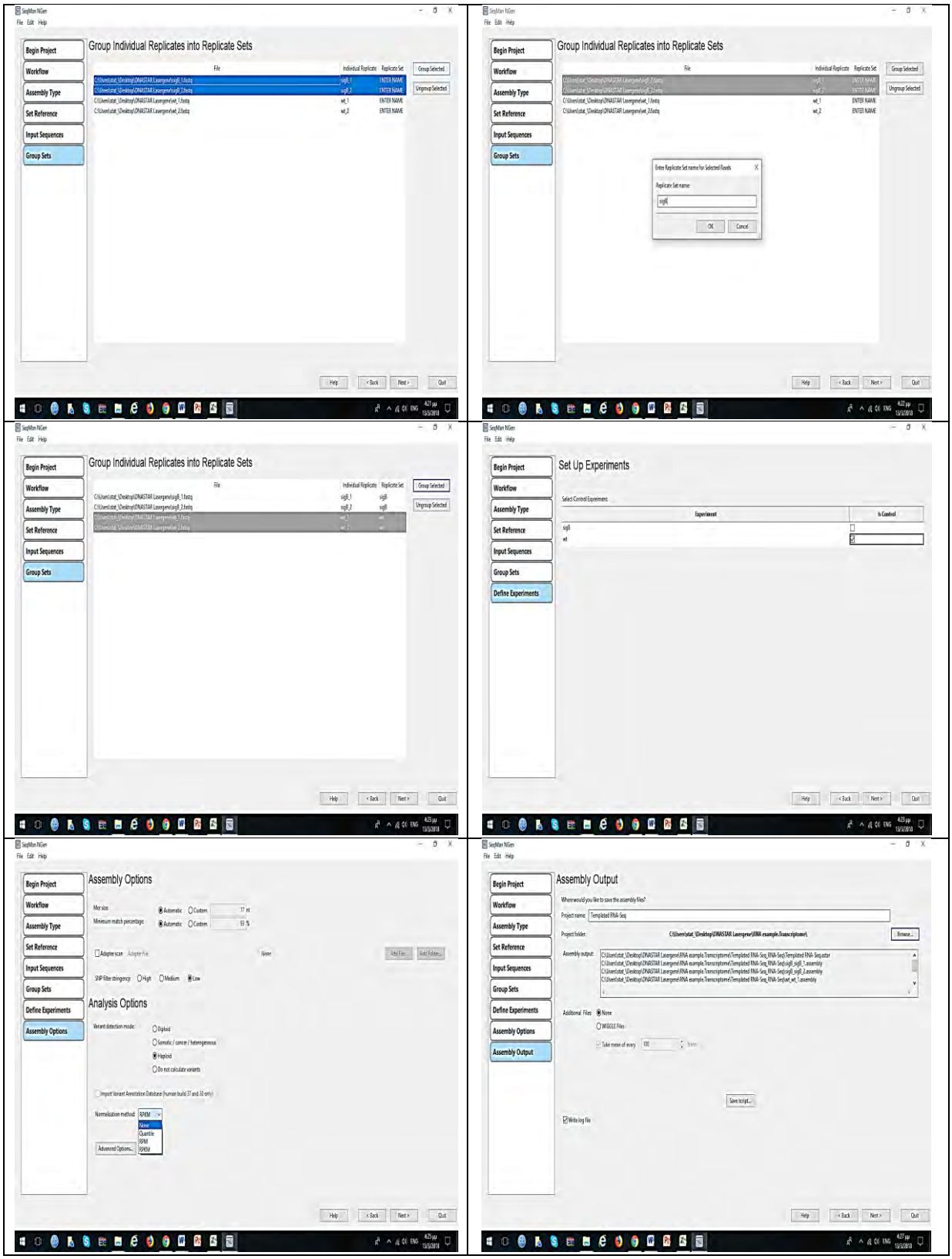
Δεν ήταν δυνατή η παράθεση μεταξύ των διαφόρων προγραμμάτων κάποιας εικόνας που να αναφέρεται στη διαφορική έκφραση των γονιδίων λόγω ανυπέρβλητων δυσκολιών για έναν αρχάριο υπολογιστικά χρήστη παρά τον ισχυρισμό των αρμοδίων ότι η δυνατότητα αυτή παρέχεται από το λογισμικό.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



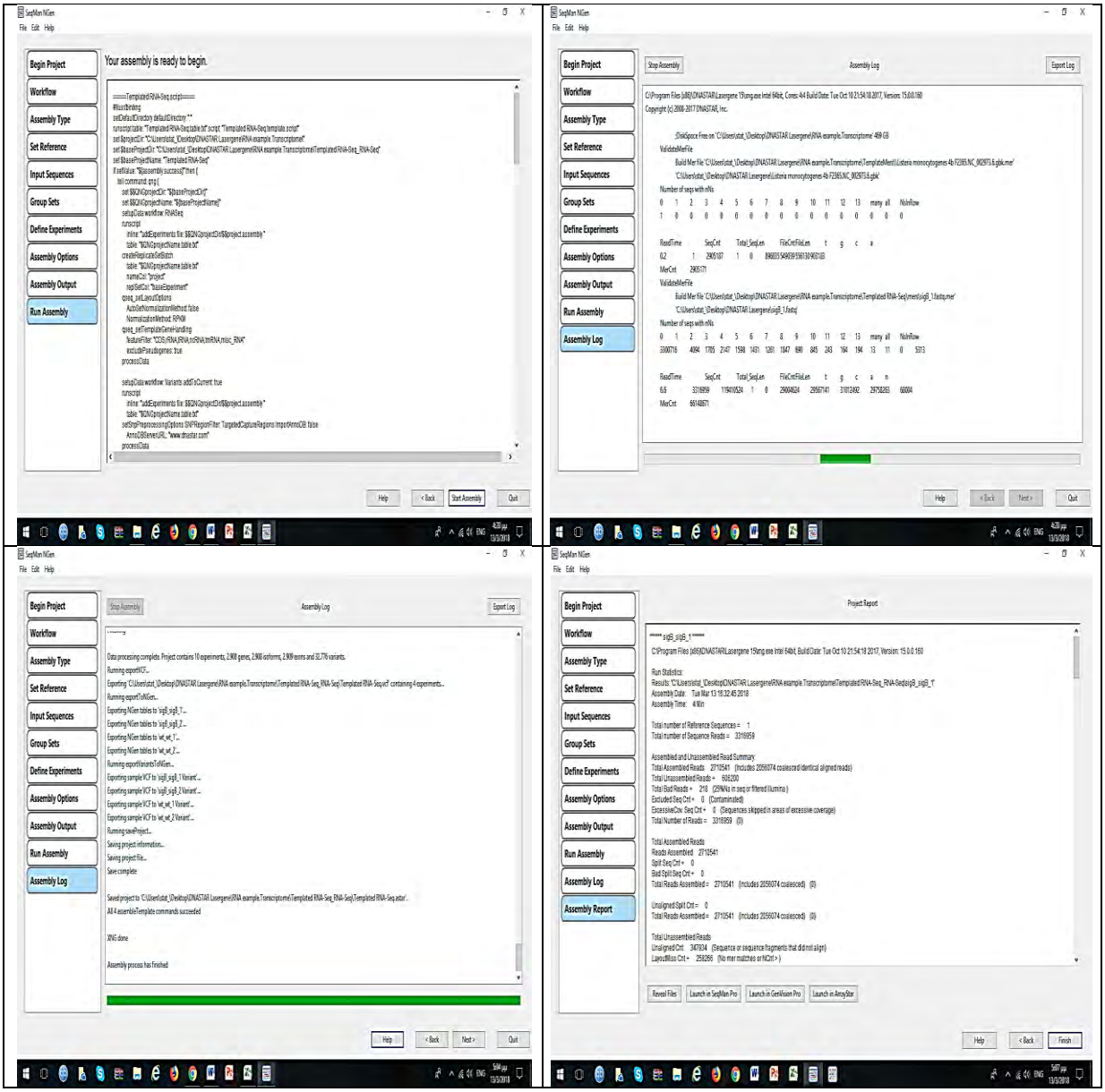
Εικόνα 2.2.2: Διαδικασία εισαγωγής FASTQ αρχείου στο πρόγραμμα SeqMan NGen (I)

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.2.3: Διαδικασίας εισαγωγής FASTQ αρχείου στο πρόγραμμα SeqMan NGen (II)

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.2.4: Διαδικασία εισαγωγής FASTQ αρχείου στο πρόγραμμα SeqMan NGen (III)

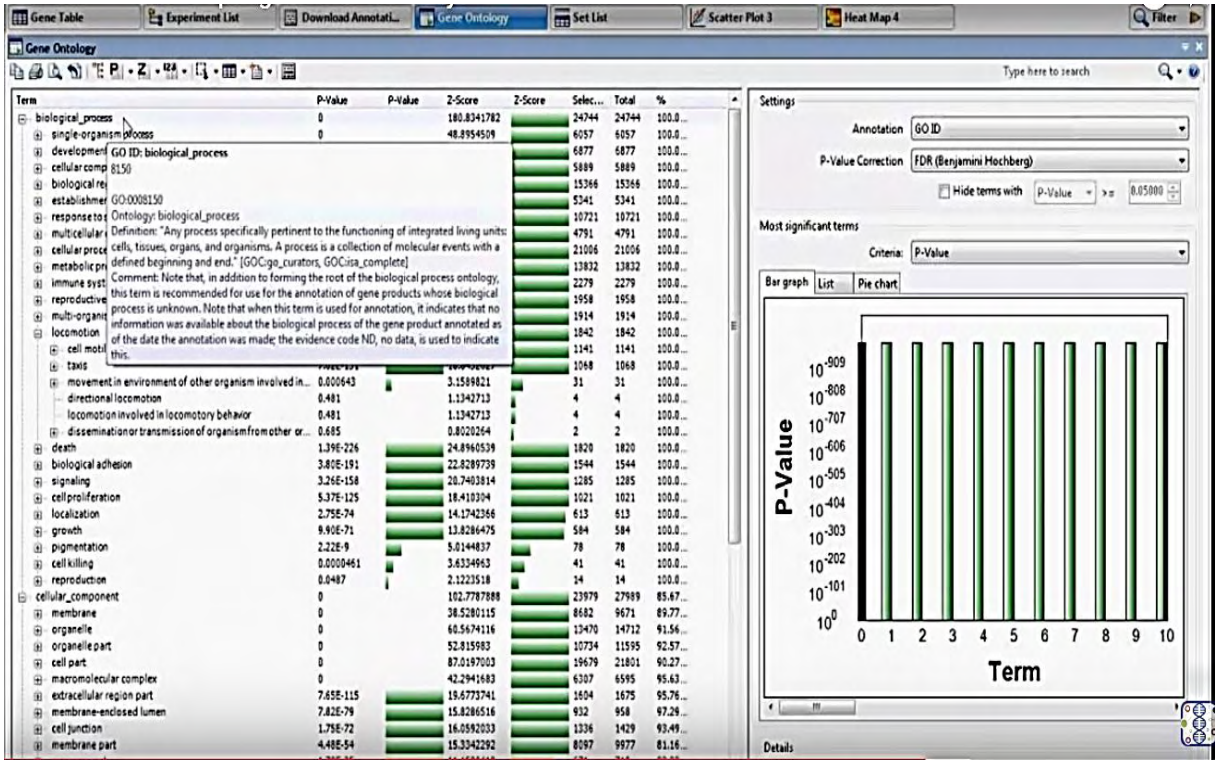
- External Annotation Databases

Η μόνη πληροφορία που αφορά στις εξωτερικές βάσεις δεδομένων είναι η διαβεβαίωση του εκπροσώπου της εταιρείας ότι το λογισμικό υποστηρίζει τις βάσεις GenBank και Protein Data Bank.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- GO Analysis

Η δυνατότητα ανάλυσης εμπλουτισμού οντολογιών των γονιδίων υποστηρίζεται από το πρόγραμμα ArrayStar που ανήκει στην οικογένεια των προγραμμάτων της εταιρείας DNASTAR. Ένα στιγμιότυπο του προηγούμενου οργανισμού φαίνεται στην Εικόνα 2.2.5.



Εικόνα 2.2.5: Αποτέλεσμα της GO ανάλυσης της Listeria monocytogenes από το ArrayStar

- Biochemical Pathway Analysis

Η δυνατότητα ανάλυσης βιοχημικών μονοπατιών δεν παρέχεται από τη συγκεκριμένη έκδοση του λογισμικού σύμφωνα με ενημέρωση, μέσω email, των ανθρώπων της εταιρείας.

- Workflows Creations

Η δημιουργία ροής εργασιών θεωρείται από τους ιθύνοντες της εταιρείας ότι παρέχεται αν και δεν κατέστη εφικτή η συγκέντρωση κάποιας αναφοράς, κάποιας εικόνας ή στιγμιότυπου.

- Clustering Analysis

Η ανάλυση κατά συστάδες παρέχεται από το πρόγραμμα ArrayStar με δυνατότητα επιλογής είτε μέσω Hierarchical Clustering είτε μέσω του αλγορίθμου K-means και μάλιστα σε συσχέτιση με τους χάρτες θερμότητας. Για το λόγο αυτό δεν παρατίθεται σε αυτό το

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

χαρακτηριστικό κάποια εικόνα αλλά αντίθετα παρουσιάζεται ένα στιγμιότυπο στην ανάλυση των heat maps που βρίσκεται στην επόμενη κατηγορία χαρακτηριστικών.

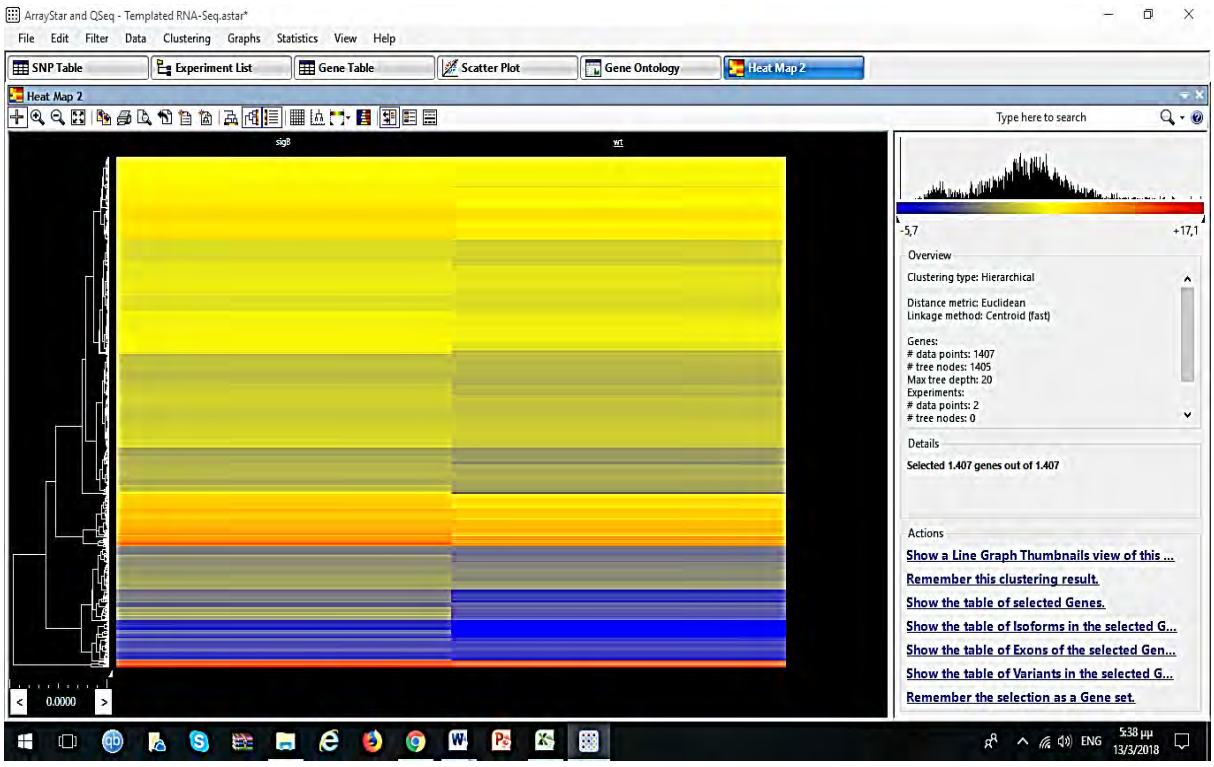
- Custom Reports

Η δημιουργία αναφορών για συγκεκριμένα γονίδια που ενδιαφέρουν το χρήστη προσφέρεται όπως αναφέρουν οι εκπρόσωποι της εταιρείας DNASTAR αν και δεν κατέστη δυνατός ο εντοπισμός μίας ή περισσότερων τέτοιων αναφορών κατά τη διάρκεια επεξεργασίας των προγραμμάτων του λογισμικού.

Τρίτη κατηγορία χαρακτηριστικών

- Heat maps

Η δημιουργία χαρτών θερμότητας είναι μία δυνατότητα που υποστηρίζεται από το πρόγραμμα ArrayStar. Η Εικόνα 2.2.6 παρουσιάζει έναν τέτοιο χάρτη που προκύπτει ύστερα από την ανάλυση κατά συστάδες με χρήση του αλγόριθμου K-means.



Εικόνα 2.2.6: Heat map για τον οργανισμό Listeria monocytogenes

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Volcano/MA plots

Η κατασκευή τέτοιου είδους γραφικών παραστάσεων δεν υποστηρίζεται από τη συγκεκριμένη έκδοση του λογισμικού Full Lasergene Suite σύμφωνα με τους υπεύθυνους της εταιρείας.

- Genome Browser

Η δυνατότητα περιήγησης γονιδιωμάτων των οργανισμών προσφέρεται από το πρόγραμμα GenVision Pro σύμφωνα με την ενημέρωση των υπεύθυνων του λογισμικού. Ωστόσο δεν κατέστη δυνατή ούτε η εκτέλεση αυτής της διαδικασίας για τον εξεταζόμενο οργανισμό ούτε η εύρεση κάποιου παραδείγματος από το manual που να αφορά το χαρακτηριστικό αυτό λόγω πολυπλοκότητας της διαδικασίας για έναν άπειρο σε τέτοιου είδους λογισμικά χρήστη.

- PCA/MDS plots

Η ανάλυση κύριων συνιστωσών όπως και τα γραφήματα πολυδιάστατης κλιμάκωσης δεν υποστηρίζονται ακόμη από την υπάρχουσα έκδοση του λογισμικού όπως τονίζουν οι εκπρόσωποι της εταιρείας.

- Box plots

Η κατασκευή τέτοιου τύπου διαγραμμάτων δεν πραγματοποιείται από κάποιο πρόγραμμα της πλατφόρμας όπως υποστηρίζουν οι άνθρωποι της εταιρείας με τους οποίους υπάρχει ταύτιση απόψεων καθότι δεν εντοπίστηκε κάποιο πρόγραμμα που να παρέχει τη δυνατότητα για οπτικοποίηση των δεδομένων μέσω boxplot.

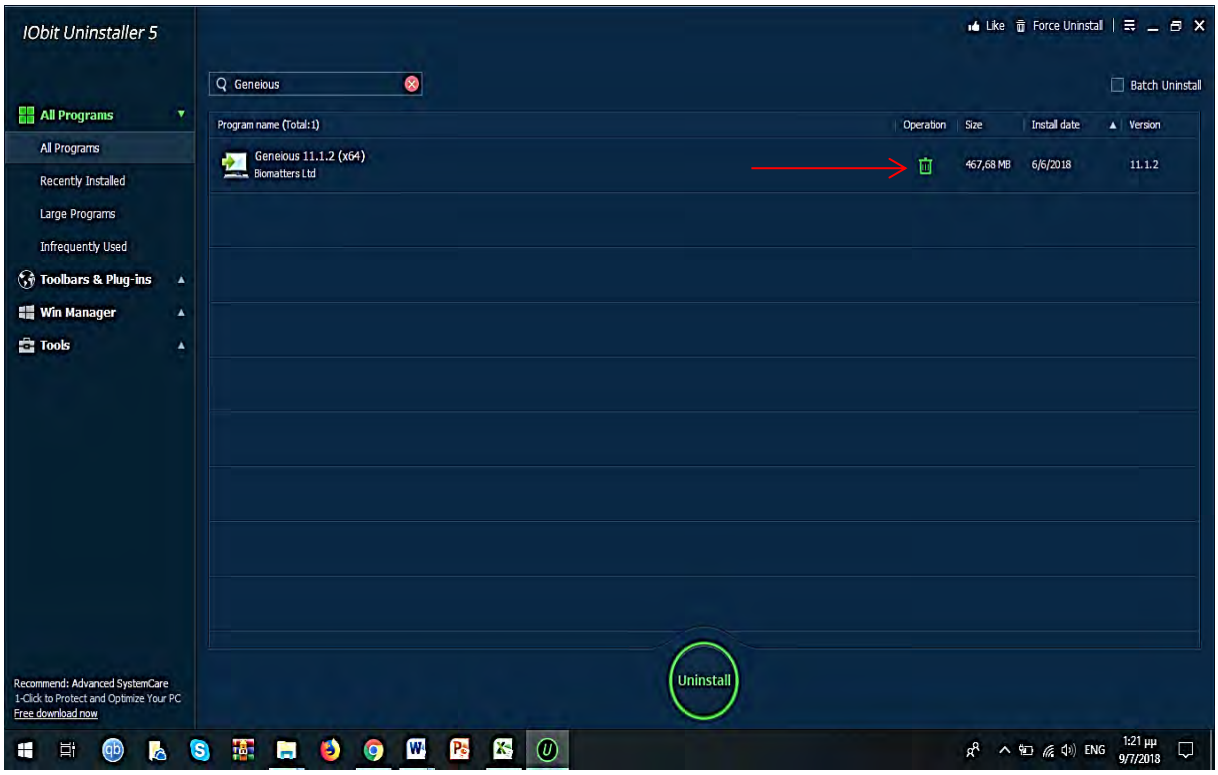
Όσον αφορά στο μηνιαίο κόστος απόκτησης άδειας χρήσης για έναν ακαδημαϊκό χρήστη, αυτό ανέρχεται στα 217 € ή στα 5095 € εφ όρου ζωής και προκύπτει ύστερα από προσωπική επικοινωνία μέσω email με το τμήμα Μάρκετινγκ και Πωλήσεων της εταιρείας DNASTAR.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.3 Geneious (Biomatters)

Πρώτη κατηγορία χαρακτηριστικών

Στην Εικόνα 2.3.1 διαφαίνεται ότι ο χώρος που καταλαμβάνει το λογισμικό Geneious της εταιρείας Biomatters είναι 468 MB περίπου. Η αρχιτεκτονική συστήματος είναι Desktop ενώ το λογισμικό υποστηρίζει και τους 3 τύπους των λειτουργικών συστημάτων σύμφωνα με την ιστοσελίδα της εταιρείας που έχει την έδρα της στη Νέα Ζηλανδία.



Εικόνα 2.3.1: Χωρητικότητα του λογισμικού Geneious

Δεύτερη κατηγορία χαρακτηριστικών

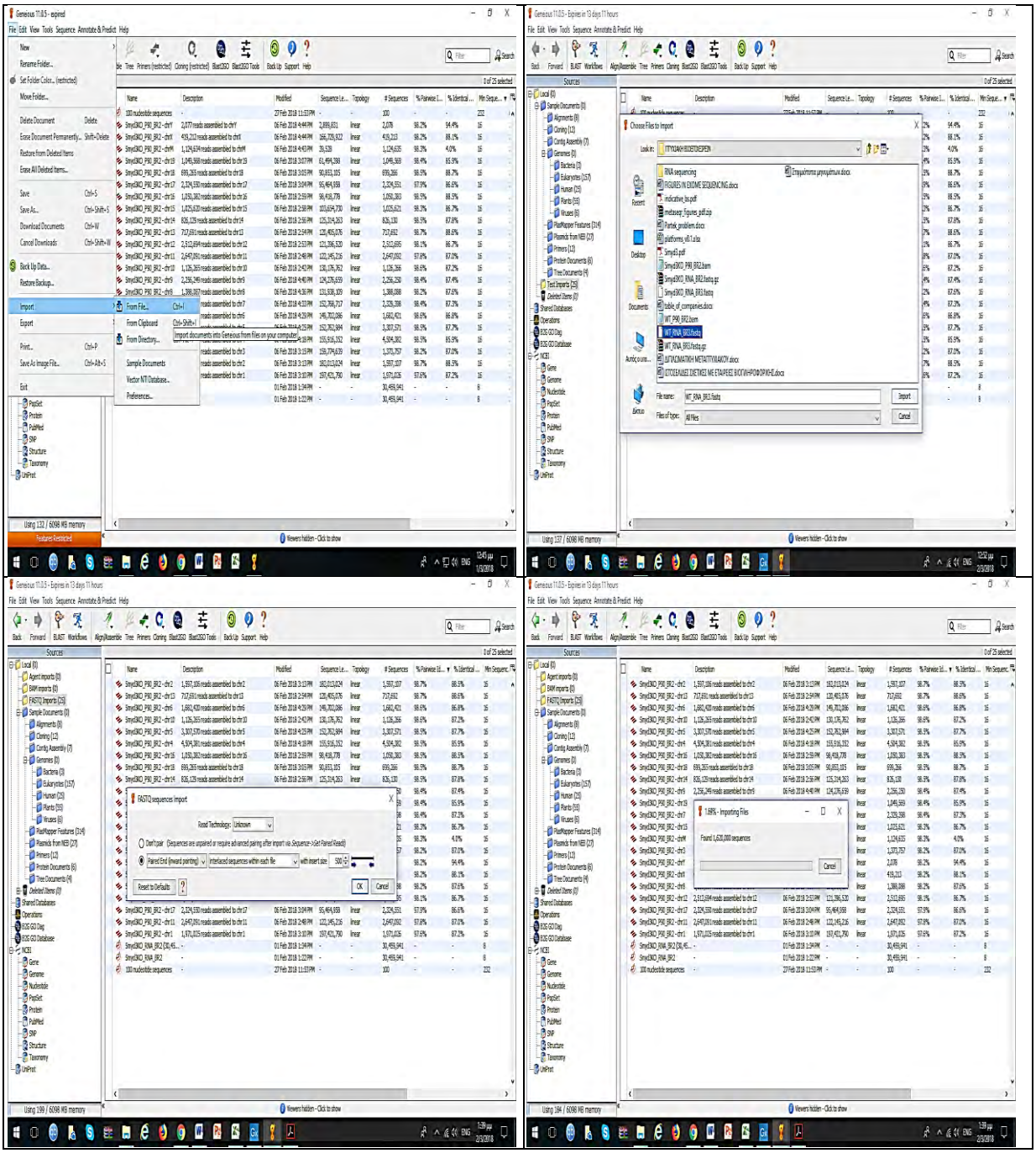
- FASTQ/BAM αρχεία

Στις Εικόνες 2.3.2 και 2.3.3 παρουσιάζεται η διαδικασία εισαγωγής από τον υπολογιστή δοκιμής ενός FASTQ αρχείου που αφορά στον εξεταζόμενο οργανισμό.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

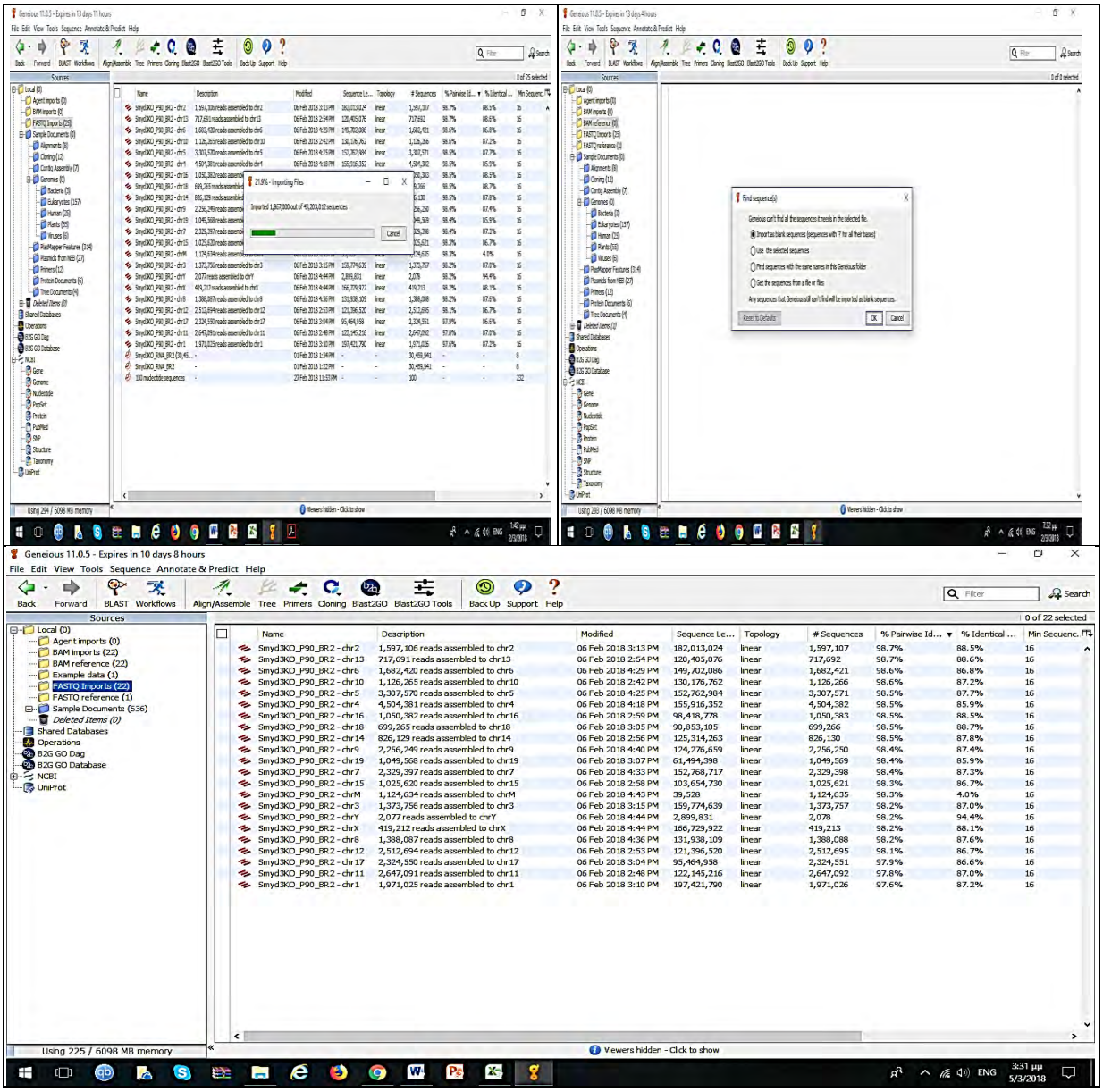
- Organisms

Τα γονιδιώματα όλων των ευκαρυωτικών οργανισμών όπως επίσης πολλών φυτών, ιών και βακτηρίων παρέχονται από το λογισμικό Geneious σύμφωνα με το manual και τους επιστημονικούς υπεύθυνους της εταιρείας.



Εικόνα 2.3.2: Στιγμιότυπα από τη διαδικασία εισαγωγής ενός FASTQ αρχείου (I)

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

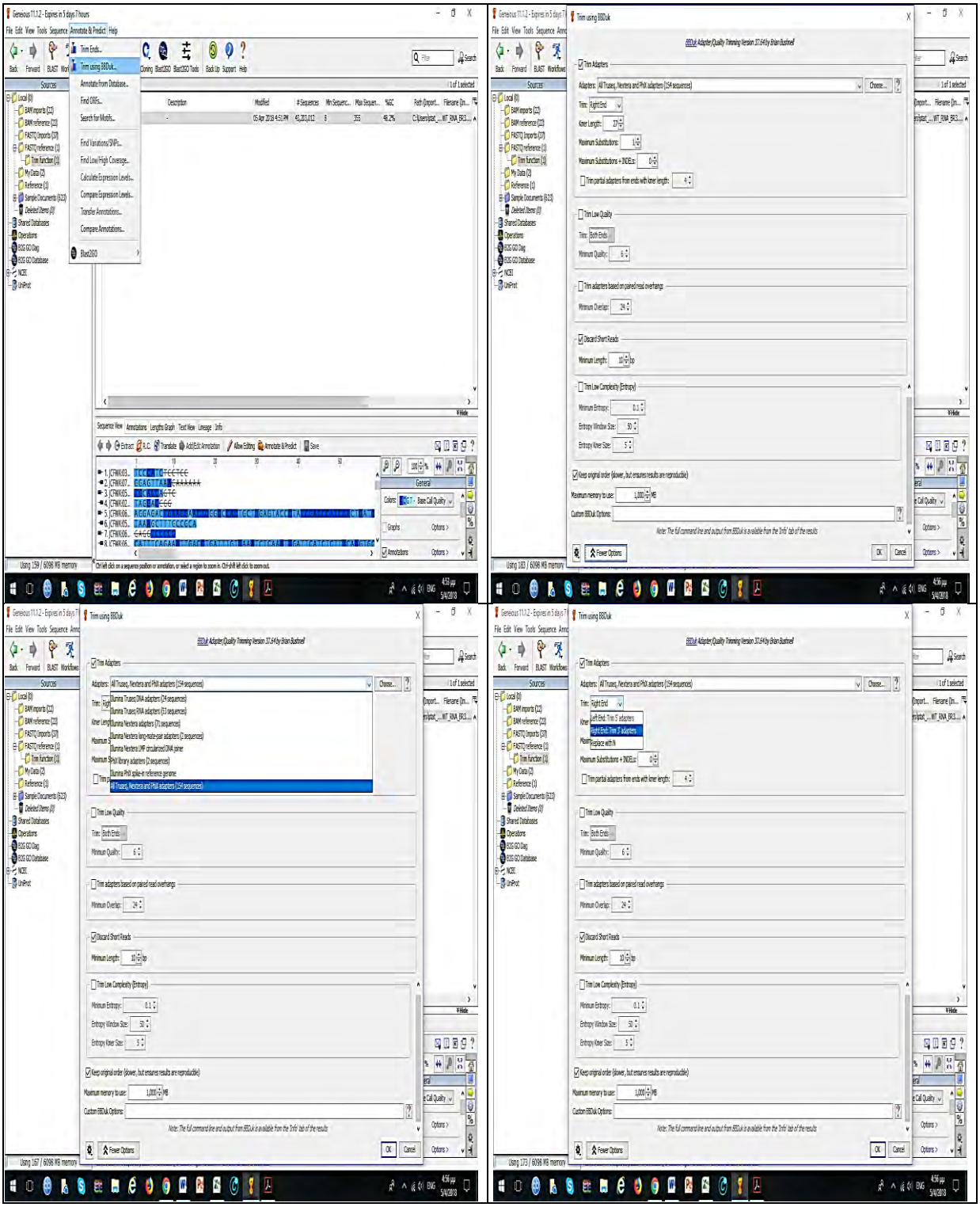


Εικόνα 2.3.3: Στιγμιότυπα από τη διαδικασία εισαγωγής ενός FASTQ αρχείου (II)

- Quality Control

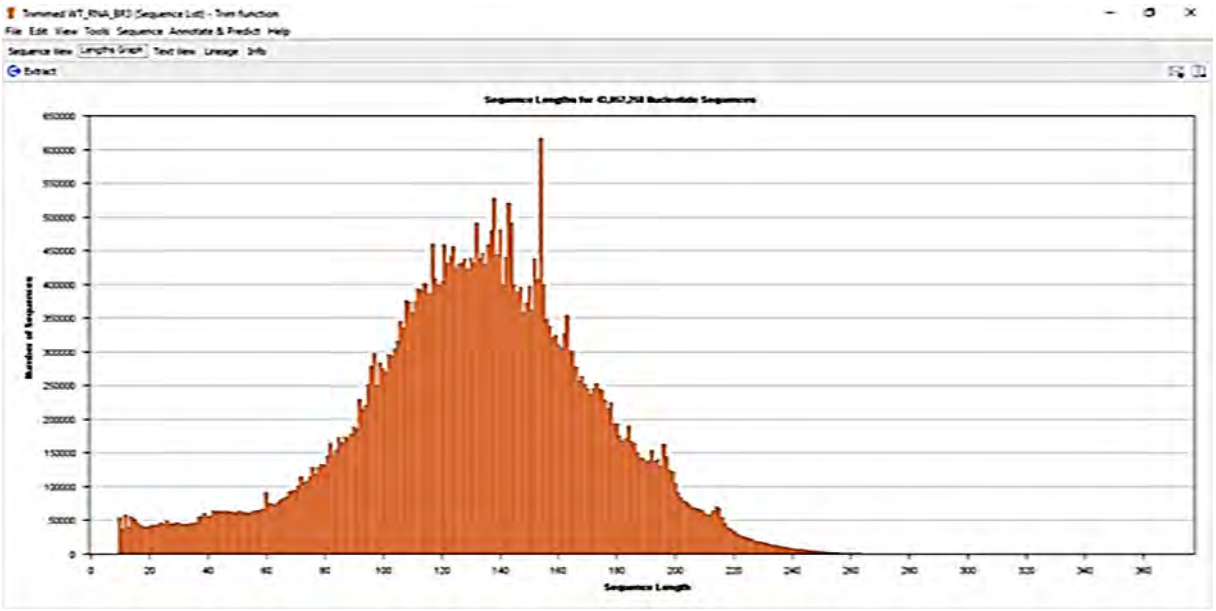
Ο ποιοτικός έλεγχος διεξάγεται μέσω ενός εργαλείου (plugin) του λογισμικού που φέρει το όνομα BBDuk Trimmer και τα βήματα της διαδικασίας που ακολουθείται παρατίθενται οπτικά στην Εικόνα 2.3.4 ενώ η Εικόνα 2.3.5 απεικονίζει ένα διάγραμμα που αφορά το length distribution των αλληλουχηθέντων ολιγονουκλεοτιδίων του mus musculus. Ειδικότερα, αυτό που φαίνεται από την Εικόνα 2.3.5 είναι ότι η μέση κατανομή μήκους σε ζεύγη νουκλεοτιδικών βάσεων ανέρχεται περίπου στις 160 βάσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.3.4: Η διαδικασία ποιοτικού ελέγχου του Geneious

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.3.5: Στιγμιότυπο από τον ποιοτικό έλεγχο του mus musculus

- Differential Expression

Η διαδικασία διαφορικής έκφρασης των γονιδίων εφαρμόζεται μόνο σε προγραμματιστικό περιβάλλον R του πακέτου DESeq. Αν και έγινε προσπάθεια για εκτέλεση αυτού του προγράμματος σε γλώσσα προγραμματισμού R δεν στέφθηκε από επιτυχία και επομένως δεν προέκυψε κάποιο στιγμιότυπο.

- External Annotation Databases

Σύμφωνα με τους επιστημονικούς συνεργάτες της εταιρείας Biomatters οι εξωτερικές βάσεις δεδομένων που συνδέονται με το λογισμικό είναι η RefSeq και η PubMed.

- GO Analysis

Η ανάλυση του εμπλουτισμού των οντολογιών της διαφορικής έκφρασης γονιδίων υποστηρίζεται από ένα άλλο plugin του λογισμικού Geneious που ονομάζεται GO-Slim. Ωστόσο δεν κατέστη εφικτή η ολοκλήρωση της ανάλυσης λόγω ανεπάρκειας σε μνήμη του υπολογιστή δοκιμής.

- Biochemical Pathway Analysis

Η δυνατότητα ανάλυσης βιοχημικών μονοπατιών δεν παρέχεται σε αυτήν την έκδοση του λογισμικού έπειτα από ενημέρωση των επιστημονικών υπεύθυνων της εταιρείας.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Workflows Creations

Σύμφωνα με τους ισχυρισμούς των ανθρώπων της εταιρείας η δημιουργία διαγραμμάτων ροής υποστηρίζεται μερικώς. Ωστόσο δεν κατέστη δυνατή η εύρεση κάποιας τέτοιας δυνατότητας κατά τη διάρκεια εξερεύνησης του λογισμικού.

- Clustering Analysis

Η ανάλυση κατά συστάδες είναι δυνατή μόνο με χρήση του plugin Blast2GO το οποίο δεν περιλαμβάνεται στη βασική έκδοση του λογισμικού για αυτό και απαιτείται ξεχωριστή μηνιαία συνδρομή.

- Custom Reports

Η δυνατότητα δημιουργίας αναφορών για συγκεκριμένα γονίδια που ενδιαφέρουν το χρήστη δεν παρέχεται στην παρούσα τουλάχιστον έκδοση του Geneious και αυτή η παρατήρηση επιβεβαιώνεται και από τους ανθρώπους της εταιρείας Biomatters.

Τρίτη κατηγορία χαρακτηριστικών

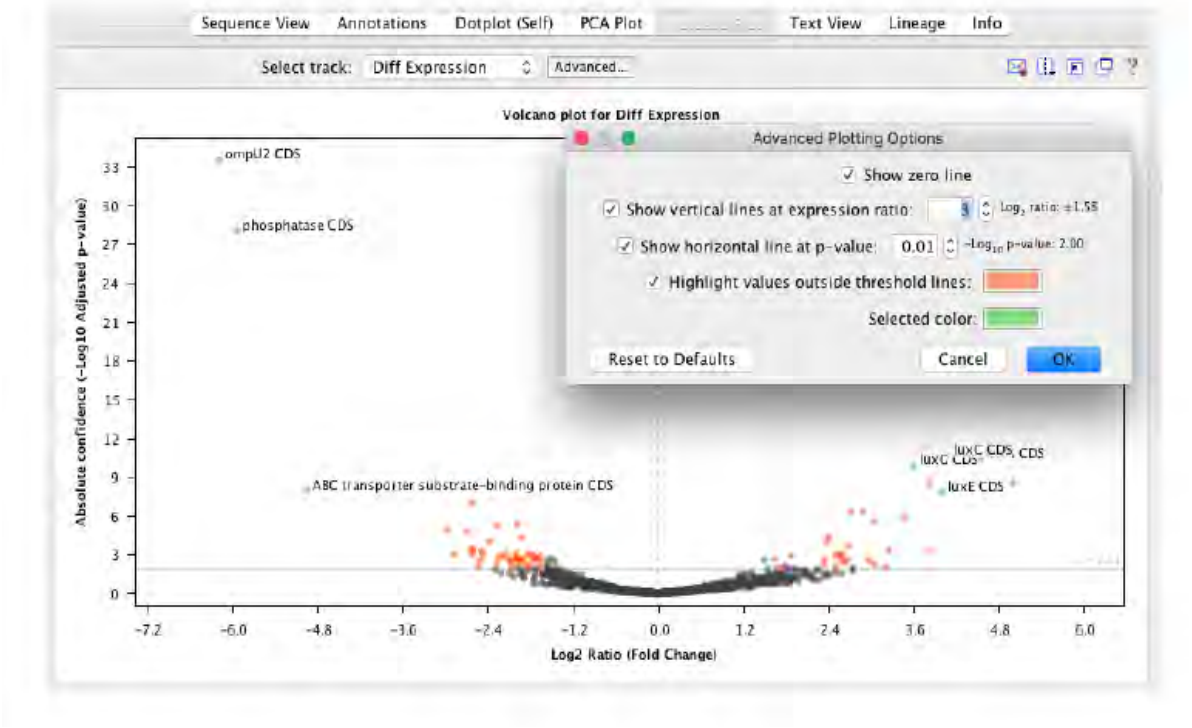
- Heat maps

Η κατασκευή χαρτών θερμότητας είναι μία από τις δυνατότητες που παρέχονται από το λογισμικό όπως επισημαίνουν οι αρμόδιοι. Παρόλα αυτά θεωρείται αρκετά περίπλοκη η διαδικασία κατασκευής τέτοιων χαρτών για έναν αρχάριο υπολογιστικά χρήστη και αυτός είναι ο λόγος που δεν υπάρχει κάποιο στιγμιότυπο από τη διαδικασία ανάλυσης του εξεταζόμενου οργανισμού.

- Volcano plots

Η διαδικασία δημιουργίας Volcano γραφικών παραστάσεων είναι μία δυνατότητα που παρέχεται από το Geneious όπως αναφέρεται και στο manual του λογισμικού και ένα στιγμιότυπο μιας τέτοιας γραφικής παράστασης παρουσιάζεται στην Εικόνα 2.3.6.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.3.6: Volcano γραφική παράσταση από το manual του λογισμικού

- MA plots

Η κατασκευή MA γραφικών παραστάσεων δεν υποστηρίζεται από το λογισμικό όπως επιβεβαιώνουν και οι αρμόδιοι της εταιρείας.

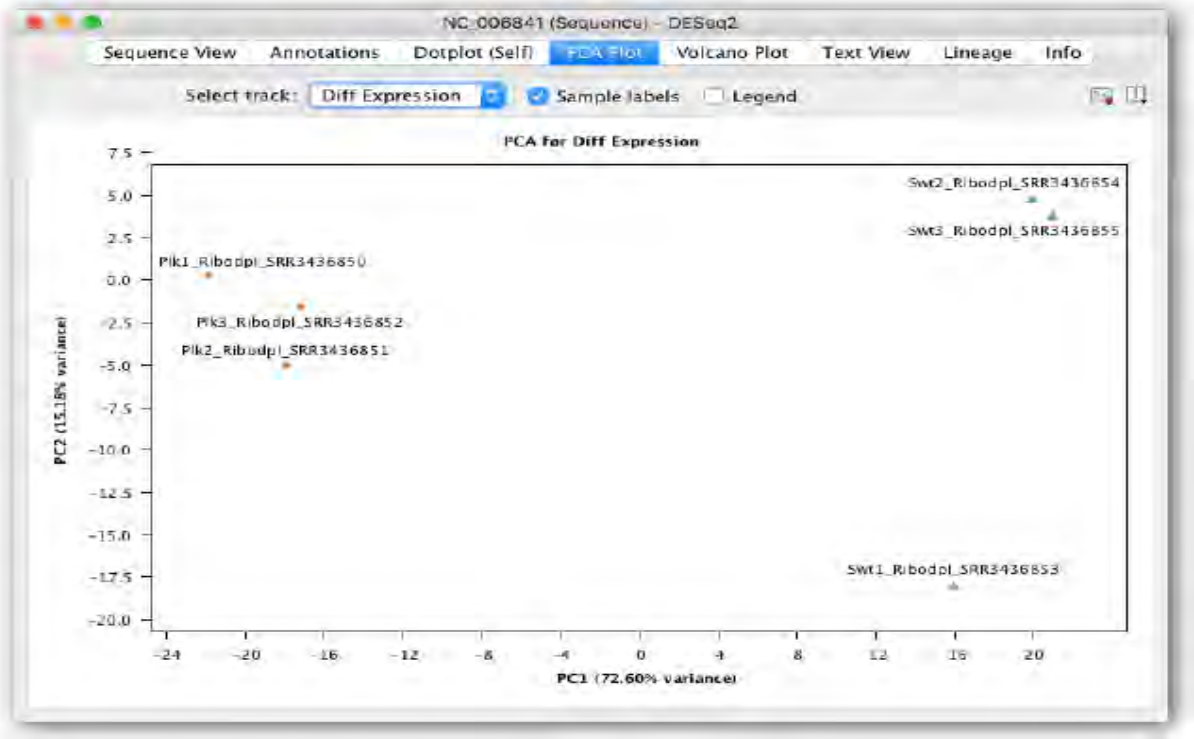
- Genome Browser

Σύμφωνα με τους ισχυρισμούς των ανθρώπων της Biomatters το λογισμικό Geneious παρέχει κάποιο υποστηρικτικό πρόγραμμα περιήγησης γονιδιωμάτων το οποίο είναι παρόμοιο με το UCSC αλλά όχι το ίδιο λειτουργικό με αυτό. Ωστόσο δεν κατέστη εφικτή η συλλογή κάποιου στιγμιότυπου του χαρακτηριστικού αυτού.

- PCA plots

Η δυνατότητα δημιουργίας PCA γραφημάτων παρέχεται από το συγκεκριμένο λογισμικό και αυτό δείχνει η Εικόνα 2.3.7. Παρ' όλα αυτά θεωρείται αρκετά περίπλοκο για έναν μη έμπειρο υπολογιστικά χρήστη να κατανοήσει και να εφαρμόσει τη διαδικασία δημιουργίας τέτοιων γραφημάτων και για αυτό το λόγο δεν υπάρχει κάποιο στιγμιότυπο που να αφορά τον εξεταζόμενο οργανισμό.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.3.7: PCA γράφημα του βακτηρίου Vibrio fischeri

- MDS

Η δημιουργία γραφημάτων πολυδιάστατης κλιμάκωσης δεν προσφέρεται από το λογισμικό και σε αυτό το συμπέρασμα καταλήγουν και οι ίδιοι οι άνθρωποι της.

- Box plots

Η κατασκευή τέτοιου είδους διαγραμμάτων δεν παρέχεται από την παρούσα έκδοση του λογισμικού Geneious όπως επιβεβαιώνουν και οι ιθύνοντες της εταιρείας.

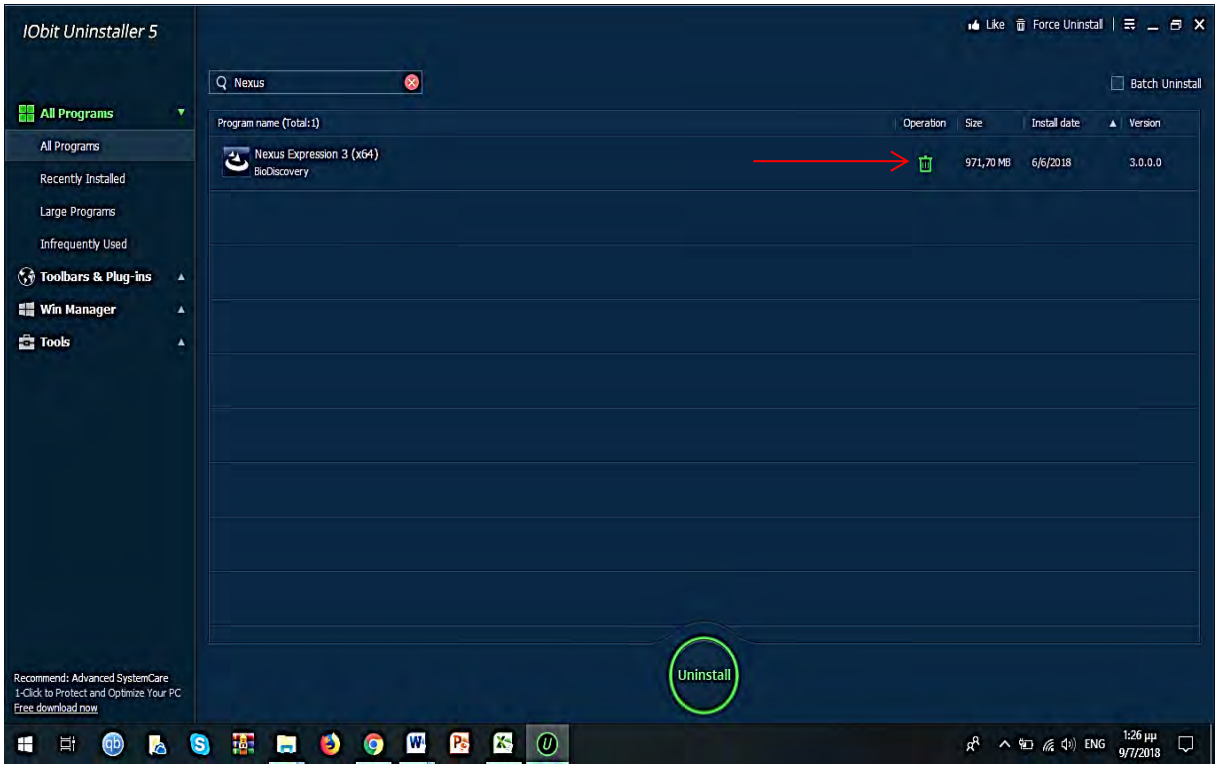
Όσον αφορά στο μηνιαίο κόστος απόκτησης άδειας χρήσης για έναν ακαδημαϊκό χρήστη αυτό ανέρχεται στα 16,25 $ ενώ κυμαίνεται από 33 $ έως 83 $ περίπου για ένα πανεπιστήμιο ή κάποια μη κυβερνητική οργάνωση αντίστοιχα και ανεβαίνει ανάλογα στα 104 ή 208 $ για κάποια κερδοσκοπική οργάνωση σύμφωνα με την ιστοσελίδα της εταιρείας. Όσον αφορά στο μηνιαίο κόστος του plugin Blast2GO για έναν ακαδημαϊκό χρήστη αυτό ανέρχεται στα 75$ περίπου σύμφωνα με την αντίστοιχη ιστοσελίδα.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.4 Nexus Expression (Biodiscovery)

Πρώτη κατηγορία χαρακτηριστικών

Στην Εικόνα 2.4.1 φαίνεται ότι ο χώρος που καταλαμβάνει το λογισμικό Nexus Expression της εταιρείας Biodiscovery είναι 972 MB περίπου. Η αρχιτεκτονική συστήματος είναι Desktop. Επιπλέον, όσον αφορά στο λειτουργικό σύστημα το λογισμικό υποστηρίζει και τους 3 τύπους των συστημάτων σύμφωνα με την ιστοσελίδα της αμερικάνικης εταιρείας.



Εικόνα 2.4.1: Χωρητικότητα του λογισμικού Nexus Expression

Δεύτερη κατηγορία χαρακτηριστικών

- FASTQ/BAM αρχεία

Σύμφωνα με τους επιστημονικούς υπεύθυνους του λογισμικού δεν προσφέρεται άμεση δυνατότητα εισαγωγής των αρχείων μορφής FASTQ και BAM. Αντιθέτως, είναι αναγκαία η μετατροπή των αλληλουχιών σε κανονικοποιημένη μορφή RPKM (Reads Per Kilobase Million) ώστε να είναι δυνατή η περαιτέρω επεξεργασία τους.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Organisms

Οι υπεύθυνοι του λογισμικού Nexus Expression ύστερα από ανταλλαγή μηνυμάτων μέσω email υποστήριξαν ότι το συγκεκριμένο λογισμικό μπορεί να χρησιμοποιηθεί για οποιονδήποτε οργανισμό είναι διαθέσιμη μία από τις γονιδιωματικές δομήσεις (genomic build) που προσφέρει η συγκεκριμένη ιστοσελίδα της εταιρείας Biodiscovery. Ακόμη, υπήρξε διαβεβαίωση ότι είναι δυνατή η προσθήκη γονιδιωμάτων για διαφορετικούς από τους ήδη υπάρχοντες οργανισμούς σε περίπτωση που τους ζητηθεί.

- Quality Control

΄Υστερα από επικοινωνία που διεξήχθη μέσω email με αρμόδιους της εταιρείας τονίσθηκε ότι η παρούσα έκδοση του λογισμικού δεν παρέχει τη δυνατότητα εκτέλεσης ποιοτικού ελέγχου για δεδομένα αλληλούχισης δεύτερης γενιάς.

- Differential Expression

Η δυνατότητα αυτή παρέχεται από το λογισμικό σύμφωνα με τον ισχυρισμό των επιστημονικών υπεύθυνων της εταιρείας. Ωστόσο δεν κατέστη εφικτός ο εντοπισμός μίας ή περισσότερων εικόνων που να αναφέρονται στη διαφορική έκφραση των γονιδίων λόγω υπολογιστής ανεπάρκειας σε μνήμη.

- External Annotation Databases

Η μόνη πληροφορία που προέκυψε ύστερα από εξερεύνηση του λογισμικού είναι ότι προσφέρεται τη δυνατότητα σχολιασμού του επιπέδου της διαφορικής έκφρασης γονιδίων χρησιμοποιώντας τη βάση δεδομένων RefSeq καθώς και συνδέσμους με άλλες εξωτερικές βάσεις δεδομένων.

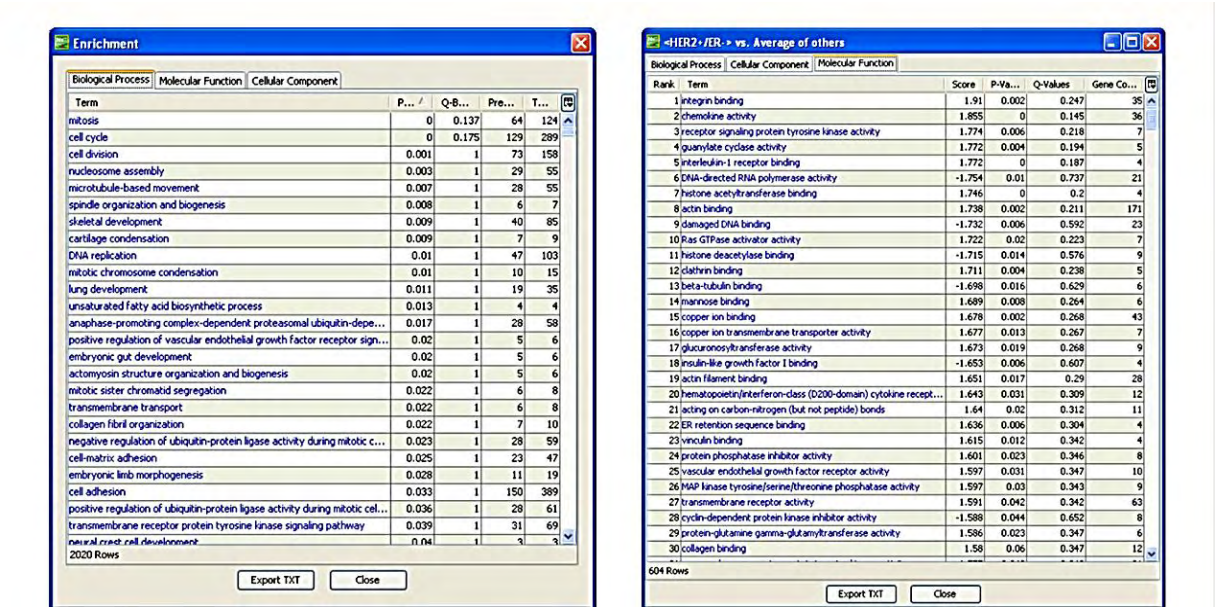
- GO Analysis

Η ανάλυση των οντολογιών διαφορικής έκφρασης γονιδίων υποστηρίζεται από το λογισμικό της εταιρείας Biodiscovery και ορισμένα στιγμιότυπα που προέρχονται από το manual του λογισμικού φαίνονται στην Εικόνα 2.4.2.

- Biochemical Pathway Analysis

Η δυνατότητα ανάλυσης βιοχημικών μονοπατιών δεν παρέχεται σε αυτήν την έκδοση του λογισμικού έπειτα από ενημέρωση των επιστημονικών συνεργατών της εταιρείας.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.4.2: Στιγμιότυπα GO Enrichment Analysis από το manual του λογισμικού

- Workflows Creations

Η δημιουργία ροών εργασιών δεν προσφέρεται από το Nexus Expression όπως επισημαίνουν οι εκπρόσωποι της.

- Clustering Analysis

Από τους ιθύνοντες της εταιρείας και από προσωπική εμπειρία που αποκτήθηκε από την εξερεύνησης του λογισμικού προκύπτει ότι αυτό προσφέρει τη δυνατότητα συσταδοποίησης μέσω επιλογής είτε του αλγόριθμου Hierarchical Clustering είτε του αλγόριθμου K-means. Ωστόσο δεν κατέστη δυνατή η ολοκλήρωση της ανάλυσης κατά συστάδες εξαιτίας της πολυπλοκότητας που εμφάνιζε αυτή για έναν μη έμπειρο σε τέτοιες διαδικασίες χρήστη.

- Custom Reports

Αναφορικά με αυτό το χαρακτηριστικό, οι υπεύθυνοι επισημαίνουν ότι η δυνατότητα που παρέχεται είναι η επιτομή των πληροφοριών της διαφορικής γονιδιακής έκφρασης σε έναν πίνακα με ξεχωριστές καρτέλες ο οποίος μπορεί να εξάγεται από το Nexus Expression με τη μορφή ενός αρχείου κειμένου.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Τρίτη κατηγορία χαρακτηριστικών

- Heat maps

Η δυνατότητα δημιουργίας Heat maps προσφέρεται από το λογισμικό όπως τονίζουν και οι επιστημονικοί συνεργάτες της εταιρείας Biomatters. Ειδικότερα, είναι δυνατή η συλλογή οποιονδήποτε πληροφοριών που σχετίζονται με τον τύπος του ιστού, τις υποομάδες και την απόκριση στα φάρμακα των δειγμάτων των εκάστοτε εξεταζόμενων οργανισμών. Η Εικόνα 2.4.3 απεικονίζει ορισμένα στιγμιότυπα χαρτών θερμότητας που προέρχονται από το manual του λογισμικού.

- Volcano/MA plots

Δεν υποστηρίζεται η δημιουργία Volcano και MA γραφικών παραστάσεων σύμφωνα με τους επιστημονικούς υπεύθυνους της εταιρείας.

- Genome Browser

Η δυνατότητα περιήγησης γονιδιωμάτων των οργανισμών δεν παρέχεται από το λογισμικό όπως τονίζουν οι εκπρόσωποι της εταιρείας.

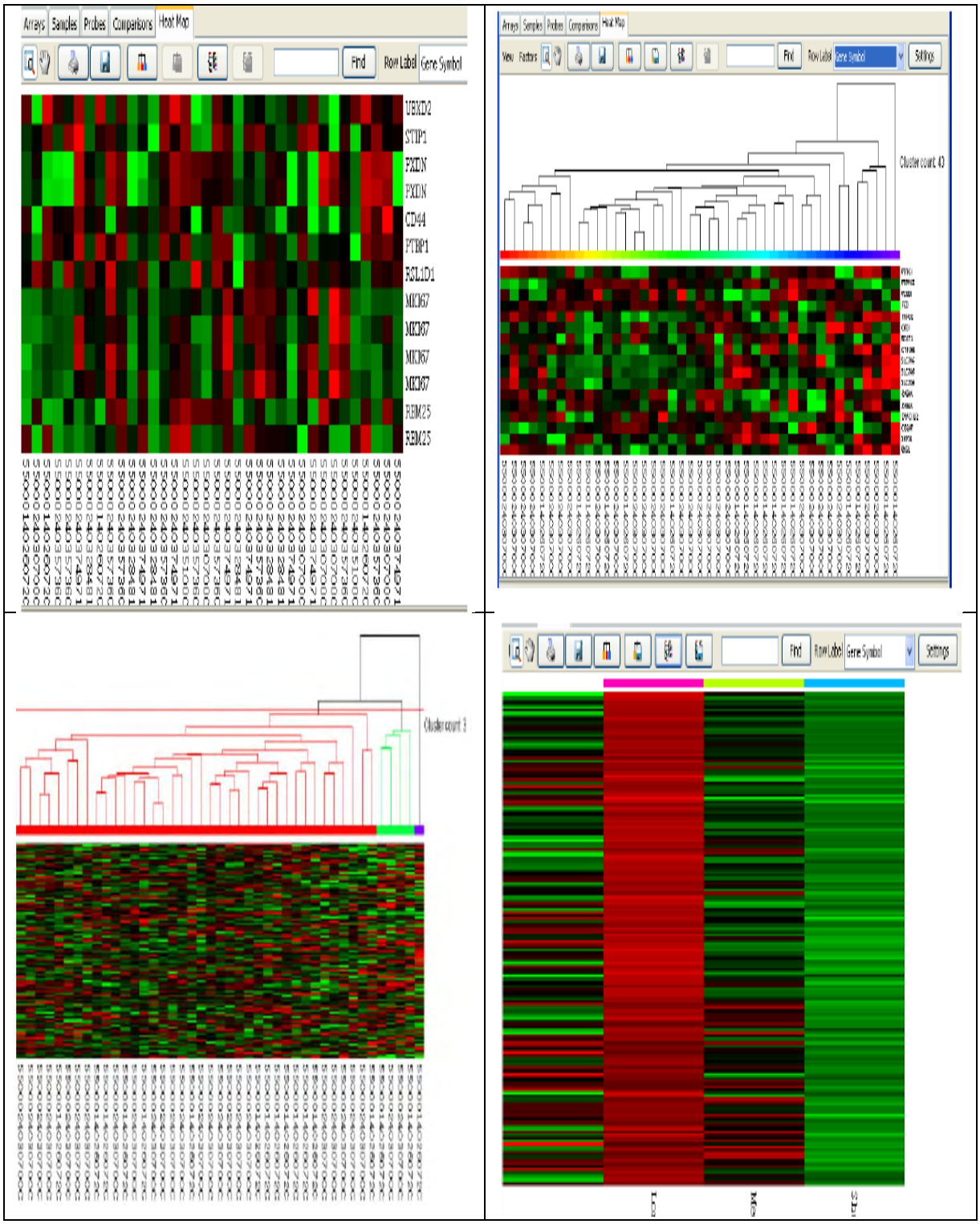
- PCA/MDS plots

Δεν είναι δυνατή η κατασκευή PCA και MDS γραφημάτων όπως προκύπτει από την επικοινωνία μέσω email με τους εξειδικευμένους συνεργάτες του λογισμικού.

- Box plots

Δεν είναι εφικτή η κατασκευή Boxplot γραφημάτων όπως υποστηρίζουν και οι αρμόδιοι της εταιρείας.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.4.3: Στιγμιότυπα Heat maps από το manual του λογισμικού Nexus Expression

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Όσον αφορά στο μηνιαίο κόστος απόκτησης άδειας χρήσης για έναν ακαδημαϊκό χρήστη αυτό ανέρχεται στα 232 € σύμφωνα με την Εικόνα 2.4.4 και αυτό προκύπτει ύστερα από επικοινωνία μέσω email με την εταιρεία Biodiscovery η οποία αποτελεί συνεργάτη του λογισμικού Nexus Expression.

Sheila Shaw <scshaw@biodiscovery.com>

προς Εμένα

Αγγλικά
Ελληνικά

Μετάφραση μηνύματος

Απενεργοποίηση για: Αγγλικά

Hi Stathis,

Thanks for your patience. Attached are the answers to your software questions. Please let me know if you have any further questions.

Regarding the price of Nexus Expression, it's a subscription based pricing model. To give you a general idea, the cost of 1-year Nexus Expression Node-Lock (single) license with a 40% academic discount is EUR 2,779.80. Please let me know if you would like a quote.

Also, below is a 7day trial license extension:
Activation code: 32cc-15e4-c361-44cc-8f59-a1d3-2e10-3dcb

Thanks
Sheila

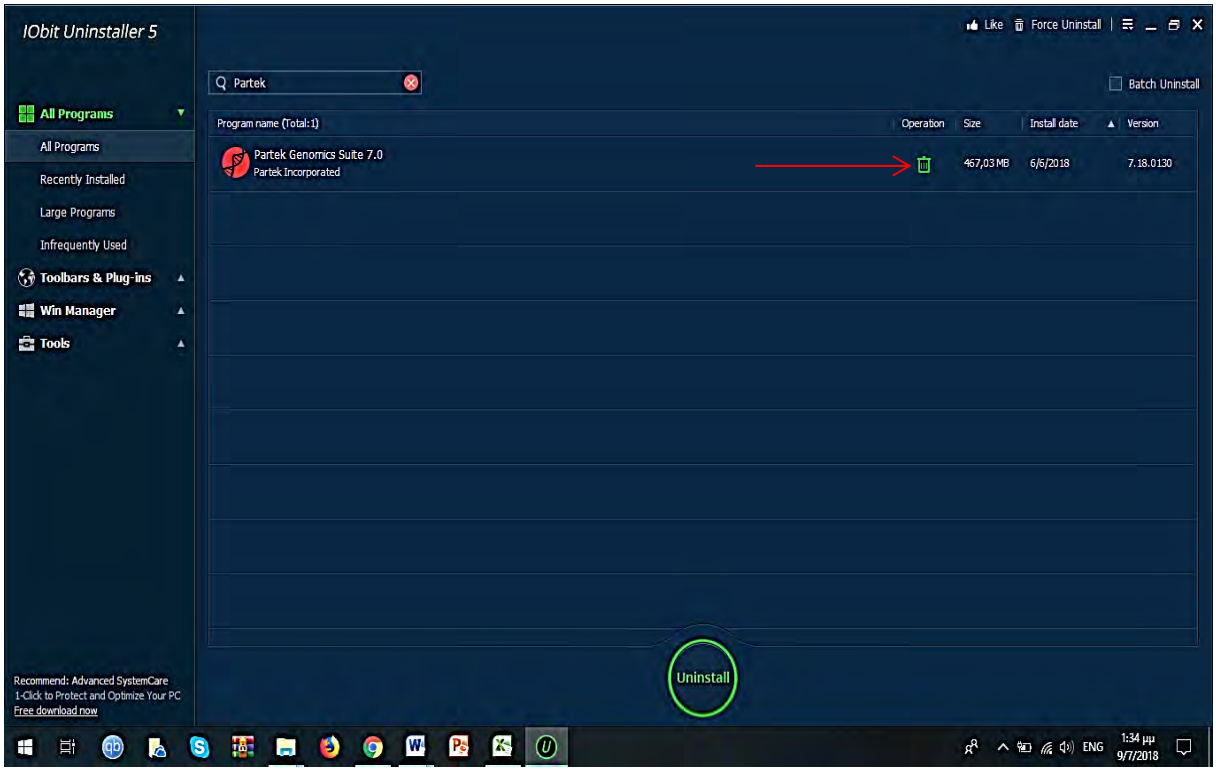
Εικόνα 2.4.4: Το ετήσιο κόστος άδειας χρήσης του λογισμικού Nexus Expression

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.5 Partek Flow (Partek)

Πρώτη κατηγορία χαρακτηριστικών

Στην Εικόνα 2.5.1 φαίνεται ότι ο χώρος που καταλαμβάνει το λογισμικό Partek Genomics Suite της εταιρείας Partek είναι 467 MB περίπου. Ωστόσο, η χωρητικότητα του Partek Flow μοιάζει να είναι μηδενική. Η αρχιτεκτονική συστήματος είναι Desktop. Επιπροσθέτως, όσον αφορά στο λειτουργικό σύστημα το λογισμικό υποστηρίζει και τους 3 τύπους των συστημάτων σύμφωνα με την ιστοσελίδα της αμερικάνικης εταιρείας.



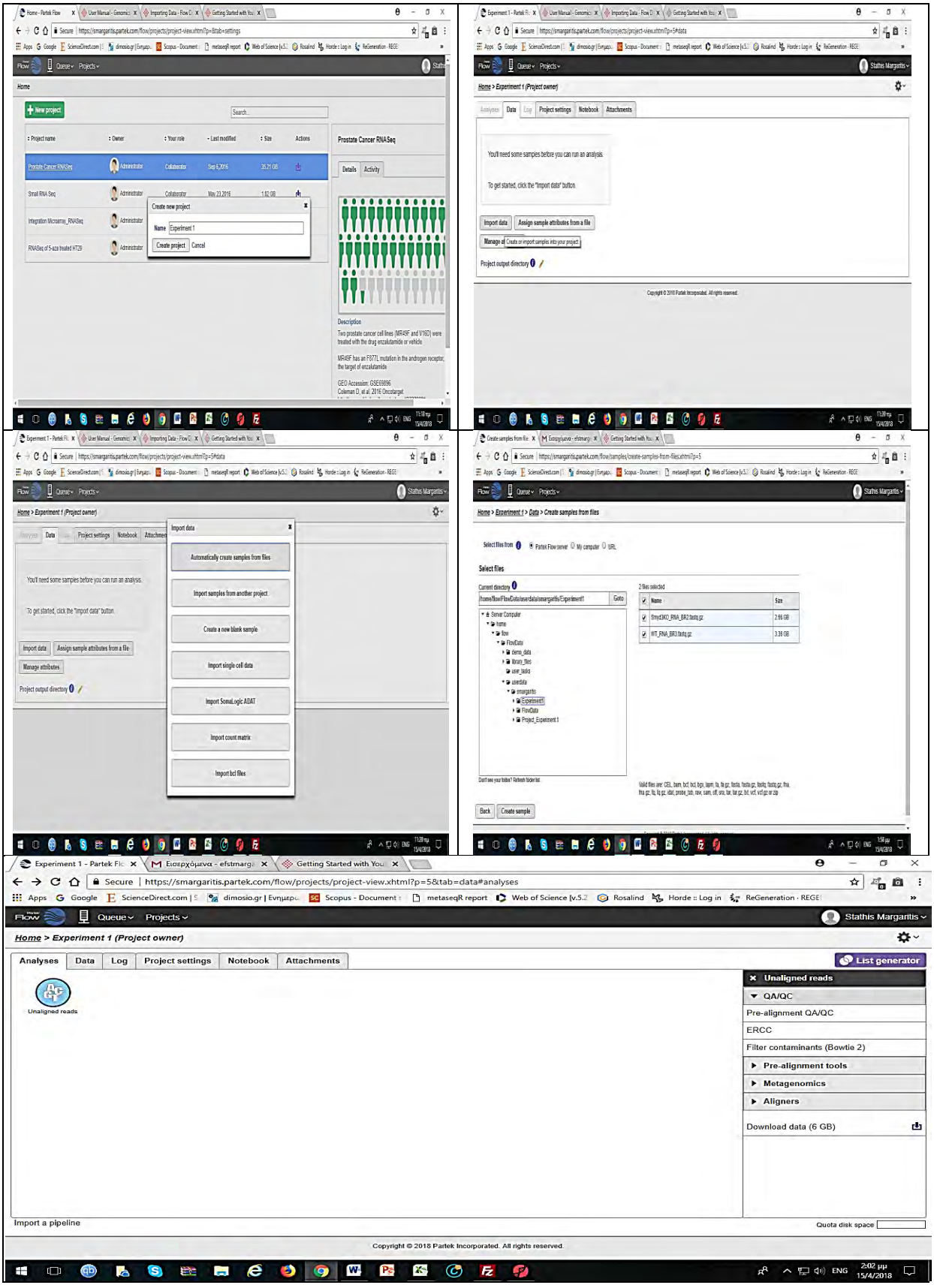
Εικόνα 2.5.1: Χωρητικότητα του λογισμικού Partek Genomics Suite

Δεύτερη κατηγορία χαρακτηριστικών

- FASTQ/BAM αρχεία

Σύμφωνα με τους επιστημονικούς συνεργάτες της Partek το λογισμικό Partek Flow δέχεται ως είσοδο τόσο FASTQ αρχεία (μη ευθυγραμμισμένα, unaligned) όσο και BAM/SAM αρχεία (ευθυγραμμισμένα). Επιπλέον, το λογισμικό αυτό δέχεται και άλλες μορφές συμπιεσμένων αρχείων όπως είναι τα gzipped fastq αρχεία. Η Εικόνα 2.5.2 δείχνει τα βήματα εισαγωγής ενός FASTQ αρχείου σε αυτό το λογισμικό.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.5.2: Διαδικασία εισαγωγής FASTQ αρχείων στο λογισμικό Partek Flow

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Organisms

Οι εξειδικευμένοι συνεργάτες της εταιρείας επισημαίνουν ότι παρέχεται η δυνατότητα επεξεργασίας γονιδιωμάτων αρκεί να υπάρχει ένα γονιδίωμα αναφοράς του οργανισμού και ένα αρχείο σχολιασμού για αυτόν. Ειδικότερα, υποστηρίζονται οι αυτόματες λήψεις γονιδιωμάτων ανθρώπων, ποντικών και αρουραίων.

- Quality Control

Το λογισμικό προσφέρει ποιοτικό έλεγχο τόσο στην αρχή όσο και στο τέλος της διαδικασίας της ευθυγράμμισης όπως τονίζουν οι υπεύθυνοι. Η Εικόνα 2.5.3 απεικονίζει τη διαδικασία του ποιοτικού ελέγχου για ένα Smyd3KO δείγμα και ένα WT δείγμα του εξεταζόμενου ποντικού mus musculus. Από το αριστερό διάγραμμα του τελευταίου στιγμιότυπου της εικόνας φαίνεται ότι το PHRED score του Smyd3KO δείγματος ισούται με 22,7 περίπου ενώ του WT δείγματος είναι ίσο με 23,5 περίπου. Σε κάθε περίπτωση αφού και τα δύο σκορ είναι μικρότερα από 30 αυτό που μπορεί να προκύψει ως συμπέρασμα είναι ότι η συγκεκριμένη κατανομή δεν έχει την απαιτούμενη ακρίβεια για να γίνει αποδεκτή. Στο ίδιο συμπέρασμα μπορεί να καταλήξει κάποιος μελετώντας το διάγραμμα της δεξιάς εικόνας στο οποίο διακρίνεται ότι το PHRED score για την αλληλούχιση κάθε νουκλεοτιδικής βάσης του WT δείγματος ανέρχεται σε 27 έναντι 22,5 του Smyd3KO δείγματος και η κατανομή που ακολουθεί αποκλίνει περισσότερο από την κανονική συγκριτικά με την κατανομή του Smyd3KO δείγματος.

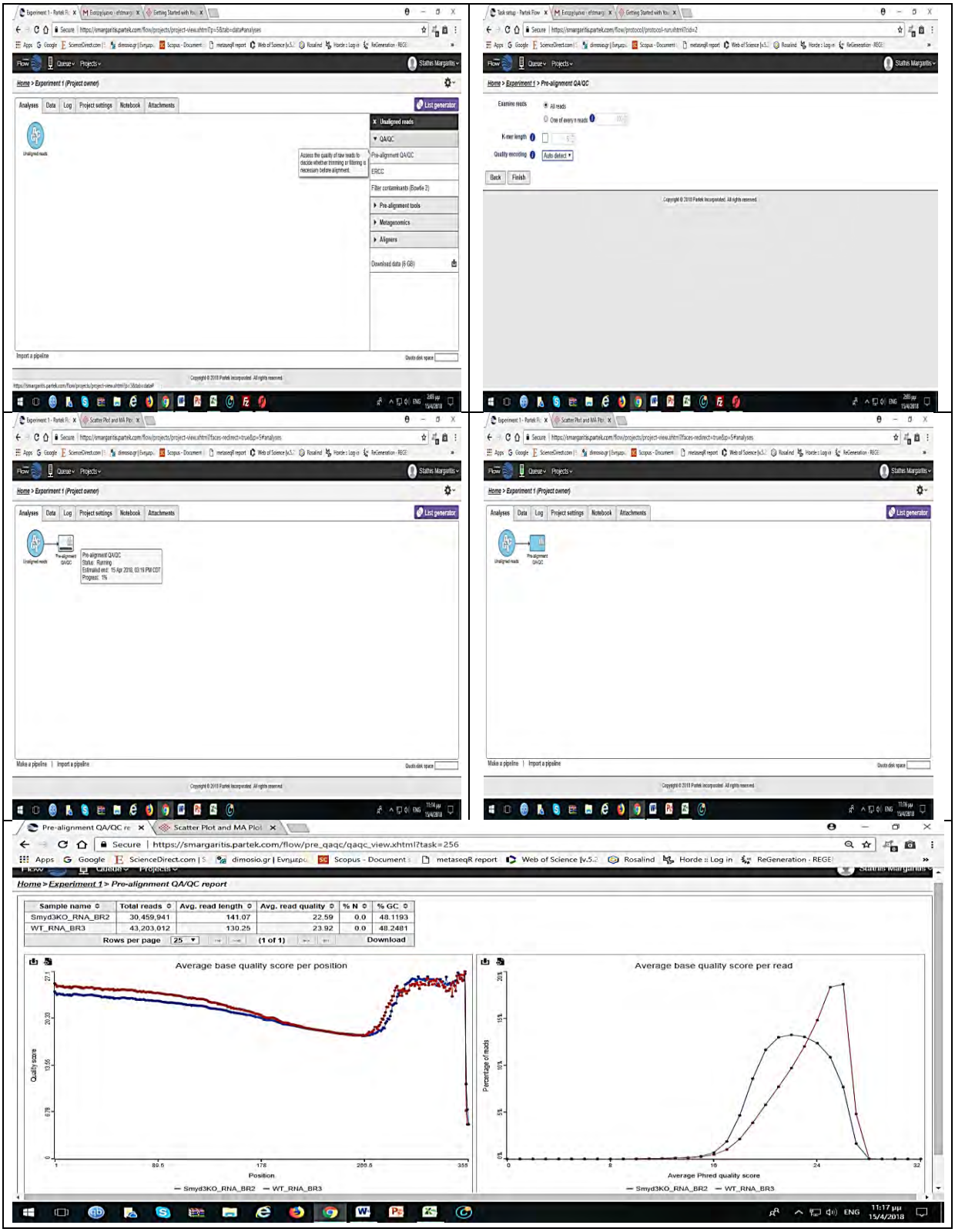
- Differential Expression

Σύμφωνα με τους ιθύνοντες της εταιρείας Partek διατίθεται ένα ευρύ φάσμα στατιστικών μεθόδων που είναι ειδικά σχεδιασμένες για RNA-Seq ανάλυση διαφορικής γονιδιακής έκφρασης. Η Εικόνα 2.5.4 συνοψίζει τη διαδικασία που ακολουθείται σε 2 δείγματα του εξεταζόμενου οργανισμού.

- External Annotation Databases

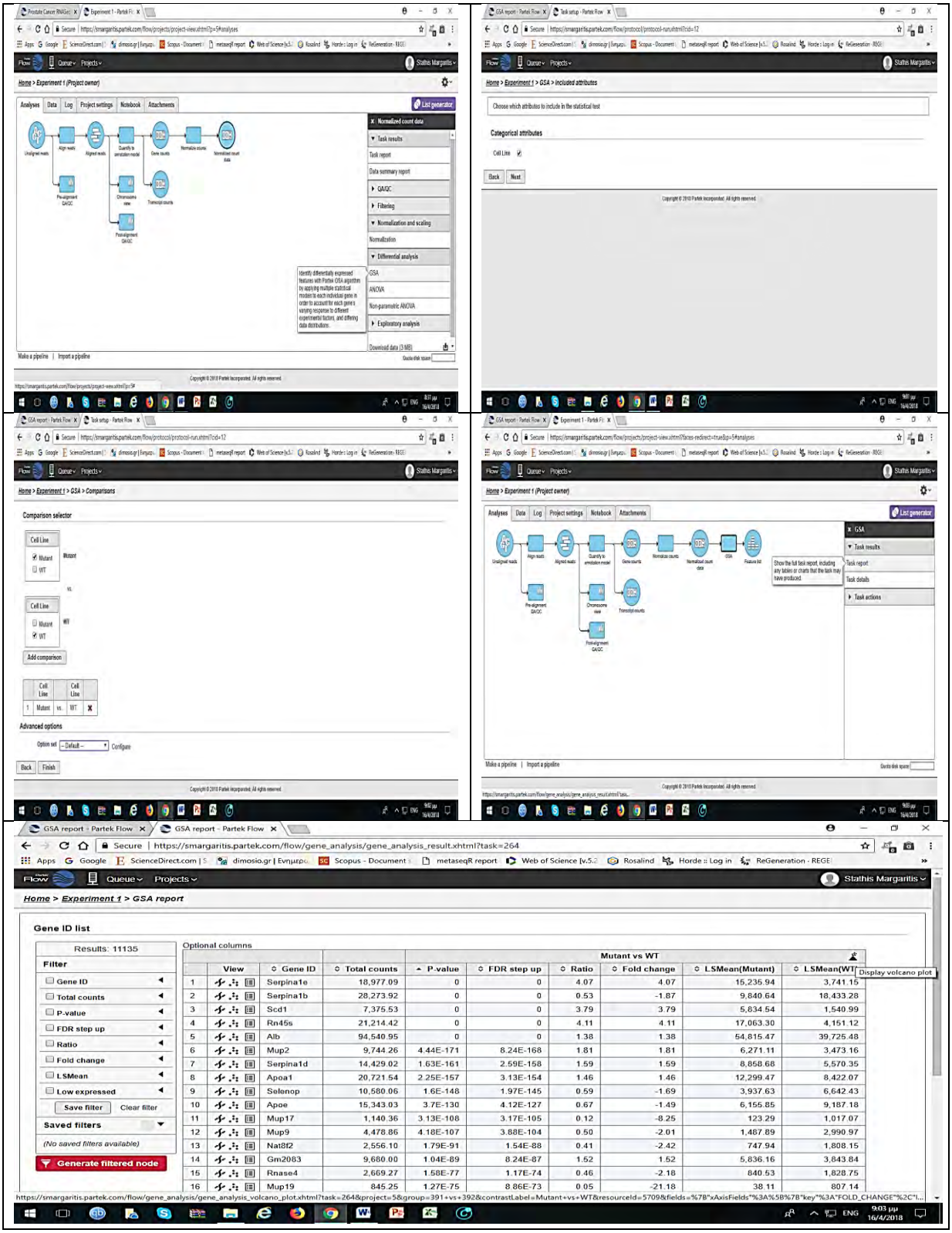
Οι αρμόδιοι της εταιρείας διατείνονται ότι για πολλούς οργανισμούς-μοντέλα παρέχονται αυτόματες λήψεις σχολιασμού της διαφορικής έκφρασης των γονιδίων από τις βάσεις δεδομένων ENSEMBL, RefSeq, GENCODE, miRBase και geneontology.org.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.5.3: Στιγμιότυπα από τη διαδικασία ποιοτικού ελέγχου 2 δειγμάτων του mus musculus

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

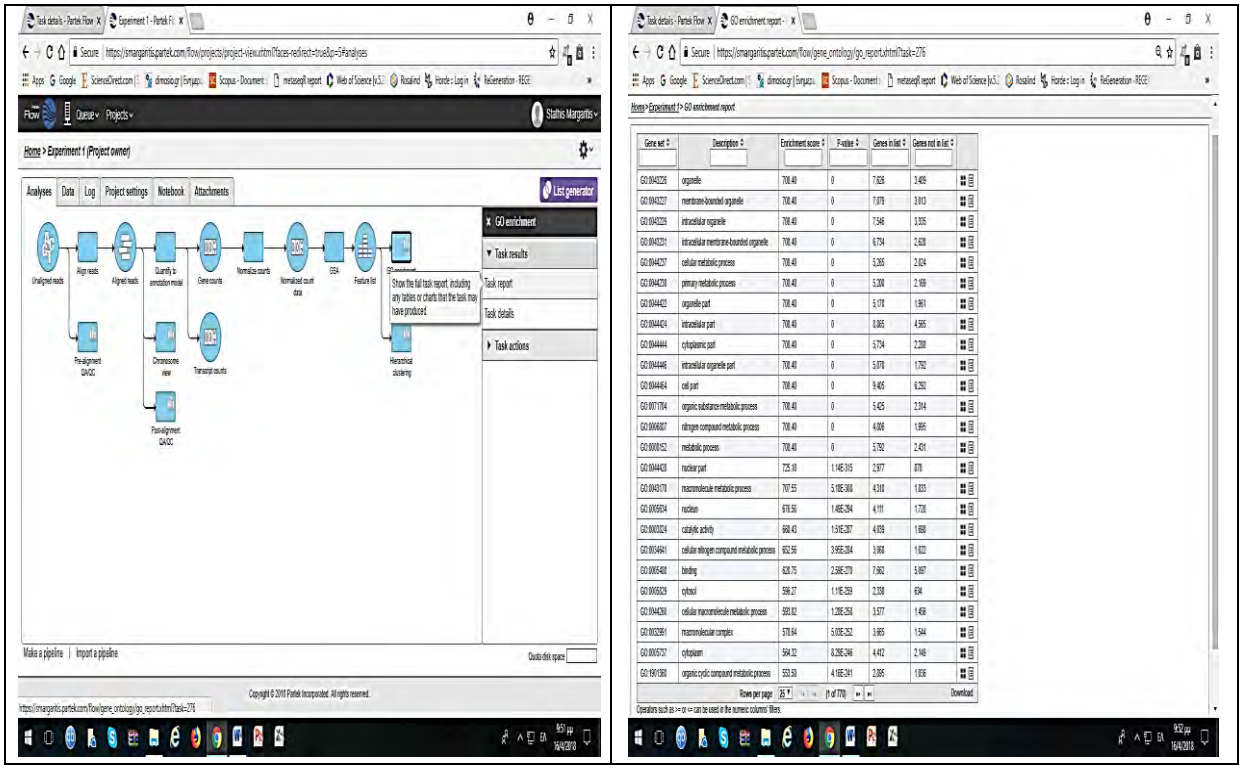


Εικόνα 2.5.4: Στιγμιότυπα από τη διαδικασία διαφορικής γονιδιακής έκφρασης

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- GO Analysis

Η ανάλυση εμπλουτισμένης οντολογίας της διαφορικής έκφρασης γονιδίων μπορεί να εκτελεστεί εφόσον έχει προκύψει μία τελική λίστα γονιδίων. Η Εικόνα 2.5.5 δείχνει μία τέτοια λίστα γονιδίων που αφορά 2 σε δείγματα του εξεταζόμενου οργανισμού.



Εικόνα 2.5.5: Ανάλυση εμπλουτισμένης γονιδιακής οντολογίας για 2 δείγματα του mus musculus

- Biochemical Pathway Analysis

Η ανάλυση βιοχημικών μονοπατιών δεν παρέχεται από το Partek Flow. Αντιθέτως, είναι απαραίτητη η εκτέλεση ενός άλλου λογισμικού με συνδρομή για το χρήστη που ονομάζεται Partek Pathway και αποτελεί υποπρόγραμμα του λογισμικού Partek Genomics Suite.

- Workflows Creations

Η δημιουργία ροών εργασιών είναι ένα χαρακτηριστικό που υποστηρίζεται από το λογισμικό όπως ισχυρίζονται οι επιστημονικοί υπεύθυνοι της εταιρείας. Ειδικότερα, υποστηρίζουν ότι είναι απαραίτητη η ακολουθία μιας σειράς βημάτων για την επεξεργασία και την ανάλυση των δεδομένων που μπορεί να αποθηκευθεί και να επαναχρησιμοποιηθεί σε μελλοντικές περιπτώσεις.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Clustering Analysis

Η ανάλυση κατά συστάδες είναι δυνατή μέσω της εφαρμογής μιας σειράς μεθόδων όπως είναι ο αλγόριθμος Hierarchical clustering και ο αλγόριθμος K-means. Παρ' όλα αυτά δεν κατέστη δυνατή η εύρεση κάποιας εικόνας που να προέρχεται είτε από το manual ή από την ανάλυση του εξεταζόμενου οργανισμού εξαιτίας της πολυπλοκότητας της διαδικασίας για έναν άπειρο με αυτά τα λογισμικά χρήστη.

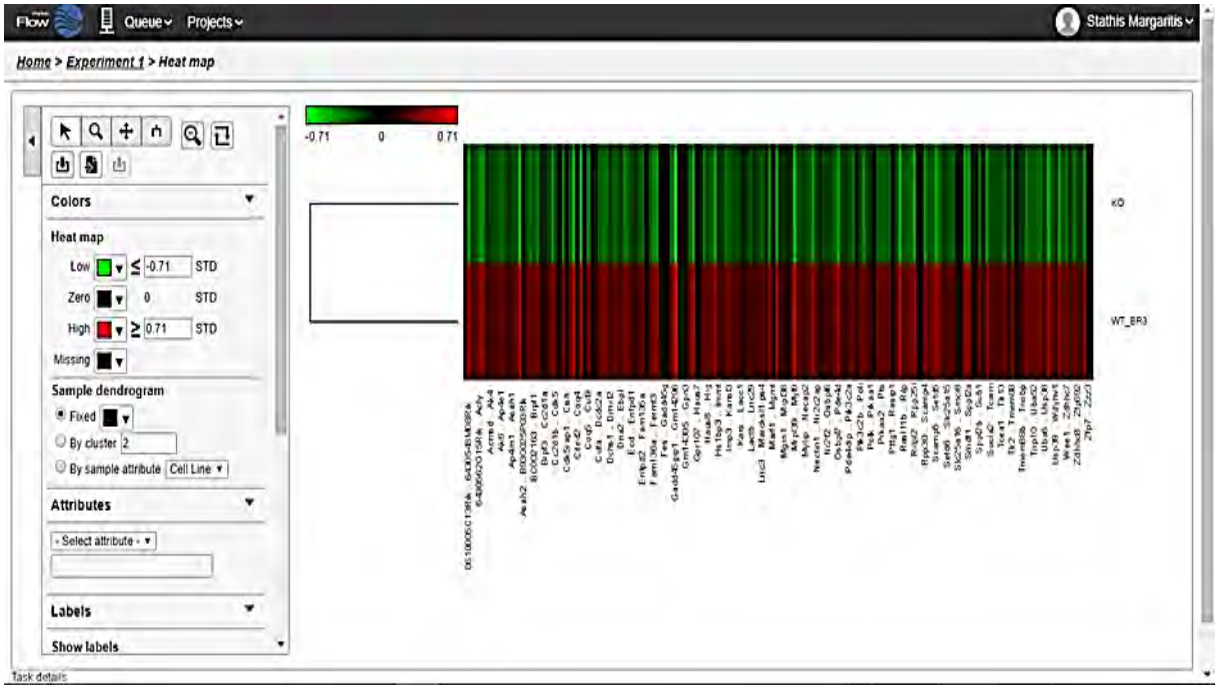
- Custom Reports

Η δημιουργία αναφορών επιλεγμένων γονιδίων που αποτελούν αντικείμενο ενδιαφέροντος του εκάστοτε χρήστη δεν είναι διαθέσιμη στην παρούσα έκδοση του λογισμικού αλλά αναμένεται να κυκλοφορήσει μία νέα έκδοση με ενσωματωμένη αυτή τη δυνατότητα.

Τρίτη κατηγορία χαρακτηριστικών

- Heat maps

Η δυνατότητα δημιουργίας χαρτών θερμότητας παρέχεται από το λογισμικό σύμφωνα και με την Εικόνα 2.5.6 η οποία απεικονίζει Heat maps διαγράμματα που προκύπτουν από την ανάλυση της διαφορικής γονιδιακής έκφρασης 2 δειγμάτων του εξεταζόμενου οργανισμού.

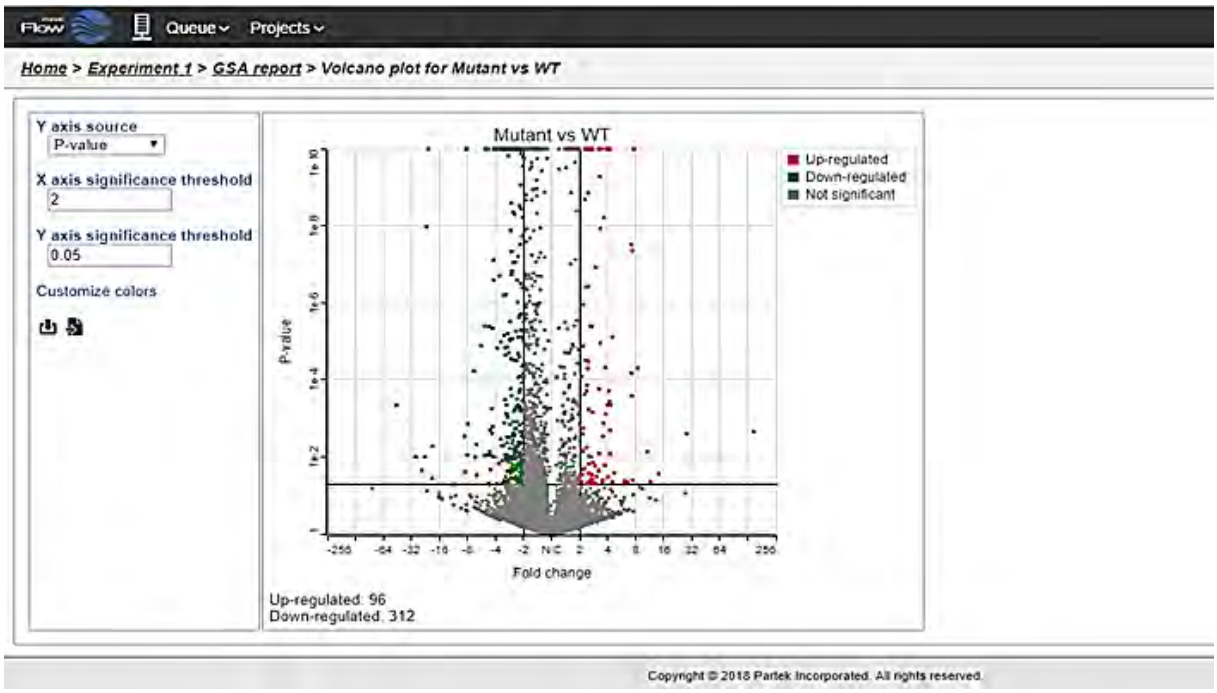


Εικόνα 2.5.6: Heat maps για 2 δείγματα (WT,Smyd3KO) του mus musculus

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Volcano plots

Το συγκεκριμένο λογισμικό παρέχει τη δυνατότητα δημιουργίας Volcano γραφικών παραστάσεων όπως δείχνει και η Εικόνα 2.5.7. Η εικόνα αυτή δείχνει τη διαφορά που παρατηρείται στην έκφραση των γονιδίων του εξεταζόμενου οργανισμού μεταξύ 2 διαφορετικών βιολογικών συνθηκών όπου το Smyd3KO δείγμα αναφέρεται ως Mutant. Ακόμη, από την ίδια εικόνα φαίνεται ότι τα υπο-εκφρασμένα γονίδια είναι πολύ περισσότερα σε σχέση με τα υπερ-εκφρασμένα (312 έναντι 96).



Εικόνα 2.5.7: Volcano διάγραμμα για 2 δείγματα του εξεταζόμενου ποντικού

- MA plots

Η κατασκευή MA γραφικών παραστάσεων είναι δυνατή μόνο με τη χρησιμοποίηση του λογισμικού Partek Genomics Suite. Βέβαια, παρέχεται η δυνατότητα για κατασκευή άλλων γραφικών παραστάσεων που είναι παρόμοιες με τις MA για τις οποίες δεν κατέστη δυνατή η συλλογή κάποιας εικόνας.

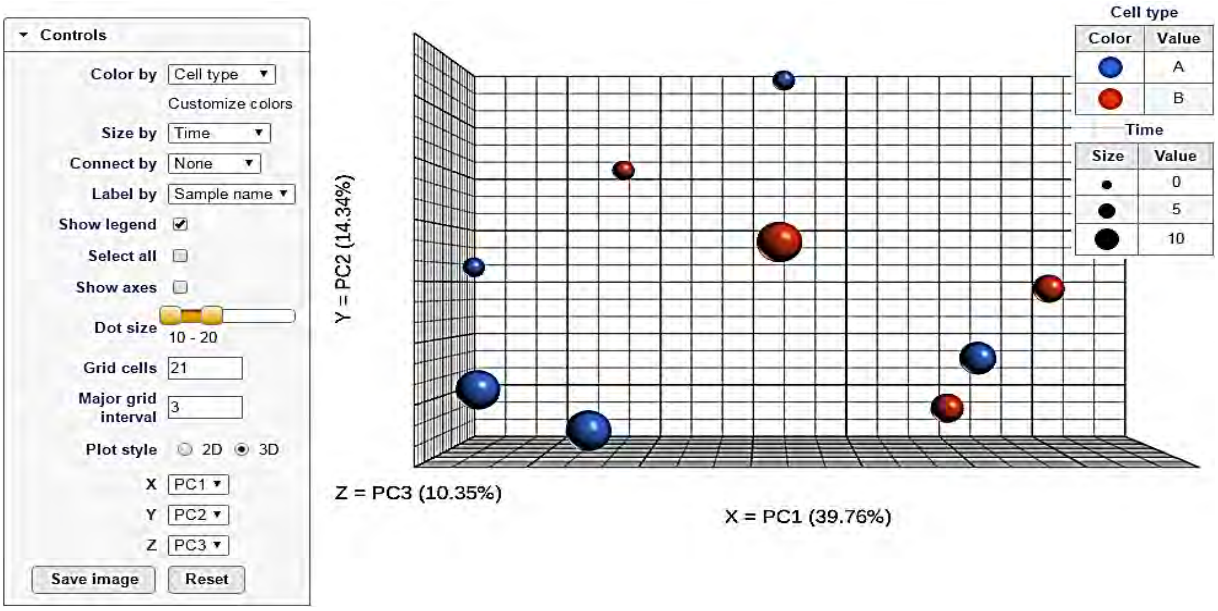
- Genome Browser

Οι επιστημονικοί συνεργάτες του λογισμικού ισχυρίζονται ότι το Partek Flow είναι εξοπλισμένο με το δικό του πρόγραμμα επεξεργασίας χρωμοσωμάτων το οποίο δεν διευκρινίζεται αν είναι παρόμοιο με το UCSC Genome Browser ή με το JBrowse.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- PCA plots

Η ανάλυση κύριων συνιστωσών προσφέρεται από το λογισμικό και ένα στιγμιότυπο αυτής σε τρισδιάστατη μορφή, όπως φαίνεται στο manual, ακολουθεί στην Εικόνα 2.5.8. Δεν κατέστη δυνατή η δημιουργία τέτοιων γραφημάτων για τον εξεταζόμενο οργανισμό λόγω ανεπάρκειας σε μνήμη του υπολογιστή δοκιμής.



Εικόνα 2.5.8: 3-D PCA γράφημα από το manual του λογισμικού

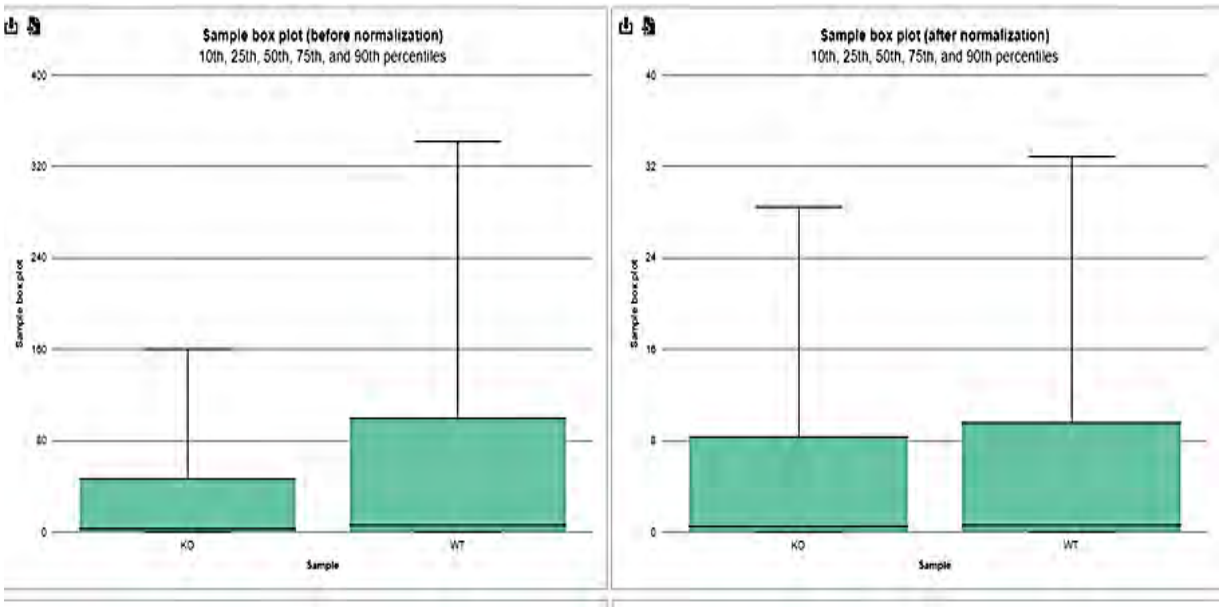
- MDS plots

Η ανάλυση πολυδιάστατης κλιμάκωσης υποστηρίζεται μόνο από το λογισμικό Partek Genomics Suite όπως επισημαίνουν και οι ειδικοί της Partek. Ωστόσο, δεν κατέστη δυνατή η συλλογή κάποιας εικόνας ή στιγμιότυπου λόγω τεχνικών δυσκολιών κατά τη διαδικασία δημιουργίας τέτοιου είδους γραφημάτων.

- Box plots

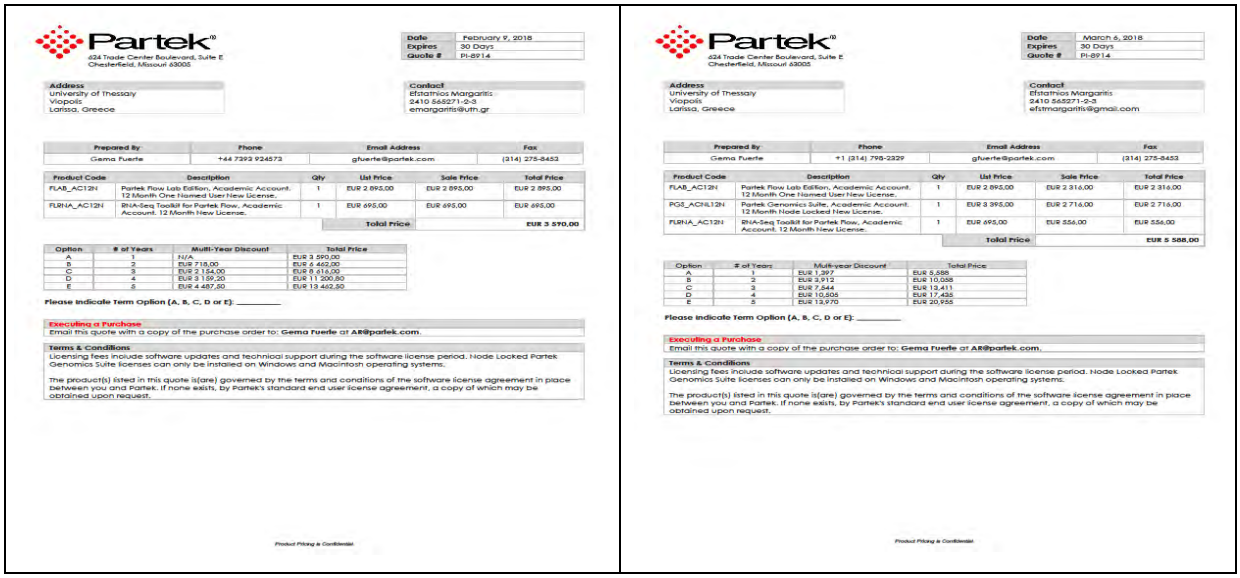
Η κατασκευή Box plot διαγραμμάτων αποτελεί τμήμα του ποιοτικού ελέγχου που παρέχει το λογισμικό Partek Flow και η Εικόνα 2.5.9 φανερώνει την ύπαρξη αυτών των διαγραμμάτων. Αυτό που φαίνεται από την παρακάτω εικόνα είναι ότι η μορφή των κανονικοποιημένων Box plot διαγραμμάτων που προέρχονται από την ανάλυση της διαφορικής γονιδιακής έκφρασης 2 δειγμάτων του εξεταζόμενου οργανισμού είναι παρόμοια. Ειδικότερα, το ελάχιστο (10%) φαίνεται να είναι το ίδιο όπως και το πρώτο τεταρτημόριο (25%) ενώ η διάμεσος (50%), το τρίτο τεταρτημόριο (75%) και το μέγιστο (90%) τείνουν να λαμβάνουν υψηλότερες τιμές στο WT δείγμα κάτι το οποίο είναι αναμενόμενο λόγω ύπαρξης του γονιδίου Smyd3.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.5.9: Box plot διαγράμματα για 2 δείγματα (Smyd3KO, WT) του mus musculus

Όσον αφορά στο μηνιαίο κόστος απόκτησης άδειας χρήσης του λογισμικού Partek Flow για έναν ακαδημαϊκό χρήστη, αυτό ανέρχεται περίπου στα 300 € σύμφωνα με την αριστερή στήλη της Εικόνας 2.5.10 και αυτό προκύπτει ύστερα από επικοινωνία μέσω email με την εταιρεία Partek. Στην ίδια εικόνα φαίνεται και το μηνιαίο κόστος απόκτησης άδειας χρήσης συνολικά για το λογισμικό Partek Genomics Suite το οποίο ανέρχεται περίπου στα 282 €.



Εικόνα 2.5.10: Το ετήσιο κόστος άδειας χρήσης για τα λογισμικά Partek Flow και Partek Genomics Suite

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.6 Rosalind (OnRamp Bioinformatics)

Πρώτη κατηγορία χαρακτηριστικών

Η επιθεώρηση του πιθανού χώρου που καταλαμβάνει το λογισμικό Rosalind της εταιρείας OnRamp φαίνεται να είναι μηδενική. Η αρχιτεκτονική συστήματος είναι Cloud και σε αυτήν την περίπτωση. Όσον αφορά στο λειτουργικό σύστημα το λογισμικό υποστηρίζει πιθανότατα και από τους 3 τύπους των λειτουργικών συστημάτων χωρίς όμως να υπάρχει κάποια συγκεκριμένη πληροφορία στην ιστοσελίδα της αμερικάνικης εταιρείας.

Δεύτερη κατηγορία χαρακτηριστικών

- FASTQ αρχεία

Είναι δυνατή η εισαγωγή FASTQ αρχείων στα οποία στη συνέχεια ακολουθείται η διαδικασία του trimming και της ευθυγράμμισης όπως επισημαίνει ο υπεύθυνος του λογισμικού. Ωστόσο δεν κατέστη εφικτή η συλλογή κάποιου στιγμιότυπου που να αφορά αυτό το χαρακτηριστικό.

- BAM αρχεία

Αντιθέτως, η εισαγωγή BAM αρχείων δεν υποστηρίζεται άμεσα από το λογισμικό Rosalind αλλά μόνο ύστερα από προσωπική επικοινωνία με κάποιον επιστημονικό υπεύθυνο της OnRamp Bioinformatics.

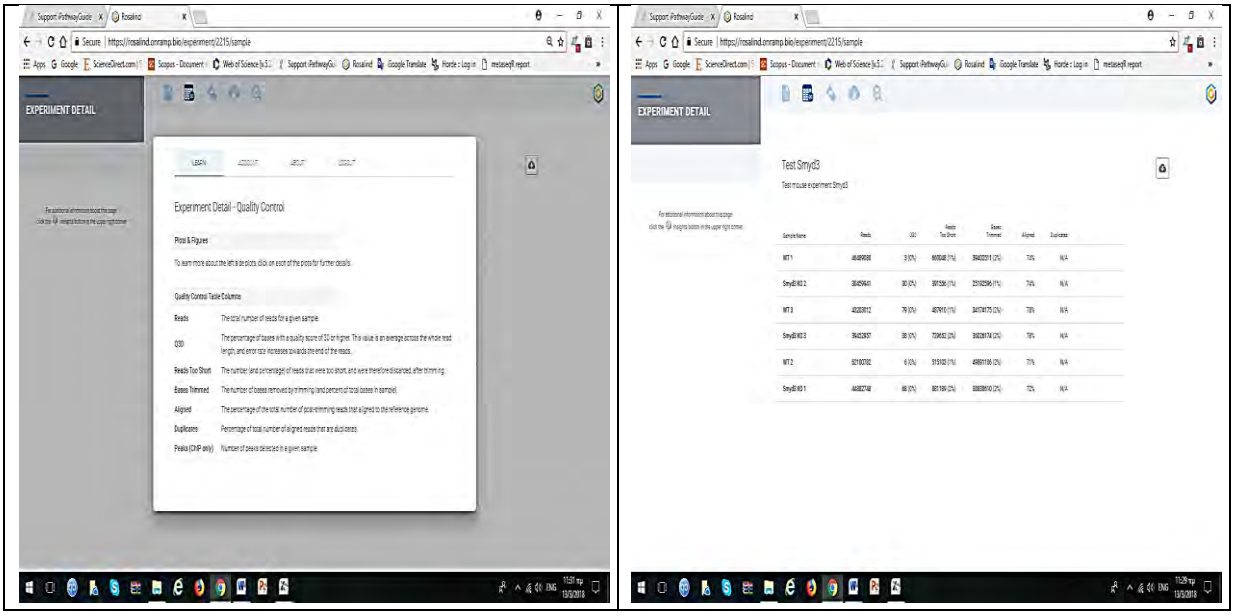
- Organisms

Το λογισμικό περιλαμβάνει γονιδιώματα ανθρώπων, ποντικών, αρουραίων, σκουληκιών, μαγιάς (yeast) καθώς και γονιδιώματα των οργανισμών drosophilia και Zebrafish. Επίσης, προσφέρεται η δυνατότητα για προσθήκη γονιδιώματος οποιουδήποτε οργανισμού ζητηθεί υπό τον όρο να συνοδεύεται από αποδεκτή αλληλουχία γονιδιώματος και σχολιασμό μεταγραφής της διαφορικής γονιδιακής έκφρασης (transcript annotation).

- Quality Control

Η Εικόνα 2.6.1 παρουσιάζει τα αποτελέσματα του ποιοτικού ελέγχου για τον εξεταζόμενο οργανισμό όπως φαίνονται στο λογισμικό. Πιο συγκεκριμένα, στην αριστερή στήλη δίνονται οι ορισμοί των 6 κριτηρίων ποιοτικού ελέγχου ενώ στη δεξιά στήλη τα αποτελέσματα που αφορούν 6 δείγματα (3 WT και 3 Smyd3KO) του εξεταζόμενου ποντικού.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.6.1: Στιγμιότυπα ποιοτικού ελέγχου των δειγμάτων του mus musculus

Από μεγέθυνση του στιγμιότυπου της δεξιάς στήλης φαίνεται ότι ο αριθμός των αναγνώσεων για το WT2 δείγμα είναι κατά 25% μεγαλύτερος από τον αντίστοιχο αριθμό του WT1 δείγματος και κατά 30% από το WT3 δείγμα. Στα Smyd3KO δείγματα ο αριθμός των αναγνώσεων του πρώτου δείγματος είναι 23% μεγαλύτερος από τον αριθμό των αναγνώσεων του δεύτερου δείγματος και 12% από τον αντίστοιχο αριθμό του τρίτου Smyd3KO δείγματος. Όσον αφορά στο δεύτερο κριτήριο το οποίο σχετίζεται με το ποσοστό των βάσεων που σημειώνουν quality PHRED score ≥ 30, αυτό που προκύπτει από τα αποτελέσματα της δεξιάς στήλης είναι ότι κανένα δείγμα δεν ξεπερνά αυτό το όριο και επομένως η κατανομή της διαφορικής έκφρασης των γονιδίων δεν έχει την απαραίτητη ακρίβεια για να γίνει αποδεκτή.

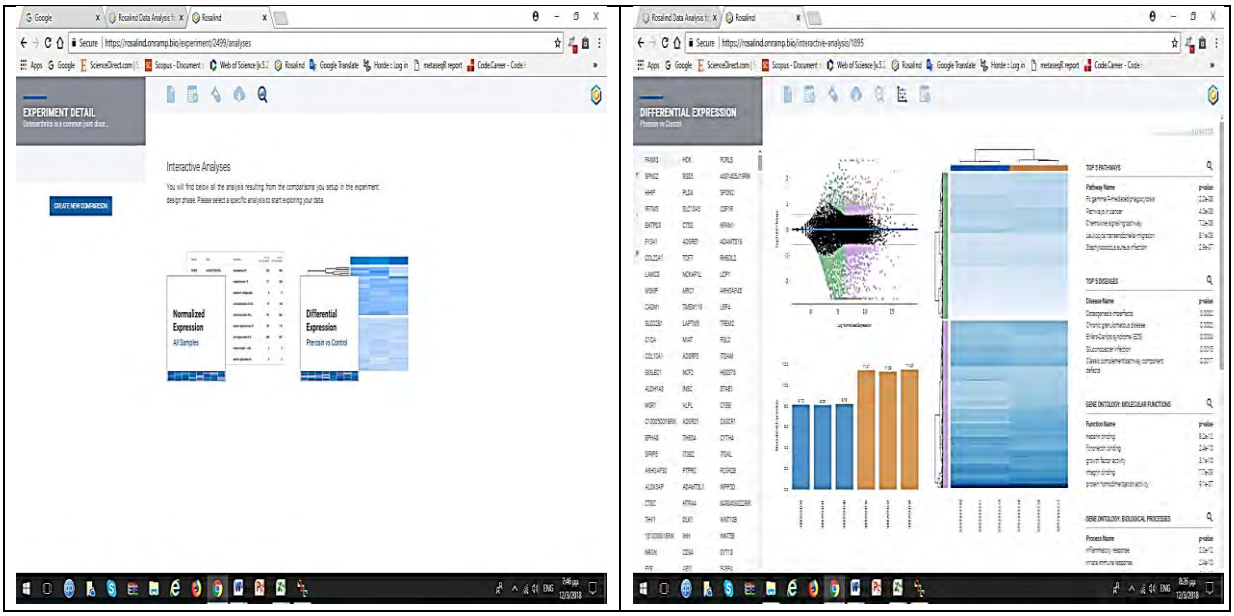
Το ποσοστό των αναγνώσεων των αλληλουχιών που ήταν πολύ μικρές και άρα αμελητέες σε σχέση με τις υπόλοιπες αλληλουχίες ανέρχεται μόλις στο 1-2% σε όλα τα δείγματα. Στα ίδια επίπεδα κυμαίνεται και το ποσοστό των βάσεων ,για όλα τα δείγματα, που αφαιρέθηκαν με τη διαδικασία trimming.

Αντιθέτως το ποσοστό των αναγνώσεων των αλληλουχιών του εξεταζόμενου οργανισμού που παρέμειναν μετά τη διαδικασία του trimming διακυμαίνεται από 71-78% για όλα τα δείγματα και των δύο διαφορετικών βιολογικών συνθηκών. Τέλος, δεν προκύπτει κάποιο αξιοσημείωτο αποτέλεσμα όσον αφορά στο ποσοστό των ευθυγραμμισμένων αναγνώσεων των αλληλουχιών που επαναλαμβάνονται περισσότερες από μία φορά.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Differential Expression

Η ανάλυση της διαφορικής έκφρασης γονιδίων πραγματοποιείται μέσω συγκρίσεων με διαφορετικό επίπεδο φιλτραρίσματος και βασίζεται στις διαφορές στο επίπεδο έκφρασης των γονιδίων (fold changes) και στις τιμές του στατιστικού μέτρου σημαντικότητας (pvalues). Η Εικόνα 2.6.2 περιέχει κάποια ενδεικτικά αποτελέσματα της ανάλυσης των δειγμάτων του εξεταζόμενου οργανισμού.



Εικόνα 2.6.2: Στιγμιότυπα από την ανάλυση διαφορικής έκφρασης γονιδίων των δειγμάτων του mus musculus

- External Annotation Databases

Η δυνατότητα σύνδεσης της διαφορικής γονιδιακής έκφρασης με εξωτερικές βάσεις δεδομένων υποστηρίζεται μόνο στην περίπτωση που είναι επαρκείς οι πληροφορίες οι οποίες σχετίζονται με το γονιδίωμα του εκάστοτε εξεταζόμενου οργανισμού. Από την απάντηση του αρμόδιου συνεργάτη της εταιρείας προκύπτει ότι στην ανωτέρω περίπτωση είναι δυνατή η πληροφόρηση για την ύπαρξη ανεξάρτητων γονιδίων (individual genes) καθώς και για το σύνολο αυτών μέσω μιας παρεχόμενης από το λογισμικό λίστας.

- GO Analysis

Η ανάλυση οντολογιών γονιδιακής έκφρασης δεν υποστηρίζεται από την παρούσα έκδοση του λογισμικού αλλά προσφέρεται με επιπλέον χρέωση μέσω ενός άλλου λογισμικού (iPathwayGuide) το οποίο παρέχει μία εταιρεία που αποτελεί συνεργάτη της OnRamp Bioinformatics και ονομάζεται Advaitabio.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Biochemical Pathway Analysis

Η ανάλυση βιοχημικών μονοπατιών υποστηρίζεται και αυτή μέσω του προγράμματος iPathwayGuide της εταιρείας Advaitabio. Επιπρόσθετα, αξίζει να σημειωθεί ότι προσφέρεται η δυνατότητα εμπλουτισμού μονοπατιών (enrichment metrics) καθώς και διαταραχών της πορείας (pathway perturbation).

- Workflows Creations

Δεν υποστηρίζεται η δυνατότητα ροής εργασιών σύμφωνα με την απάντηση που έδωσε ο υπεύθυνος του λογισμικού Rosalind.

- Clustering Analysis

Η ανάλυση κατά συστάδες παρέχεται χρησιμοποιώντας τον αλγόριθμο K-means ή/ και με τα MDS γραφήματα όπως τονίζουν οι ειδικοί της εταιρείας. Ωστόσο, δεν κατέστη εφικτή η συλλογή εικόνων λόγω της πολυπλοκότητας της όλης διαδικασίας.

- Custom Reports

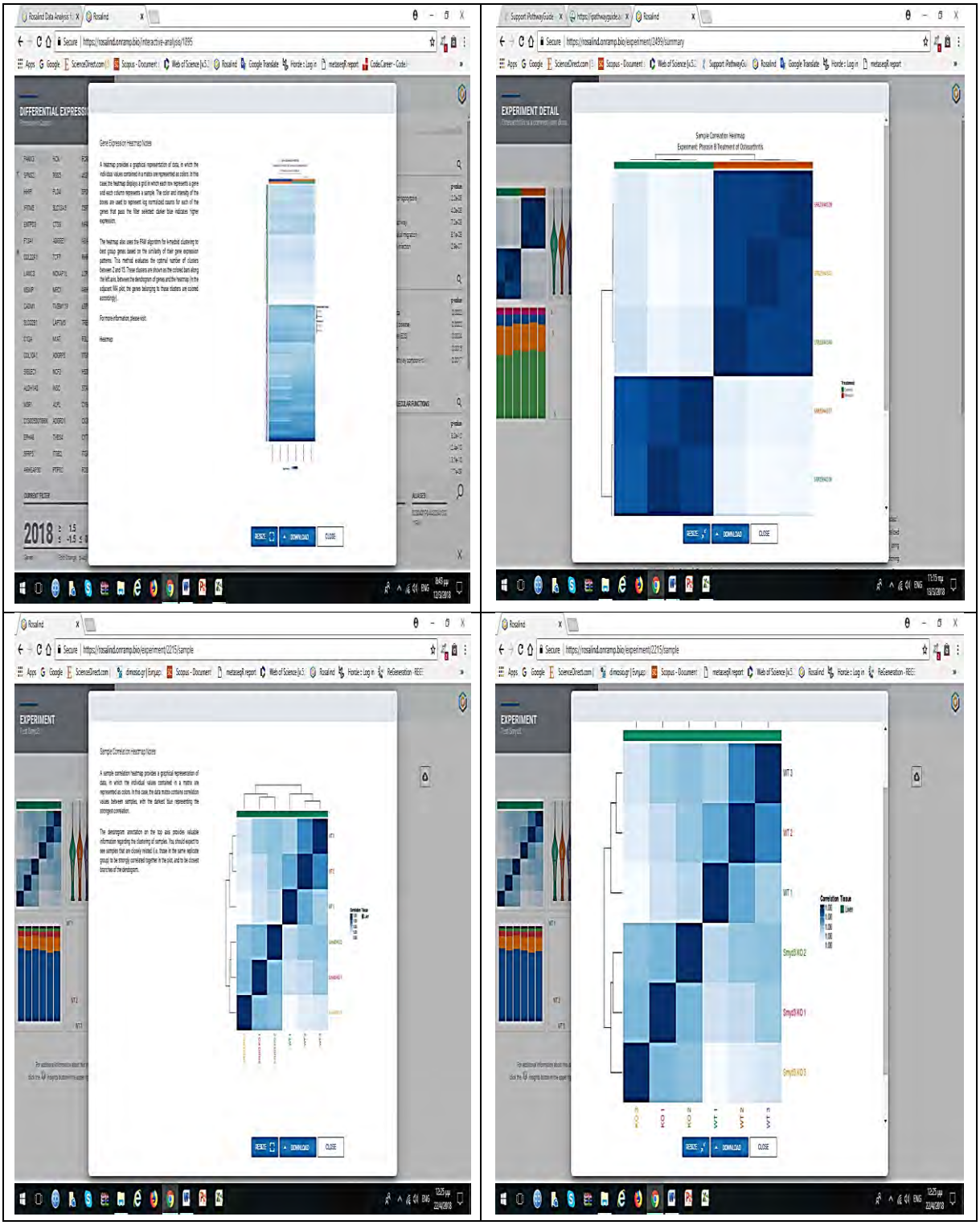
Η δημιουργία αναφορών επιλεγμένων από το χρήστη γονιδίων παρέχεται μόνο μέσω του λογισμικού iPathwayGuide της Advaita σύμφωνα με ενημέρωση των επιστημονικών συνεργατών του Rosalind.

Τρίτη κατηγορία χαρακτηριστικών

- Heat maps

Η κατασκευή Heat maps προσφέρεται από το λογισμικό και αυτό επαληθεύεται τόσο από την πληροφόρηση του υπεύθυνου του Rosalind όσο και από προσωπική εμπειρία. Η Εικόνα 2.6.3 δείχνει κάποια στιγμιότυπα χαρτών θερμότητας που προέκυψαν ύστερα από την ανάλυση κατά συστάδες που πραγματοποιήθηκε στα δείγματα του εξεταζόμενου οργανισμού.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.6.3: Heat maps διαγράμματα των 6 δειγμάτων του mus musculus

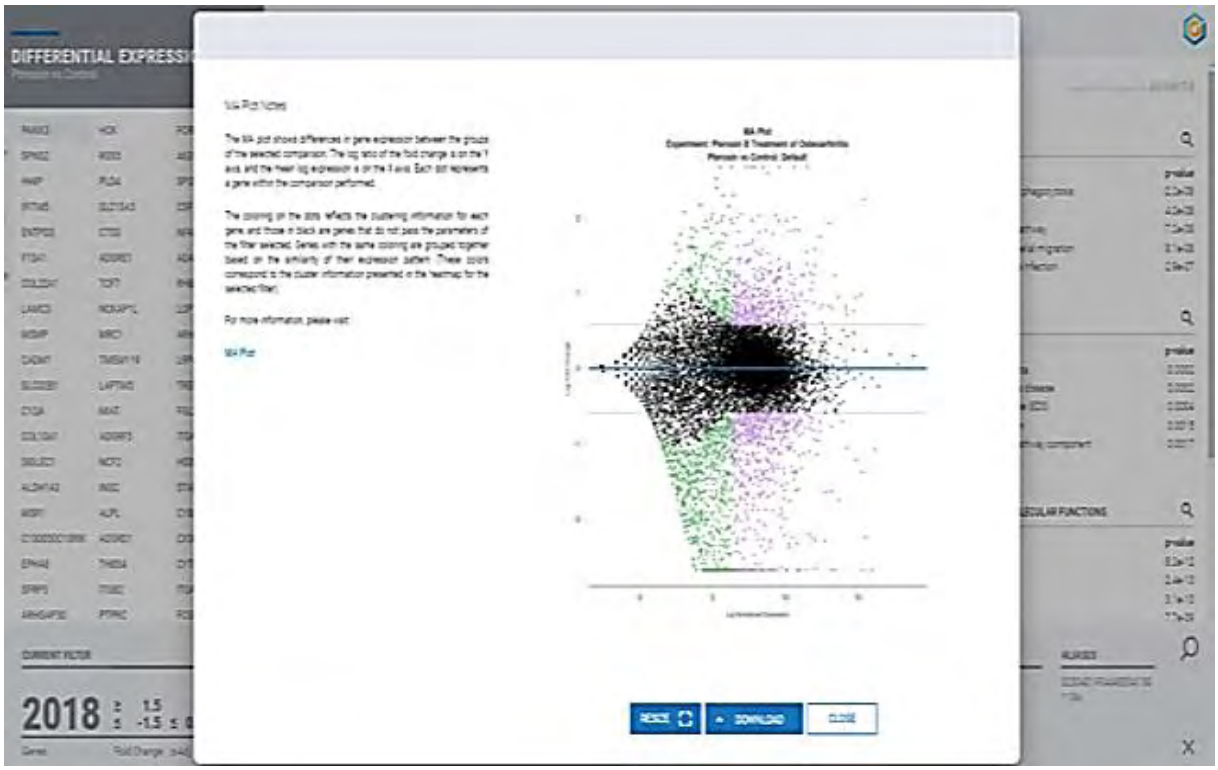
- Volcano plots

Δεν υποστηρίζεται από την παρούσα έκδοση του λογισμικού η δημιουργία Volcano διαγραμμάτων σύμφωνα και με τα όσα ισχυρίζεται ο εκπρόσωπος της εταιρείας

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- MA plots

Το λογισμικό σε αυτή του την έκδοση παρέχει τη δυνατότητα δημιουργίας MA διαγραμμάτων όπως φαίνεται και στην Εικόνα 2.6.4 η οποία δείχνει ένα MA διάγραμμα που αναφέρεται σε ένα παράδειγμα το οποίο αναλύεται από το συγκεκριμένο λογισμικό. Δυστυχώς, δεν κατέστη δυνατή η δημιουργία τέτοιων διαγραμμάτων για τον εξεταζόμενο οργανισμό λόγω πολυπλοκότητας της όλης διαδικασίας για έναν μη έμπειρο υπολογιστικά χρήστη.



Εικόνα 2.6.4: Ενδεικτικά στιγμιότυπα του λογισμικού από τη δημιουργία MA διαγραμμάτων

- Genome Browser

Η δυνατότητα περιήγησης γονιδιωμάτων παρέχεται μόνο στην αλληλούχιση ChIP-seq αλλά προβλέπεται ότι θα ενσωματωθεί μελλοντικά και στην RNA-seq αλληλούχιση όπως υποστηρίζει και ο υπεύθυνος της εταιρείας.

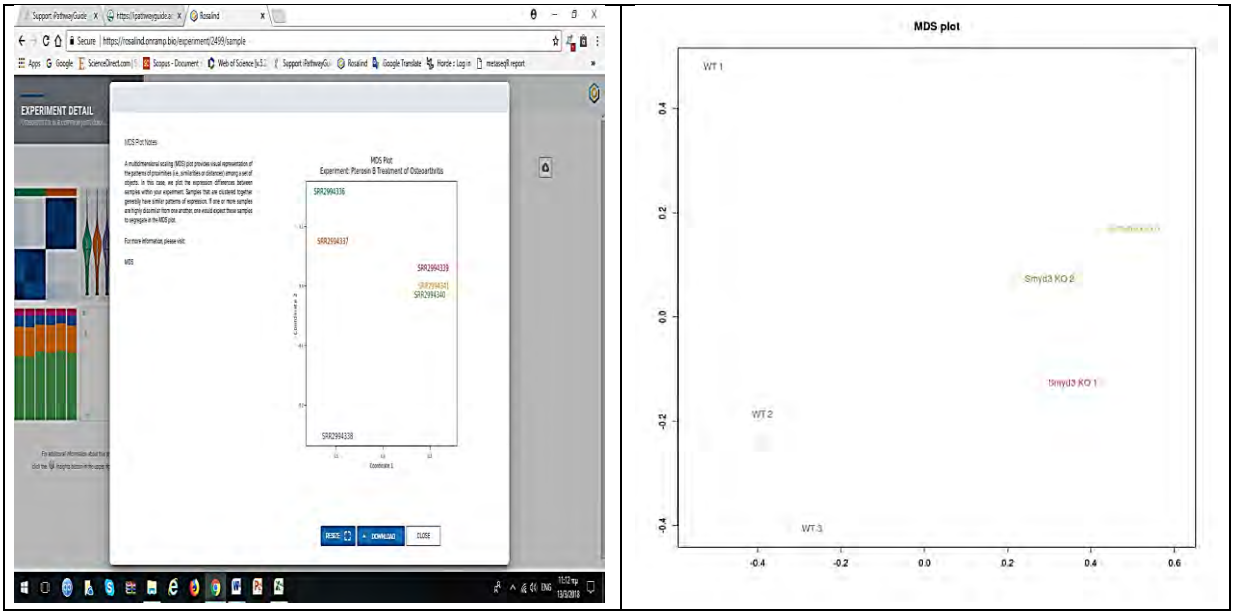
- PCA plots

Δεν περιλαμβάνονται γραφήματα ανάλυσης κύριων συνιστωσών στην παρούσα έκδοση του λογισμικού σύμφωνα και με την ενημέρωση που έγινε μέσω email από τον επιστημονικό υπεύθυνο του λογισμικού Rosalind.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- MDS plots

Το λογισμικό αυτό υποστηρίζει την ανάλυση κατά συστάδες με χρήση MDS γραφημάτων όπως αυτά που παρουσιάζονται στην Εικόνα 2.6.5 και τα οποία αφορούν στην αντιμετώπιση (treatment) της οστεοαρθρίτιδας στην αριστερή στήλη και στην αφαίρεση με τεχνητό τρόπο του γονιδίου Smyd3 από το γονιδίωμα του ποντικού mus musculus στη δεξιά στήλη. Αυτό που φαίνεται από τη δεξιά στήλη είναι ότι τα WT δείγματα βρίσκονται σε μεγαλύτερη απόσταση μεταξύ τους σε σχέση με την απόσταση που χωρίζει τα Smyd3KO δείγματα.



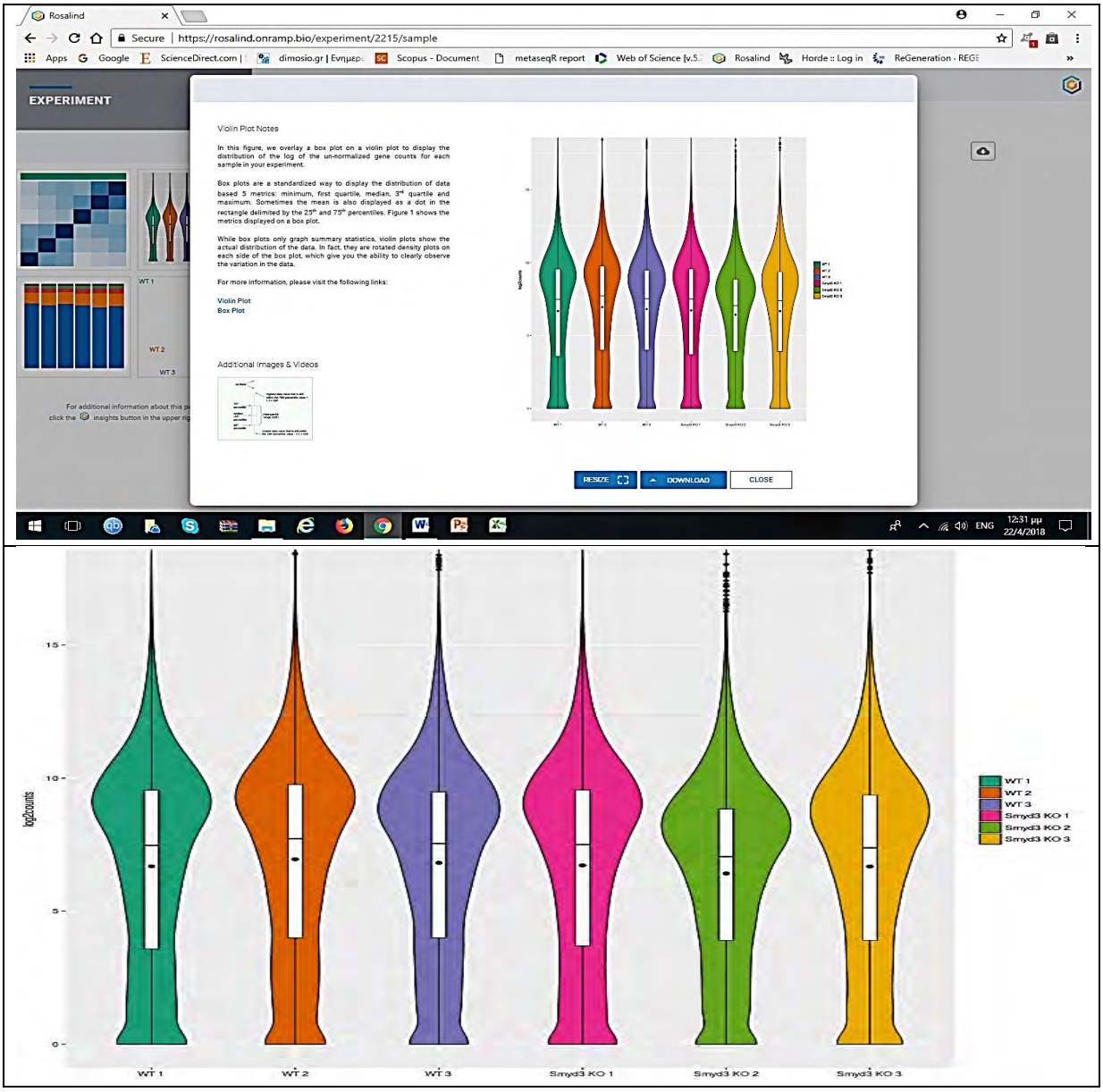
Εικόνα 2.6.5.: MDS γραφήματα του λογισμικού Rosalind

- Box plots

Η δυνατότητα κατασκευής ενός συνδυασμού Boxplot και Violin plot διαγραμμάτων προσφέρεται από την παρούσα έκδοση του λογισμικού αυτού όπως δείχνει και η Εικόνα 2.6.6. Πιο συγκεκριμένα, σύμφωνα με τις on-line σημειώσεις/οδηγίες που προσφέρει το λογισμικό Rosalind πραγματοποιείται ενσωμάτωση των box plot διαγραμμάτων στα Violin plot διαγράμματα.

Η ειδοποιός διαφορά μεταξύ των 2 τύπων των διαγραμμάτων εντοπίζεται στο ότι τα διαγράμματα τύπου Violin παρέχουν τη δυνατότητα παρατήρησης της διακύμανσης των δεδομένων με μεγαλύτερη σαφήνεια ενώ τα διαγράμματα τύπου Box plot προσφέρουν μόνο συνοπτικά στατιστικά στοιχεία.

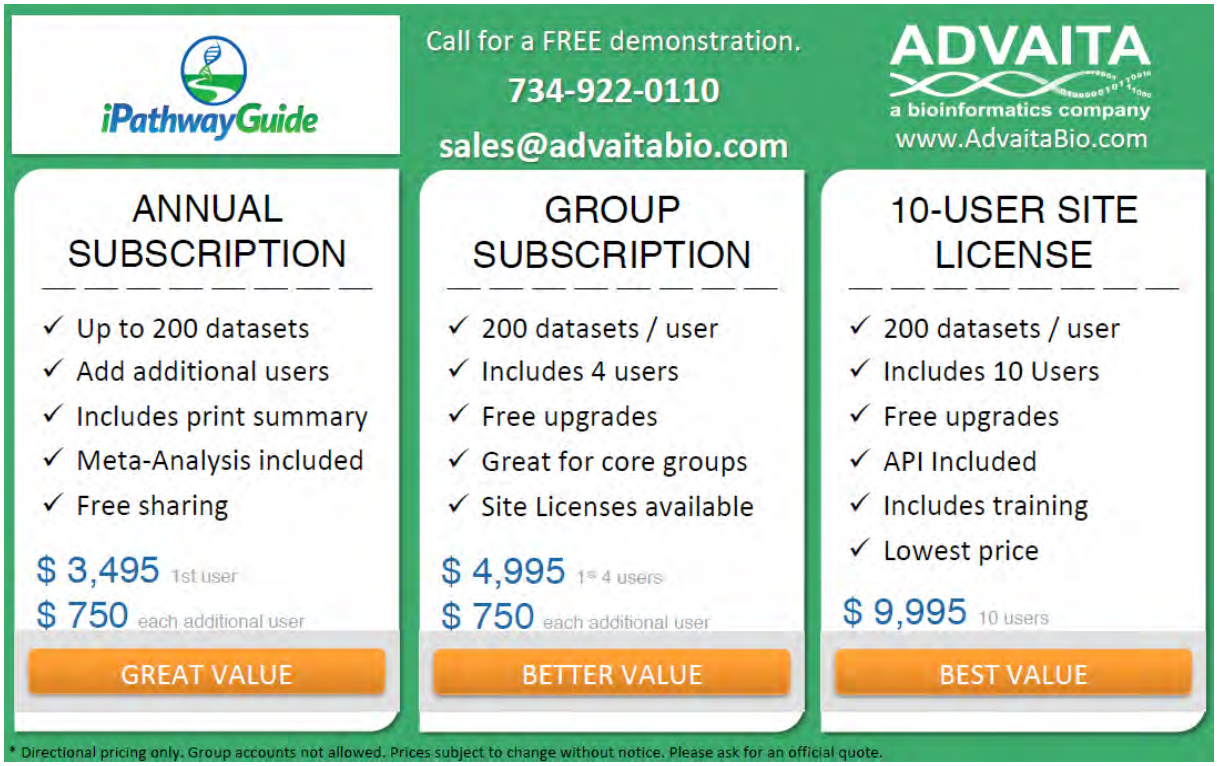
Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.6.6: Boxplot και Violin plot διαγράμματα των δειγμάτων του εξεταζόμενου οργανισμού

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Δεν υπάρχει κάποιο συγκεκριμένο κόστος άδειας χρήσης του λογισμικού σύμφωνα με την ενημέρωση του υπεύθυνου της εταιρείας ύστερα από επικοινωνία που πραγματοποιήθηκε μέσω email. Η χρέωση γίνεται ανά δείγμα και αυτήν την περίοδο το κόστος ανέρχεται στα 40$ ανά δείγμα. Το κόστος περιλαμβάνει την μετάβαση από την εισαγωγή ενός FASTQ αρχείου σε διαφορικά εκφρασμένα γονίδια με διαδραστική πρόσβαση στην ανάλυση βιοχημικών μονοπατιών, στην ανάλυση οντολογιών έκφρασης γονιδίων και στην ανάλυση πιθανών ασθενειών που σχετίζονται με συγκεκριμένα γονίδια. Ακόμη, ένα σημαντικό σημείο είναι ότι η πληρωμή πραγματοποιείται αφότου ελέγξει ο χρήστης τα αποτελέσματα του ποιοτικού ελέγχου. Τέλος, το κόστος μηνιαίας άδειας για το λογισμικό iPathwayGuide της εταιρείας Advaita, η οποία αποτελεί συνεργάτη της OnRamp, για έναν ή περισσότερους χρήστες μπορεί εύκολα να υπολογισθεί από την Εικόνα 2.6.7. Επομένως, ύστερα από υπολογισμούς προκύπτει ότι το μηνιαίο κόστος ατομικής άδειας χρήσης του iPathwayGuide ανέρχεται στα 291$ περίπου.



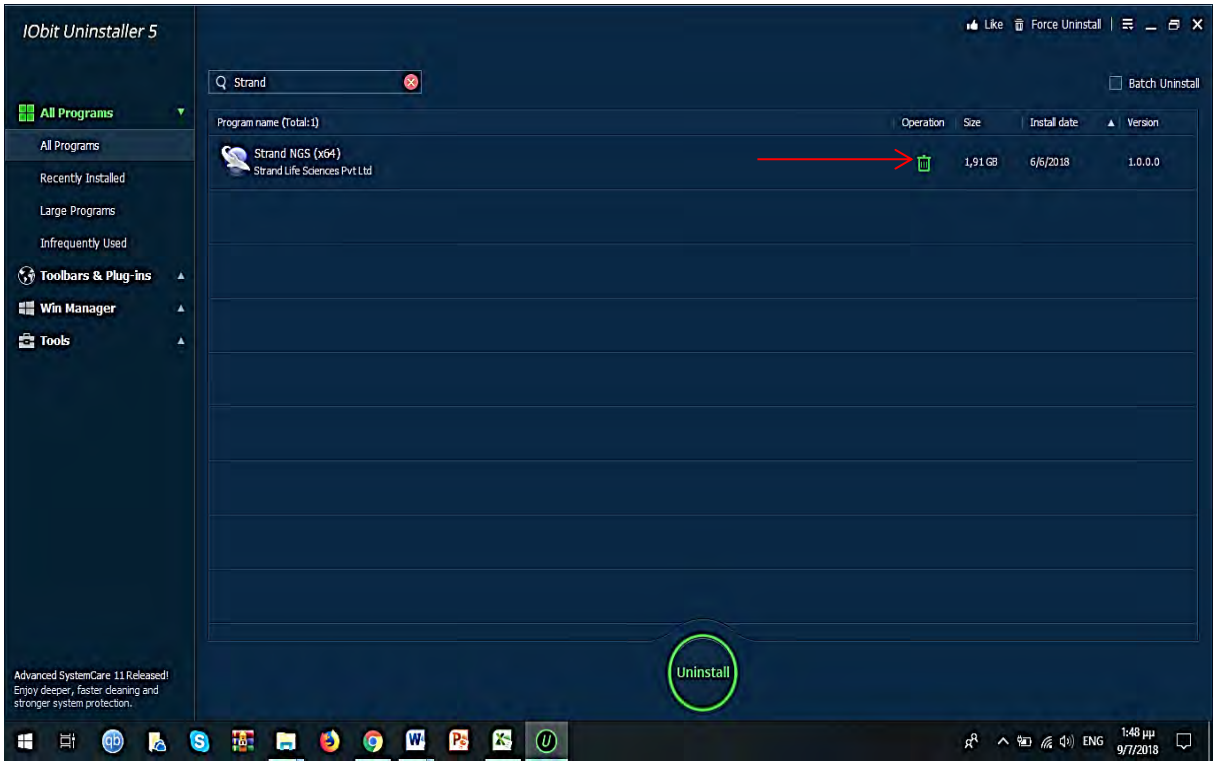
Εικόνα 2.6.7: Ετήσιο κόστος άδειας χρήσης του λογισμικού της Advaita

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

2.7 Strand NGS (Strand Scientific Intelligence)

Πρώτη κατηγορία χαρακτηριστικών

Στην Εικόνα 2.7.1 φαίνεται ότι ο χώρος που καταλαμβάνει το λογισμικό Strand NGS της εταιρείας Strand Scientific Intelligence είναι 1,91 GB περίπου. Η αρχιτεκτονική συστήματος είναι Desktop ενώ όσον αφορά στο λειτουργικό σύστημα το λογισμικό υποστηρίζει και τους 3 τύπους των συστημάτων τύποι σύμφωνα με την ιστοσελίδα της αμερικάνικης εταιρείας.



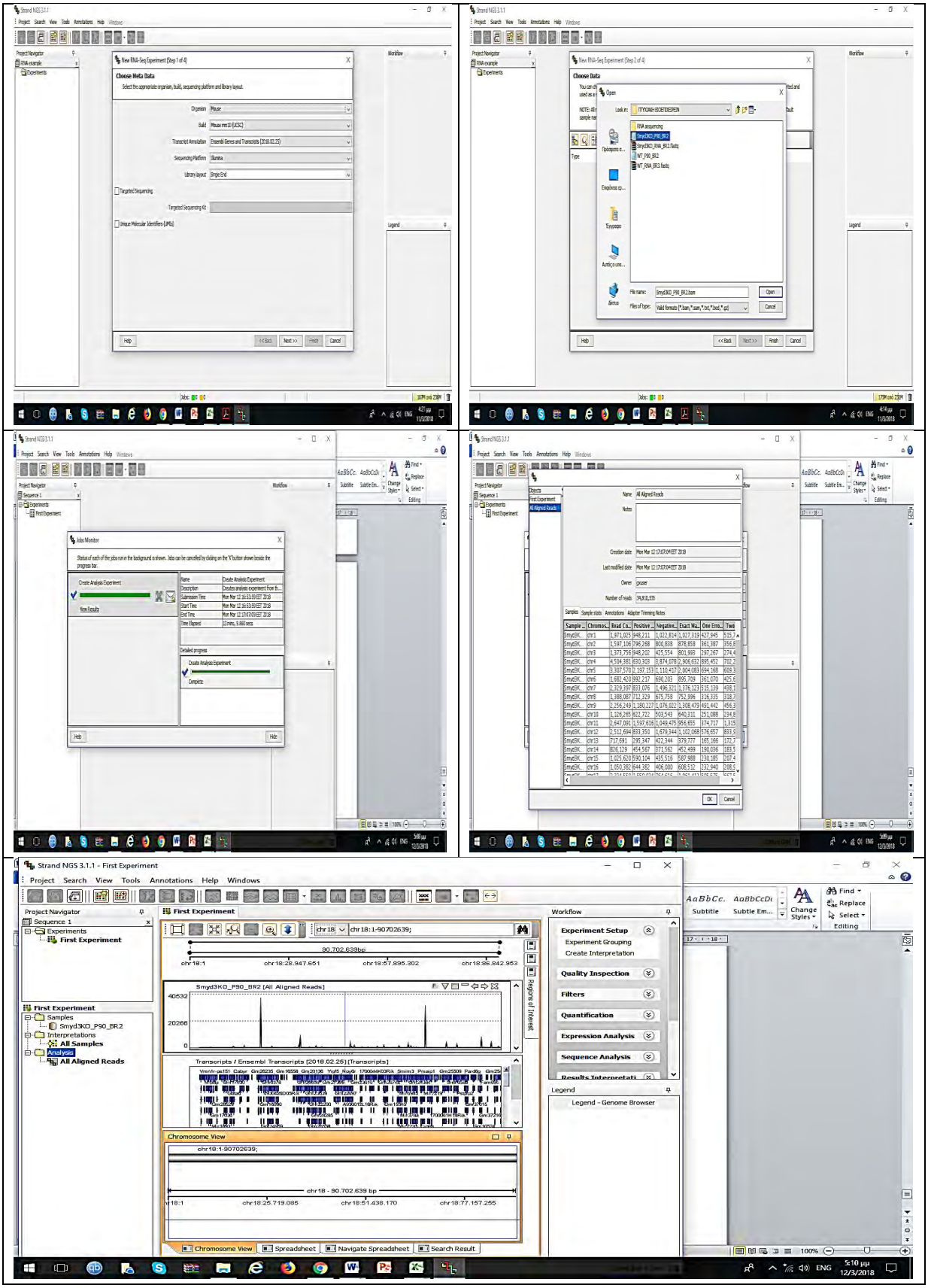
Εικόνα 2.7.1: Χωρητικότητα του λογισμικού Strand NGS

Δεύτερη κατηγορία χαρακτηριστικών

- FASTQ/BAM αρχεία

Σύμφωνα με τους υπευθύνους της εταιρείας, το λογισμικό Strand NGS υποστηρίζει αρχεία δεδομένων μορφής FASTQ, FASTA, GZIP, TXT καθώς και unaligned SAM και BAM αρχεία. Η Εικόνα 2.7.2 δείχνει την διαδικασία εισαγωγής ενός BAM αρχείου που αφορά τον εξεταζόμενο οργανισμό.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.7.2: Διαδικασία εισαγωγής ενός BAM αρχείου

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Organisms

Σύμφωνα με τους ισχυρισμούς των αρμοδίων, το λογισμικό παρέχει τη δυνατότητα για ανάλυση του γονιδιώματος οποιουδήποτε οργανισμού. Ύστερα από εξερεύνηση του manual του λογισμικού διαπιστώθηκε ότι είναι δυνατή η επεξεργασία των γονιδιωμάτων περισσότερων από 40 οργανισμών. Επίσης, διαθέτει σχολιασμούς (annotations) μεταγραφής των γονιδίων πρότυπων εργαστηριακών οργανισμών-μοντέλων.

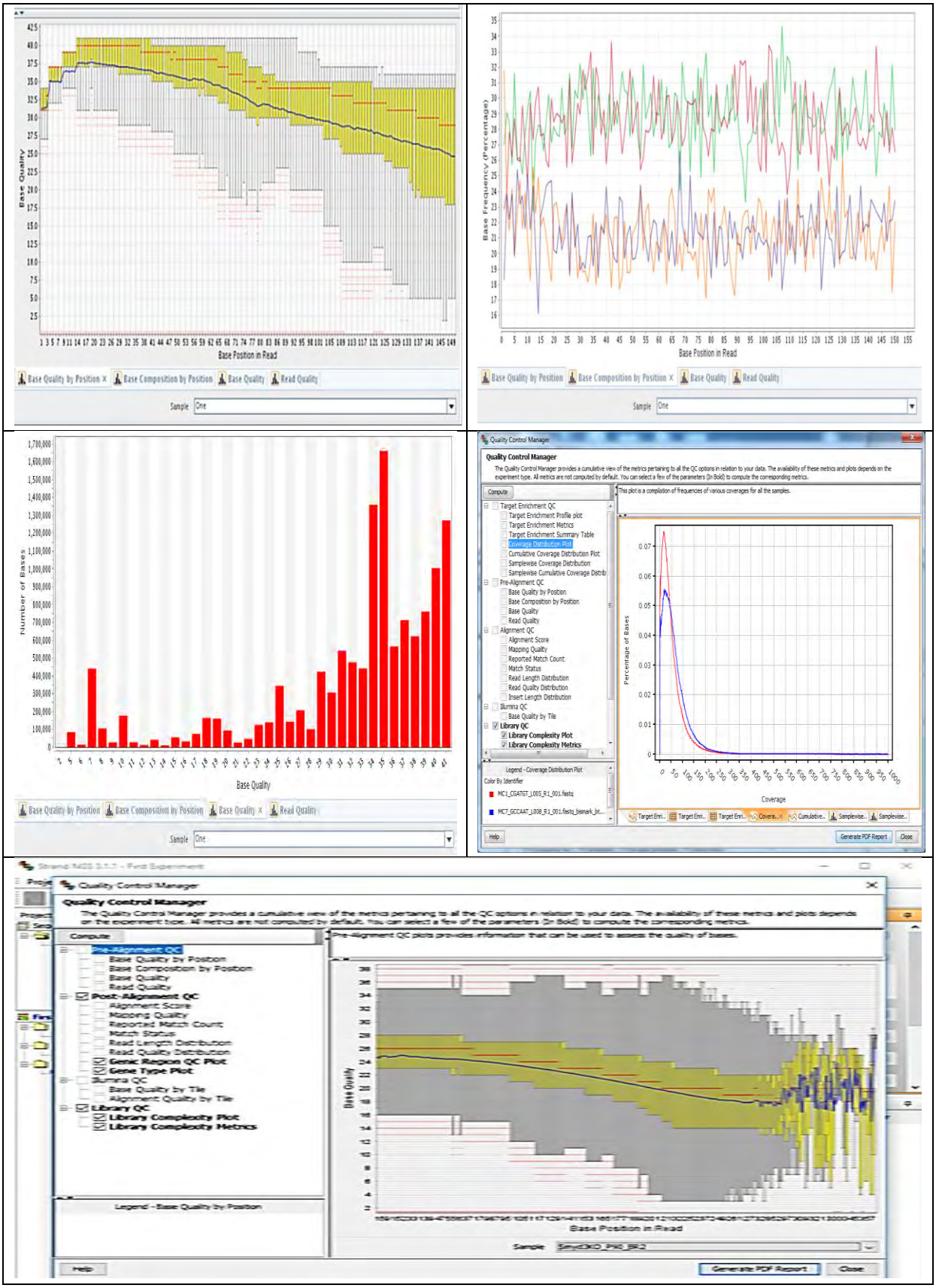
- Quality Control

Το συγκεκριμένο λογισμικό προσφέρει εκτεταμένες επιλογές ποιοτικού ελέγχου των δειγμάτων στην αρχή και στο τέλος της διαδικασίας της ευθυγράμμισης (Pre και Post Alignment QC). Ειδικότερα, μεταξύ των άλλων επιλογών του ποιοτικού ελέγχου προσφέρονται κατανομές Length, Quality και Coverage Distribution. Η Εικόνα 2.7.3 παρουσιάζει κάποια στιγμιότυπα από τον ποιοτικό έλεγχο που αφορά στην ποιότητα των βάσεων ανά θέση αλληλούχισης (Base Quality by Position), τη σύνθεση των βάσεων ανά θέση (Base Composition by Position), την κάλυψη κάθε βάσης (Base Coverage) καθώς και την κατανομή της ποιότητας των βάσεων (Base Quality) με τη μορφή ιστογράμματος.

Ειδικότερα, το πρώτο διάγραμμα της εικόνας αναφέρεται στην Base Quality by Position, που προέρχεται από το manual του λογισμικού, και είναι παρόμοιο με την κατανομή της ποιότητας των αλληλουχηθέντων ολιγονουκλεοτιδίων όπως αυτή ορίστηκε στο κύριο μέρος της εργασίας. Το δεύτερο διάγραμμα της εικόνας προέρχεται και αυτό από το εγχειρίδιο του λογισμικού και φέρει την ονομασία Base Composition by Position. Ουσιαστικά, είναι ανάλογο με το χαρακτηριστικό του ποιοτικού ελέγχου που ονομάζεται περιεκτικότητα των νουκλεοτιδικών βάσεων σε GC-content.

Το τρίτο διάγραμμα αποτελεί παράδειγμα του manual του λογισμικού Strand NGS και είναι ένα ιστόγραμμα της κατανομής της ποιότητας των βάσεων στο οποίο ο κάθετος άξονας αποτελείται από τον αριθμό των βάσεων και ο οριζόντιος από το PHRED score όπως αυτό ορίστηκε στη σελίδα 14 αυτού του κεφαλαίου. Το τέταρτο διάγραμμα αναφέρεται στην κάλυψη κάθε βάσης και προέρχεται από ένα παράδειγμα του εγχειρίδιου χρήσης του λογισμικού. Το πέμπτο και τελευταίο διάγραμμα είναι αποτέλεσμα του ποιοτικού ελέγχου ενός δείγματος Smyd3KO του εξεταζόμενου οργανισμού. Από το στιγμιότυπο του mus musculus φαίνεται ότι το PHRED score των 300 βάσεων κυμαίνεται από 24-28 και αφού επομένως είναι μικρότερο από 40 συμπεραίνεται ότι δεν πρόκειται για μια ανάλυση καλής ποιότητας. Παρ' όλα αυτά γίνεται αποδεκτή επειδή προσεγγίζει το όριο του 30 ενώ μετά τις 300 βάσεις γίνεται αντιληπτό ότι η κατανομή ακολουθεί διαφορετική τροχιά λόγω εμφάνισης θορύβου όπως αποκαλείται φασματοσκοπικά.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.7.3: Ενδεικτικά στιγμιότυπα ποιοτικού ελέγχου του λογισμικού

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Differential Expression

Η ανάλυση διαφορικής έκφρασης γονιδίων παρέχεται ως δυνατότητα από το συγκεκριμένο λογισμικό σε πειράματα που εφαρμόζονται RNA και Small RNA-Seq αλληλουχίσεις όπως επισημαίνουν οι ειδικοί της Strand NGS χωρίς όμως να καταστεί δυνατή η εκτέλεση αυτής της ανάλυσης και η παρουσίαση των αποτελεσμάτων με εικόνες εξαιτίας της ανεπάρκειας σε μνήμη του υπολογιστή δοκιμής.

- External Annotation Databases

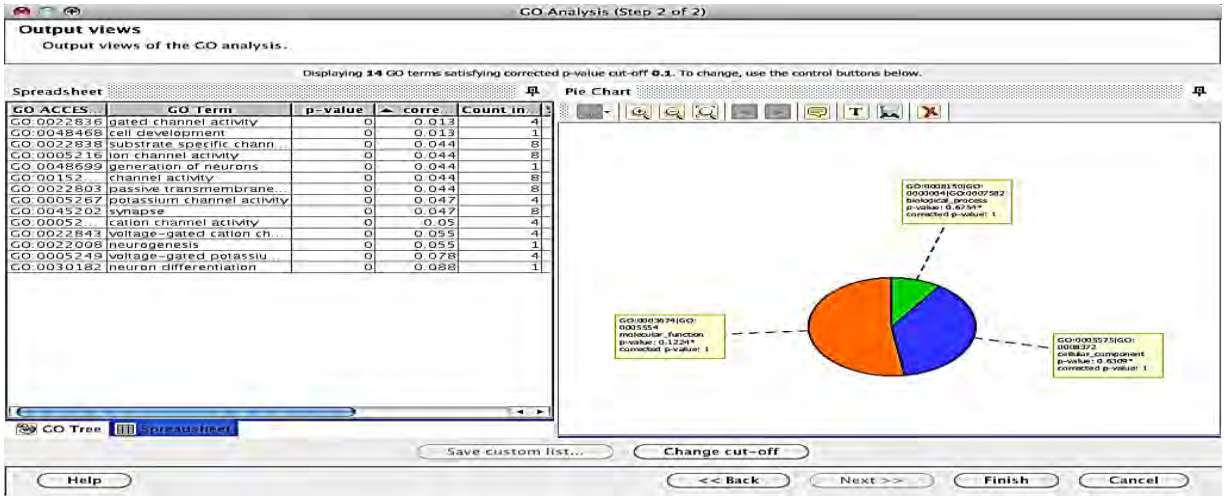
Το λογισμικό της Strand Scientific Intelligence δεν διαθέτει κάποιο υπολογιστικό εργαλείο που να συνδέεται με εξωτερικές βάσεις σχολιασμών των γονιδιωμάτων των οργανισμών. Όμως, οι σχολιασμοί πακετάρονται από τις βάσεις δεδομένων Ensembl, RefSeq και UCSC και στη συνέχεια είναι διαθέσιμοι για πειράματα ευθυγράμμισης και ανάλυσης.

- GO Analysis

Η ανάλυση εμπλουτισμένης γονιδιακής οντολογίας παρέχεται από το λογισμικό επιτρέποντας με αυτόν τον τρόπο να σχηματιστεί το βιολογικό πλαίσιο των δεδομένων του χρήστη. Η Εικόνα 2.7.4 δείχνει ένα στιγμιότυπο της ανάλυσης αυτής που προέρχεται από το manual του λογισμικού Strand NGS.

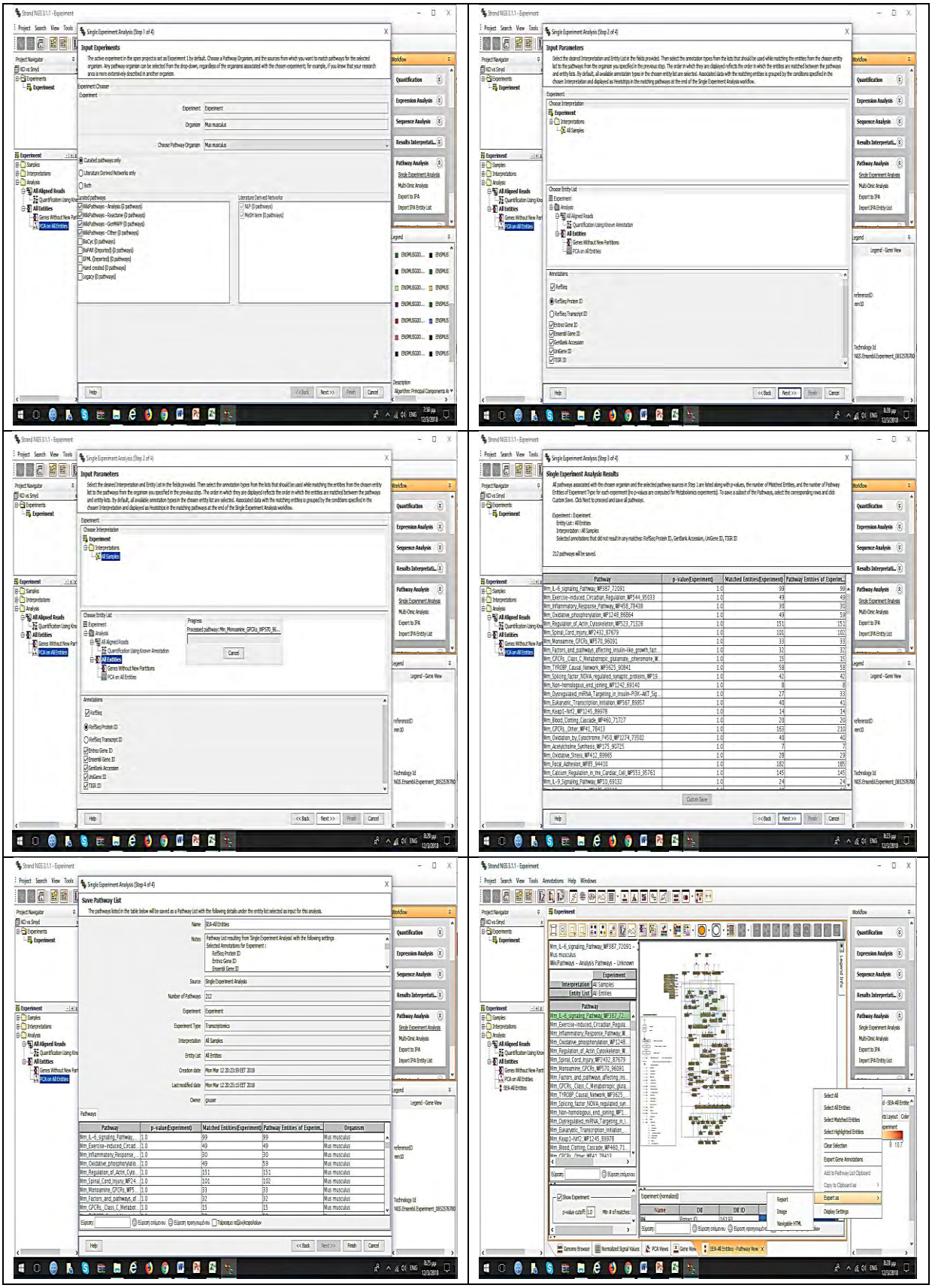
- Biochemical Pathway Analysis

Η δυνατότητα ανάλυσης βιοχημικών μονοπατιών μπορεί να πραγματοποιηθεί με τράπεζες δεδομένων όπως είναι η Wiki-Pathways και η Reactome. Η Εικόνα 2.7.5 δείχνει την ανάλυση βιοχημικών μονοπατιών των δειγμάτων του εξεταζόμενου οργανισμού.



Εικόνα 2.7.4: Στιγμιότυπο GO ανάλυσης από το manual του Strand NGS

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.7.5: Διαδικασία ανάλυσης βιοχημικών μονοπατιών για τα εξεταζόμενα δείγματα

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Workflows Creations

Το λογισμικό αυτό στην παρούσα του έκδοση δεν περιλαμβάνει τη δυνατότητα ροής εργασιών σύμφωνα με ενημέρωση των ανθρώπων της εταιρείας.

- Clustering Analysis

Η ανάλυση κατά συστάδες είναι μια δυνατότητα που υποστηρίζεται, όπως τονίζουν οι αρμόδιοι, από το λογισμικό μέσω του αλγόριθμου Hierarchical Clustering καθώς και μέσω του αλγορίθμου K-means. Ωστόσο, δεν κατέστη δυνατή η συλλογή ενός ή περισσότερων στιγμιότυπων καθώς δεν εντοπίστηκε κάποια εικόνα στο manual του λογισμικού ούτε θεωρήθηκε εύκολη η διαδικασία ανάλυσης για έναν υπολογιστικά αρχάριο χρήστη.

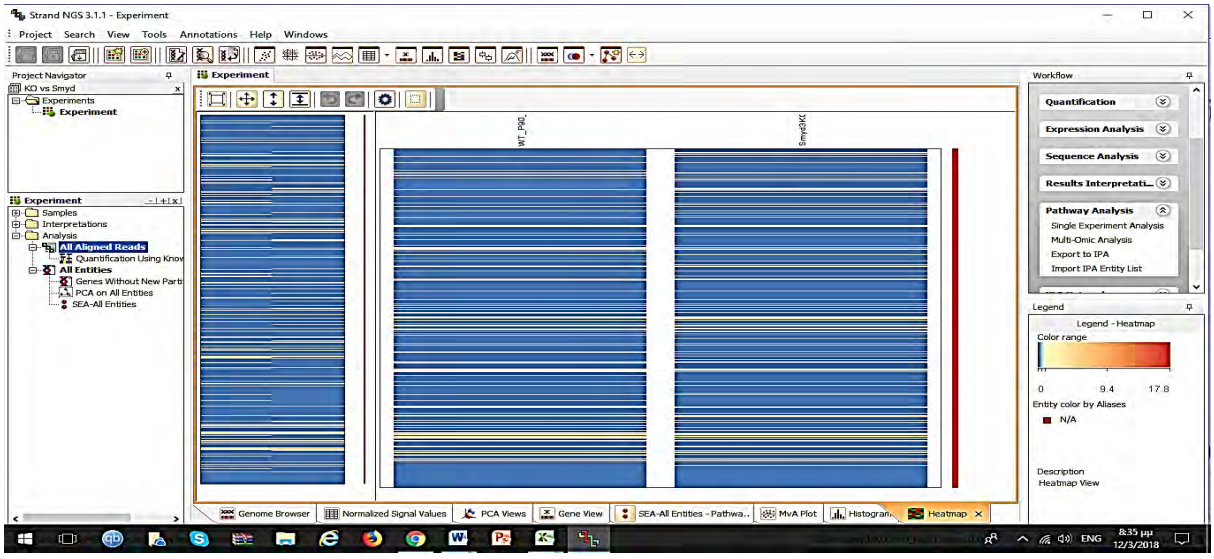
- Custom Reports

Η δυνατότητα δημιουργίας αναφορών προερχόμενων από την επεξεργασία συγκεκριμένων γονιδίων, που αποτελούν αντικείμενο ενδιαφέροντος για το χρήστη, δεν υποστηρίζεται από το λογισμικό όπως τονίζουν οι επιστημονικοί συνεργάτες της Strand Scientific Intelligence.

Τρίτη κατηγορία χαρακτηριστικών

- Heat maps

Η δημιουργία χαρτών θερμότητας υποστηρίζεται από το λογισμικό και αυτό επιβεβαιώνεται και από την Εικόνα 2.7.6.



Εικόνα 2.7.6: Heat maps για 2 από τα εξεταζόμενα δείγματα

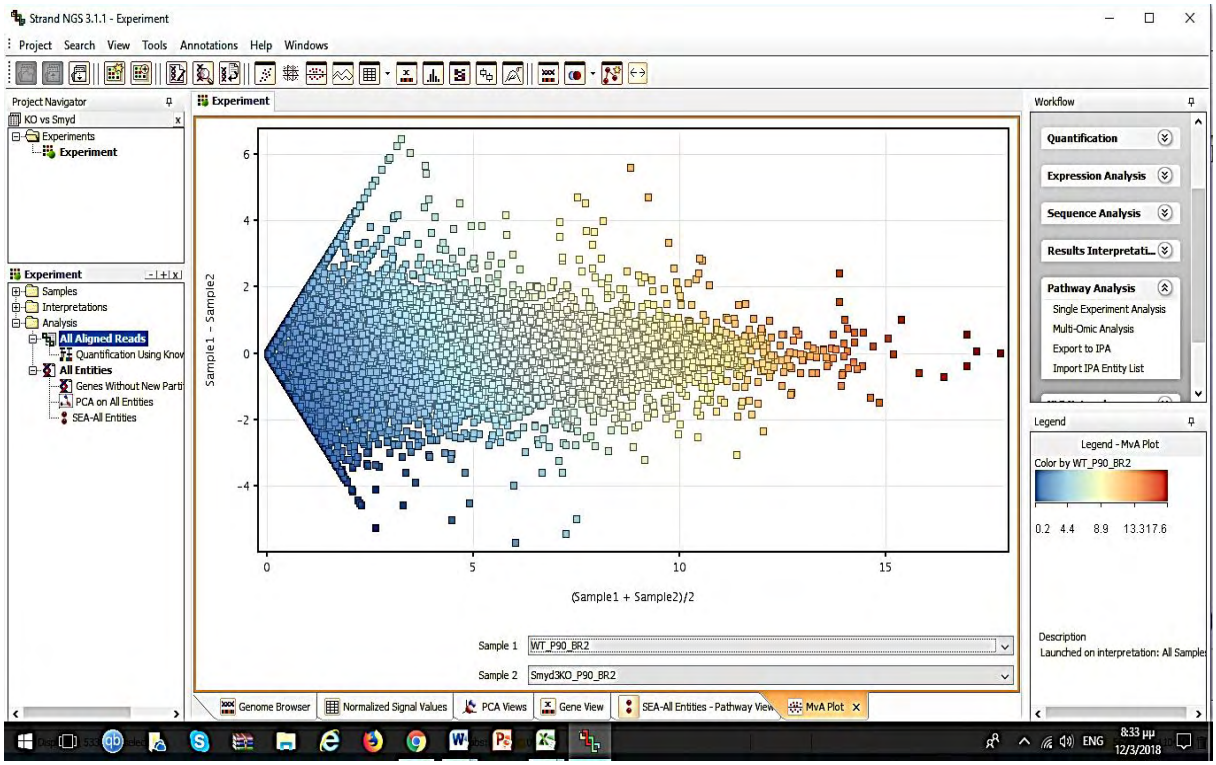
Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- Volcano plots

Οι αρμόδιοι του λογισμικού ισχυρίζονται ότι ο συγκεκριμένος τύπος γραφικών παραστάσεων υποστηρίζεται από το πρόγραμμα μερικώς δηλαδή μόνο κατά την ανάλυση αντιγράφου (replicate analysis). Για αυτό το λόγο, δεν υπάρχει κάποια σχετική εικόνα.

- MA plots

Η κατασκευή MA γραφημάτων υποστηρίζεται πλήρως από το λογισμικό όπως δείχνει και η Εικόνα 2.7.7. Το γράφημα της εικόνας αποτελεί έναν τρόπο οπτικοποίησης της διαφορικής γονιδιακής έκφρασης 2 δειγμάτων (1 WT και 1 Smyd3KO) του εξεταζόμενου οργανισμού.

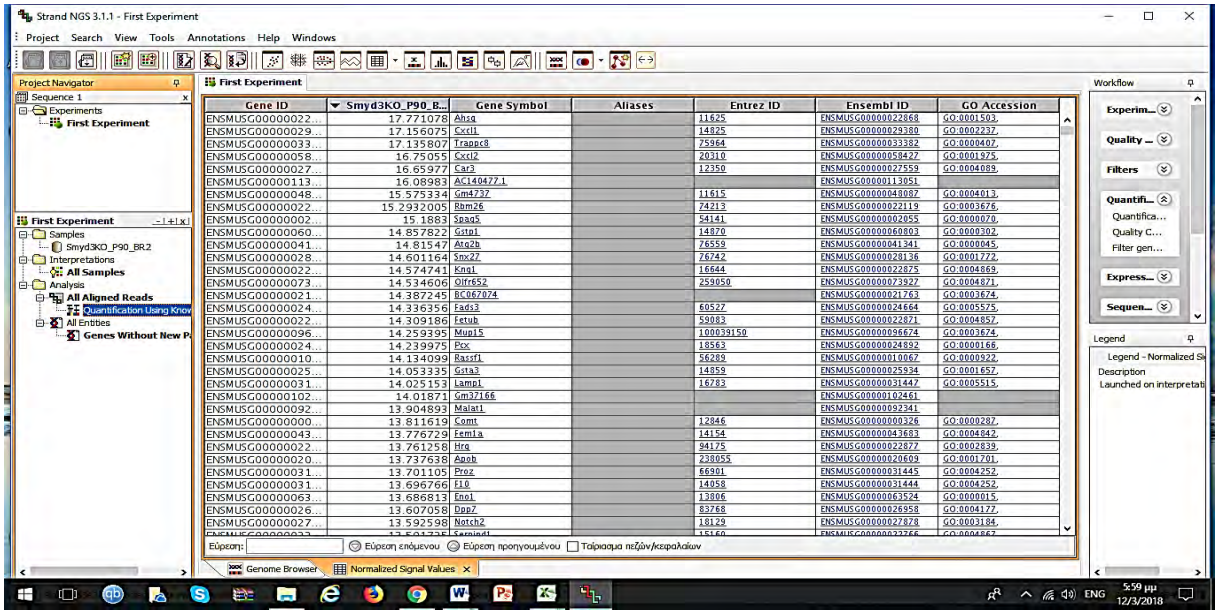


Εικόνα 2.7.7: MA γράφημα για 2 δείγματα του εξεταζόμενου οργανισμού

- Genome Browser

Το λογισμικό αυτό διαθέτει έναν πολύ ισχυρό περιηγητή γονιδιώματος ο οποίος μπορεί να χρησιμοποιηθεί και για την απεικόνιση της μεταβολής της εσωτερικής και της αναμεταξύ των χρωμοσωμάτων δομής. Η Εικόνα 2.7.8 παρουσιάζει ένα στιγμιότυπο του περιηγητή γονιδιώματος για ένα Smyd3KO δείγμα του εξεταζόμενου οργανισμού.

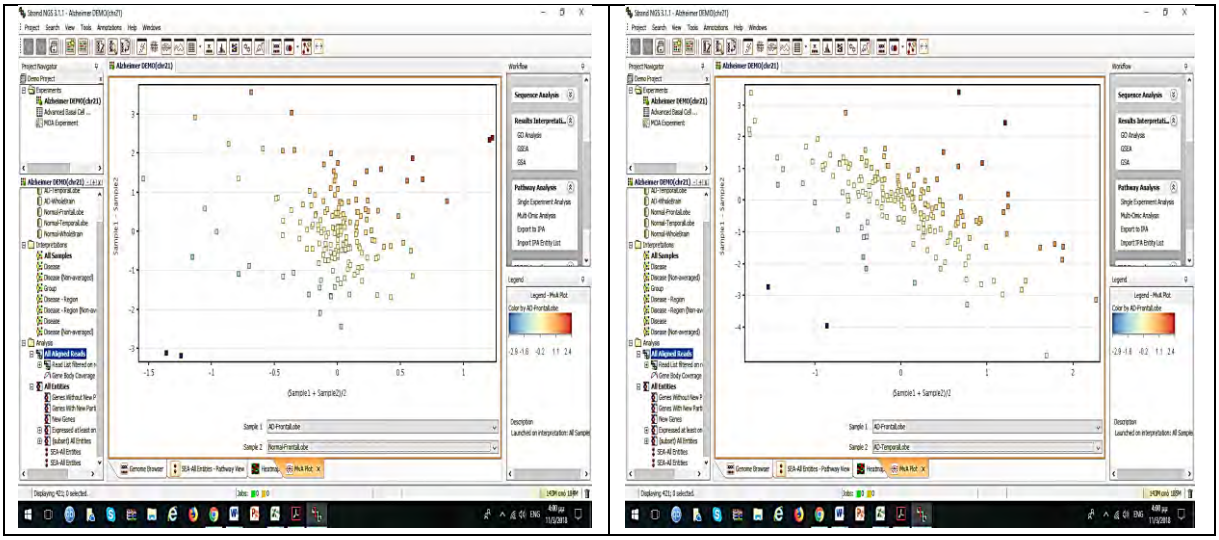
Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154



Εικόνα 2.7.8: Περιηγητής γονιδιώματος για ένα δείγμα του mus musculus

- PCA plots

Η ανάλυση κύριων συνιστωσών παρέχεται από το λογισμικό όπως φαίνεται και στην Εικόνα 2.7.9. Η εικόνα αυτή προέρχεται από το manual του λογισμικού και παρουσιάζει στην αριστερή στήλη το γράφημα του μετωπιαίου λοβού (Δείγμα 1) ο οποίος έχει προσβληθεί από τη νόσο Αλτσχάιμερ σε σχέση με το φυσιολογικό μετωπιαίο λοβό (Δείγμα 2) και στη δεξιά στήλη το γράφημα του προσβεβλημένου λοβού (Δείγμα 1) σε σύγκριση με τον κροταφικό λοβό (Δείγμα 2).



Εικόνα 2.7.9: PCA γραφήματα από το manual του λογισμικού

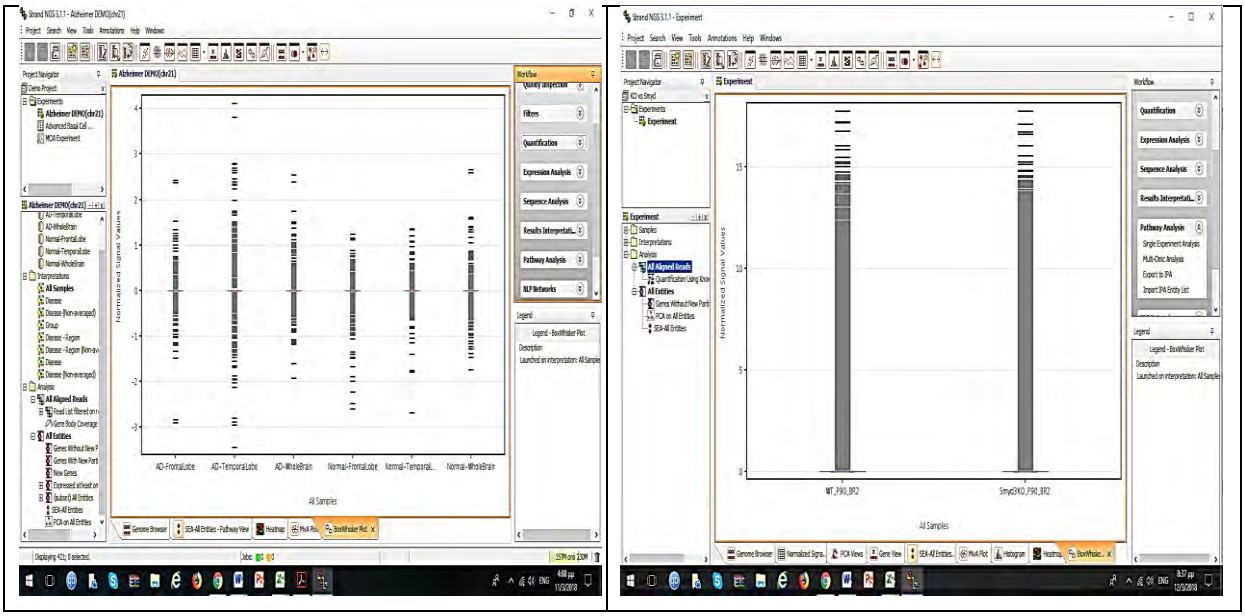
Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

- MDS plots

Γραφήματα πολυδιάστατης κλιμάκωσης δεν διατίθενται από την παρούσα έκδοση του λογισμικού της εταιρείας Strand Scientific Intelligence.

- Box plots

Η Εικόνα 2.7.10 παραθέτει στην αριστερή στήλη διαγράμματα τύπου Box plot που προέρχονται από το manual του λογισμικού. Στη δεξιά στήλη της ίδιας εικόνας φαίνονται τα Box plot διαγράμματα 2 δειγμάτων (1 WT και 1 Smyd3KO) του εξεταζόμενου οργανισμού. Τα κανονικοποιημένα Box plot διαγράμματα των 2 δειγμάτων του mus musculus παρουσιάζουν περισσότερες ομοιότητες παρά διαφορές.



Εικόνα 2.7.10: Box plot διαγράμματα του Strand NGS

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Το μηνιαίο κόστος απόκτησης άδειας χρήσης για έναν ακαδημαϊκό χρήστη, κατόπιν επικοινωνίας μέσω email με την εταιρεία Strand Scientific Intelligence, ανέρχεται στα 375 $. Το ποσό αυτό διαμορφώνεται από υπολογισμούς που βασίζονται στο ετήσιο κόστος άδειας χρήσης το οποίο φαίνεται στην Εικόνα 2.7.11.

Dear Mr. Margaritis,

This is Prathima from Strand NGS team who spoke to you today. Thank you for taking my call, It was nice talking to you.

Understand from our discussion that currently you require approx. pricing only. Would like to share that pricing of Strand NGS 1 year Individual node locked/desktop academic licenses USD 4500 (for qty 1).

As suggested, we will connect back to you by 2nd March to understand your purchase plans. In the meantime, please feel free to contact us for any queries connected to Strand NGS software, we will be glad to assist you. Thanks

Best regards,
Prathima
The Strand NGS team

Εικόνα 2.7.11: Το ετήσιο κόστος άδειας χρήσης για το λογισμικό Strand NGS

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

3. ΣΥΜΠΕΡΑΣΜΑΤΑ-ΣΥΖΗΤΗΣΗ

Στους 4 συγκεντρωτικούς πίνακες που ακολουθούν συνοψίζονται τα χαρακτηριστικά καθώς και το κόστος μηνιαίας άδειας ενός χρήστη για όλες τις πλατφόρμες αλληλούχισης δεύτερης γενιάς προκειμένου να καταστεί εφικτή και όσο το δυνατόν αποτελεσματικότερη η σύγκριση τους.

Πίνακας 3.1: Χαρακτηριστικά Συστήματος

Λογισμικό	Εταιρεία	Required Space	System Architecture	Operating System (Windows/Mac/Linux)
1. CLC Genomics Workbench	CLC bio, Qiagen	464 MB	Desktop	✓/ ✓/✓
2. Full Lasergene Suite	DNASTAR	1,67 GB	Desktop και Cloud	✓/✓/✗
3. Geneious	Biomatters	468 MB	Desktop	✓/✓/✓
4. Nexus Expression	Biodiscovery	972 MB	Desktop	✓/✓/✓
5. Partek Flow	Partek	467 MB	Desktop	✓/✓/✓
6. Rosalind	OnRamp Bioinformatics	0 MB	Cloud	✓/✓/✓
7. Strand NGS	Strand Scientific Intelligence	1,91 GB	Desktop	✓/✓/✓

Από τον Πίνακα 3.1 μπορούν να προκύψουν τα ακόλουθα συμπεράσματα:
Πρώτον, όσον αφορά στο χώρο που καταλαμβάνει το κάθε λογισμικό στο δίσκο του υπολογιστή, φαίνεται ότι το Strand NGS και το Full Lasergene Suite είναι τα 2 λογισμικά των οποίων το μέγεθος ξεπερνά το 1 GB ενώ κοντά σε αυτό το όριο βρίσκεται και το λογισμικό Nexus Expression. Το λογισμικό Rosalind καταλαμβάνει σχεδόν μηδενικό χώρο στον υπολογιστή και αυτό το καθιστά περισσότερο ανταγωνιστικό σε σχέση με τα υπόλοιπα λογισμικά, ωστόσο, λόγω της cloud υποδομής ενδέχεται να είναι εξαιρετικά επίπονη η μεταφορά πολλών αρχείων από τον υπολογιστή του εκάστοτε χρήστη.

Όσον αφορά στην αρχιτεκτονική του συστήματος, το κύριο συμπέρασμα που μπορεί να εξάγει κάποιος κοιτάζοντας τον παραπάνω πίνακα είναι ότι 2 από τα 7 λογισμικά είναι σχεδιασμένα για να λειτουργούν σε οποιονδήποτε υπολογιστή και αυτό φαίνεται από το γεγονός ότι η αρχιτεκτονική τους είναι Cloud και όχι Desktop πράγμα που σημαίνει ότι η άδεια χρήσης του αντίστοιχου λογισμικού ισχύει για την εκτέλεση του σε έναν και μόνο υπολογιστή.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Τέλος, αναφορικά με το λειτουργικό σύστημα, όλα τα λογισμικά υποστηρίζουν και τους 3 τύπους λειτουργικών συστημάτων (Windows, MAC και Linux) εκτός από το Full Lasergene Suite το οποίο δεν υποστηρίζει το Linux.

Επομένως, συνοψίζοντας τα συμπεράσματα που προέκυψαν από τον πρώτο πίνακα διαπιστώνεται ότι δεν υπάρχει κάποιο λογισμικό που να είναι συγκριτικά ανώτερο ως προς αυτήν την ομάδα χαρακτηριστικών. Αντιθέτως, το λογισμικό της εταιρείας DNASTAR δείχνει να υστερεί έναντι των υπολοίπων αφού δεν υποστηρίζει και τα 3 είδη των λειτουργικών συστημάτων.

Πίνακας 3.2: Χαρακτηριστικά ανάλυσης δεδομένων (I)

Λογισμικό	Εταιρεία	FASTQ/ BAM	Organisms	Quality Control	Differential Expression	External Annotation Databases
1. CLC Genomics Workbench	CLC bio, Qiagen	✓/✓	Mammals, Plants, Animals	✓	✓	✓ (Ensembl, RefSeq)
2. Full Lasergene Suite	DNASTAR	✓/✓	Common Organisms	✓	✓	✘
3. Geneious	Biomatters	✓/✓	Eukaryotes, Bacteria, Plants, Viruses	✘	✓	✓ (RefSeq, PubMed)
4. Nexus Expression	Biodiscovery	✘/✘	Mammals, Worms, Fish, Birds, Plants	✘	✓	✓ (RefSeq)
5. Partek Flow	Partek	✓/✓	Human, Mouse, Rat, Plants, Worms	✓	✓	✓ (GENCODE ,Ensembl, RefSeq)
6. Rosalind	OnRamp Bioinformatics	✓/ ✘	Human, mouse, rat, drosophilia, Zebrafish, worm and yeast	✓	✓	✘
7. Strand NGS	Strand Scientific Intelligence	✓/✓	40 different organisms	✓	✓	✘

Από τον Πίνακα 3.2 προκύπτουν τα ακόλουθα συμπεράσματα:

Αρχικά, όσον αφορά στα αρχεία εισαγωγής FASTQ και BAM αυτό που προκύπτει είναι ότι το λογισμικό Nexus Expression υστερεί πλήρως ενώ το Rosalind υστερεί μερικώς αφού δεν υποστηρίζει BAM αρχεία ως αρχεία εισαγωγής.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Αναφορικά με τα γονιδιώματα των οργανισμών διαπιστώνεται ότι όλα τα λογισμικά διαθέτουν γονιδιώματα για την πλειοψηφία των οργανισμών και επομένως δεν μπορεί να εξαχθεί κάποιο ασφαλές συμπέρασμα με βάση αυτό το χαρακτηριστικό.

Στην ενότητα του ποιοτικού ελέγχου τα μόνα λογισμικά που δεν υποστηρίζουν κάποιο χαρακτηριστικό γνώρισμα ελέγχου είναι το Nexus Expression και το Geneious.

Η ανάλυση διαφορικής έκφρασης γονιδίων είναι ένα χαρακτηριστικό το οποίο πληρείται από όλα ανεξαιρέτως τα λογισμικά των εταιρειών και επομένως δεν μπορεί να θεωρηθεί ότι κάποιο υπερτερεί ή υστερεί έναντι των υπολοίπων όσον αφορά σε αυτό το χαρακτηριστικό.

Αναφορικά με τις εξωτερικές βάσεις σχολιασμών διαφορικής γονιδιακής έκφρασης φαίνεται ότι δεν μπορεί να προκύψει κάποιο ασφαλές συμπέρασμα αφού 4 από τα 7 λογισμικά πληρούν αυτό το χαρακτηριστικό ενώ στα υπόλοιπα τρία δεν περιλαμβάνεται αυτό στην βασική τους έκδοση.

Συνοψίζοντας, τα λογισμικά τα οποία δείχνουν να πληρούν όλα τα χαρακτηριστικά του πίνακα 3.2 είναι το CLC Genomics Workbench της εταιρείας Qiagen καθώς και το Partek Flow της εταιρείας Partek.

Πίνακας 3.3: Χαρακτηριστικά ανάλυσης δεδομένων (II)

Λογισμικό	Εταιρεία	GO Analysis	Biochemical Pathway	Workflows Creation	Clustering Analysis	Custom Reports
1. CLC Genomics Workbench	CLC bio, Qiagen	✓	✗	✓	✓	✗
2. Full Lasergene Suite	DNASTAR	✓	✗	✓	✓	✗
3. Geneious	Biomatters	✗	✗	✓	✗	✗
4. Nexus Expression	Biodiscovery	✓	✓	✗	✓	✗
5. Partek Flow	Partek	✓	✗	✓	✓	✗
6. Rosalind	OnRamp Bioinformatics	✗	✗	✗	✓	✗
7. Strand NGS	Strand Scientific Intelligence	✓	✓	✓	✓	✗

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Από τον Πίνακα 3.3 προκύπτουν τα ακόλουθα:
Η ανάλυση εμπλουτισμού της γονιδιακής οντολογίας καλύπτεται από την πλειοψηφία των λογισμικών με εξαίρεση αυτό της εταιρείας Biomatters το οποίο προσφέρει αυτή τη δυνατότητα με χρήση όμως επιπρόσθετου υπολογιστικού εργαλείου (Blast2GO) το οποίο επιβαρύνει οικονομικά τον χρήστη.

Επιπλέον, τα μόνα λογισμικά που παρέχουν στην βασική τους έκδοση την ανάλυση βιοχημικών μονοπατιών είναι το Nexus Expression και το Strand NGS με τα υπόλοιπα λογισμικά είτε να μην την περιλαμβάνουν ή να απαιτείται η χρήση κάποιου επιπρόσθετου λογισμικού το οποίο παρέχεται στο χρήστη με επιπλέον χρέωση.

Το λογισμικό της εταιρείας Biodiscovery και αυτό της OnRamp Bioinformatics υστερούν σε λειτουργικότητα (ροές εργασιών) σε σχέση με τα λογισμικά των υπόλοιπων εταιρειών.

Το λογισμικό Geneious της εταιρείας Biomatters αποτελεί το μοναδικό από τα 7 συνολικά λογισμικά το οποίο δεν υποστηρίζει την ανάλυση των γονιδίων κατά συστάδες χωρίς επιπρόσθετη χρέωση για τον χρήστη.

Η τελευταία κατηγορία αυτού του πίνακα που σχετίζεται με τη δημιουργία αναφορών για συγκεκριμένα γονίδια, τα οποία αποτελούν αντικείμενο ενδιαφέροντος για τον εκάστοτε χρήστη, φαίνεται να μην περιλαμβάνεται σε κάποιο λογισμικό και άρα δεν αποτελεί μέτρο σύγκρισης του ανταγωνισμού.

Συνοψίζοντας, τα λογισμικά Nexus Expression και Strand NGS είναι εκείνα τα οποία υποστηρίζουν 4 από τα 5 χαρακτηριστικά αυτού του πίνακα και άρα είναι επικρατέστερα στον ανταγωνισμό έναντι των υπολοίπων.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Πίνακας 3.4: Χαρακτηριστικά οπτικοποίησης δεδομένων

Λογισμικό	Εταιρεία	Heat maps	Volcano/ MA plots	Genome Browser	PCA/MDS plots	Boxplots
1. CLC Genomics Workbench	CLC bio, Qiagen	✓	✓/✓	✓	✓/✗	✓
2. Full Lasergene Suite	DNASTAR	✓	✗/✗	✓	✗/✗	✗
3. Geneious	Biomatters	✓	✓/✗	✓	✓/✗	✗
4. Nexus Expression	Biodiscovery	✓	✗/✗	✗	✗/✗	✗
5. Partek Flow	Partek	✓	✓/✗	✓	✓/✗	✓
6. Rosalind	OnRamp Bioinformatics	✓	✗/✓	✗	✗/✓	✓
7. Strand NGS	Strand Scientific Intelligence	✓	✗/✓	✓	✓/✗	✓

Με βάση τον Πίνακα 3.4 προκύπτουν τα ακόλουθα: Όσον αφορά στους Heat maps (χάρτες θερμότητας), είναι εμφανές ότι όλα τα λογισμικά παρέχουν τη δυνατότητα κατασκευής τέτοιων χαρτών και επομένως το χαρακτηριστικό αυτό δεν αποτελεί αξιόλογο μέτρο σύγκρισης μεταξύ των λογισμικών

Η δημιουργία Volcano καθώς και MA γραφικών παραστάσεων υποστηρίζεται από 3 λογισμικά. Περιηγητές γονιδιωμάτων προσφέρονται από όλα τα λογισμικά εκτός του Nexus Expression και του Rosalind. Επομένως, τα 2 αυτά λογισμικά υστερούν σε σχέση με τα υπόλοιπα αναφορικά με αυτό το χαρακτηριστικό.

Τα 4 από τα 7 λογισμικά υποστηρίζουν τη δημιουργία PCA γραφημάτων ενώ το λογισμικό Rosalind είναι το μοναδικό που επιτρέπει τη δημιουργία MDS γραφημάτων.

Το Nexus Expression μαζί με τα λογισμικά Full Lasergene Suite και Geneious δεν παρέχουν στο χρήστη τη δυνατότητα κατασκευής Box plot διαγραμμάτων.

Συνοψίζοντας, το λογισμικό CLC Genomics Workbench υπερτερεί έναντι των υπολοίπων ως προς τα οπτικά χαρακτηριστικά δεδομένων ακολουθούμενο από το Partek Flow και το Strand NGS.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

Πίνακας 3.5: Κόστος μηνιαίας άδειας χρήσης για έναν χρήστη

Πλατφόρμα	Εταιρεία	Cost($ or euros) per month for an individual user license
1. CLC Genomics Workbench	CLC bio, Qiagen	**413 € + 1281 € (IPA) = 1694 €**
2. Full Lasergene Suite	DNASTAR	**5095 € forever or 217 €**
3. Geneious	Biomatters	**16 $ + 75 $ (Blast2GO) = 91 $**
4. Nexus Expression	Biodiscovery	**232 €**
5. Partek Flow	Partek	**299 € + 282 € (Partek Genomics Suite) = 581 €**
6. Rosalind	OnRamp Bioinformatics	**40 $ per sample + 291 $ (iPathwayGuide) = 331 $**
7. Strand NGS	Strand Scientific Intelligence	**375 $**

Όσον αφορά στο κόστος απόκτησης ατομικής μηνιαίας άδειας χρήσης συμπεραίνεται από τον Πίνακα 3.5 ότι το φθηνότερη λογισμικό είναι το Geneious ενώ το ακριβότερο είναι το CLC Genomics Workbench.

Πρόταση Ιδανικού Λογισμικού

Από τη σκοπιά ενός μη έμπειρου υπολογιστικά χρήστη το ιδανικό λογισμικό ανάλυσης δεδομένων RNA-Seq είναι απαραίτητο να πληροί τα ακόλουθα κριτήρια:

- Ελάχιστο υπολογιστικά χώρο
- Αρχιτεκτονική Cloud
- Υποστήριξη και των 3 τύπων λειτουργικών συστημάτων
- Δυνατότητες ανάλυσης των αποτελεσμάτων της RNA αλληλούχισης (FAST/BAM αρχεία, Organisms, Quality control, Differential Expression analysis, External Annotation Databases, Gene Ontology analysis, Pathway analysis, Workflows Creation, Clustering analysis, Custom Reports) χωρίς να απαιτείται η εγκατάσταση κάποιου επιπρόσθετου plugin λογισμικού με επιπλέον χρέωση για το χρήστη
- Δυνατότητες οπτικοποίησης της διαφορικής γονιδιακής έκφρασης (Heat maps, Volcano ή/και MA plots, PCA ή/και MDS plots και Box plots) χωρίς να απαιτείται η εγκατάσταση κάποιου επιπρόσθετου plugin λογισμικού με επιπλέον χρέωση για το χρήστη
- Ελάχιστο κόστος απόκτησης άδειας χρήσης

Λαμβάνοντας υπόψιν τα προηγούμενα κριτήρια αλλά και την εμπειρία που αποκόμισα από τη δοκιμή των χαρακτηριστικών των λογισμικών κατέληξα στην ακόλουθη κατάταξη των λογισμικών με φθίνουσα σειρά ως προς τη σύγκλιση τους με το ιδανικό λογισμικό:

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

1) **Strand NGS** (good value for money: Καλύτερη σχέση ποιότητας-κόστους)
Το λογισμικό της εταιρείας Strand Scietific Intelligence αν και καταλαμβάνει τον περισσότερο υπολογιστικά χώρο σε σχέση με τα υπόλοιπα, δεν υποστηρίζει αρχιτεκτονική Cloud και τη δυνατότητα σύνδεσης της διαφορικής γονιδιακής έκφρασης με εξωτερικές βάσεις δεδομένων και είναι τέταρτο στην κατηγορία κόστους, ικανοποιεί όλα τα υπόλοιπα κριτήρια χωρίς να απαιτείται η λήψη κάποιου επιπρόσθετου λογισμικού με επιπλέον χρέωση για τον χρήστη.

2) **Partek Flow**
Το λογισμικό της Partek κατατάσσεται δεύτερο αφού παρέχει σχεδόν τις ίδιες δυνατότητες με το πρώτο στο χρήστη, καταλαμβάνει μικρότερο υπολογιστικά χώρο και δεν υποστηρίζει Cloud αρχιτεκτονική. Ωστόσο, η απόκτηση άδειας χρήσης του έχει υψηλότερο κόστος για το χρήστη εξαιτίας της απαίτησης για λήψη επιπρόσθετου λογισμικού.

3) **CLC Genomics Workbench**
Το λογισμικό της Qiagen είναι το πληρέστερο λογισμικό αλλά ταυτόχρονα το κόστος απόκτησης άδειας χρήσης αυτού και του απαιτούμενου για την ανάλυση βιοχημικών μονοπατιών plugin IPA είναι το υψηλότερο.

4) **Rosalind**
Το λογισμικό της OnRamp Bioinformatics υστερεί στην εισαγωγή BAM αρχείων. Η μη υποστήριξη τέτοιων αρχείων αυτή θεωρείται αρκετά σημαντική για αυτό και βρίσκεται σε αυτήν τη θέση.

5) **Full Lasergene Suite**
Το λογισμικό της DNASTAR υποστηρίζει μόνο 2 από τις 7 δυνατότητες οπτικοποίησης δεδομένων για αυτό και είναι πέμπτο στην κατάταξη.

6) **Geneious**
Το λογισμικό της Biomatters ενώ είναι το φθηνότερο σε κόστος, δεν παρέχει τη δυνατότητα ποιοτικού ελέγχου των δειγμάτων του εκάστοτε εξεταζόμενου οργανισμού, ένα χαρακτηριστικό το οποίο κρίνεται ιδιαίτερα σημαντικό για την ανάλυση των αποτελεσμάτων της RNA-Seq αλληλούχισης.

7) **Nexus Expression**
Το λογισμικό της Biodiscovery βρίσκεται στην τελευταία θέση επειδή δεν υποστηρίζει τη δυνατότητα εισαγωγής FASTQ και BAM αρχείων καθιστώντας ουσιαστικά αδύνατη την εφαρμογή οποιουδήποτε λογισμικού.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

4. ΒΙΒΛΙΟΓΡΑΦΙΑ

1) Allan M. Maxam and Walter Gilbert. (1977). A new method for sequencing DNA. *PNAS*, 560-564.
2) Bayley Hagan. (2015). Nanopore Sequencing: From Imagination to Reality. *Clinical Chemistry*, 25-31.
3) Bell C. David et al. (2012). DNA Bulk Identification by Electron Microscopy. *Microscopy and Microanalysis*, 1049-1053.
4) Byron S.A. et al. (2016). Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nature Reviews Genetics volume*, 257-271.
5) Gharajeh, M. S. (2018). *Chapter Eight - Biological Big Data Analytics.* Pethuru Raj and Ganesh Chandra Deka.
6) Goodwin S. et al. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics volume*, 333-351.
7) Gut, I. G. (2013). New sequencing technologies. *Clinical and Transational Oncology*, 879-881.
8) Gužvić, M. (2013). THE HISTORY OF DNA SEQUENCING. *Journal of Medical Biochemistry*, 301-312.
9) Heather M. James and Chain Benjamin. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 1-8.
10) http://www.biodiscovery.com/nexus-expression/
11) http://www.partek.com/pgs
12) http://www.strand-ngs.com/signup/freetrial
13) https://www.dnastar.com/f-reg-submit.aspx
14) https://www.geneious.com/free-trial/
15) https://www.onramp.bio/self-service-bioinformatics
16) https://www.qiagenbioinformatics.com/products/clc-genomics-workbench
17) Kulski K. Jerzy. (2016). An Overview of the History, Tools, and “Omic” Applications. Στο *Next Generation Sequencing* (σσ. 1-58). Λονδίνο: https://www.intechopen.com.
18) Liu Lin et al. (2012). Comparison of Next-Generation Sequencing Systems. *Hindawi Publishing Corporation*, 11.
19) Mardis Elaine R. (2008). Next-Generation DNA Sequencing Methods. *Annu. Rev. Genomics Hum. Genet*, 387-402.
20) Pereira M. Araújo et al. (2017). Application of Next-Generation Sequencing in the Era of Precision Medicine. Στο *Applications of RNA-Seq and Omics Strategies* (σσ. 1-26). Λονδίνο: https://www.intechopen.com/.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154

21) Raza K. and Sabahuddin A. (2016). Recent advancement in Next Generation Sequencing techniques and its computational analysis. *International Journal of Bioinformatics Research and Applications*, 31.
22) Ronaghi M. et al. (1996). Real-Time DNA Sequencing Using Detection. *Analytical Biochemistry*, 84-89.
23) Sanger F. et al. (1977). DNA sequencing with chain-terminating inhibitors. *PNAS*, 5463-5467.
24) Schadt E. Eric et al. (2010). A window into third-generation sequencing. *Human Molecular Genetics*, 227-240.
25) Shan Li et al. (2014). A Survey on Evolutionary Algorithm Based Hybrid. *BioMed Research International*, 8.
26) Zhang, Z. (2014). Big data and clinical research: focusing on the area of critical care medicine in mainland China. *Quantitative Imaging in Medicine and Surgery*, 426-429.

Institutional Repository - Library & Information Centre - University of Thessaly
21/05/2024 13:53:20 EEST - 3.128.171.154