# ΕΛΛΗΝΙΚΟ PORTAL ΒΙΟΠΛΗΡΟΦΟΡΙΚΗΣ ΑΝΑΛΥΣΗΣ «ΟΜΙΚΩΝ» ΔΕΔΟΜΕΝΩΝ ΣΥΝΘΕΤΙΚΗΣ ΒΙΟΛΟΓΙΑΣ

**Ε. Λαδουκάκης[1]*, Π. Αγιουτάντης[1], Μ. Λογοθέτη[1], Δ. Κέκος[1], Φ. Κολίσης[1,]**

[1]Εργαστήριο Βιοτεχνολογίας, Σχολή Χημικών Μηχανικών, ΕΜΠ, Αθήνα, Ελλάδα

(*ladoukef@central.ntua.gr)

## ΠΕΡΙΛΗΨΗ

Με τον όρο «Συνθετική Βιολογία» ορίζουμε τον κλάδο της Βιολογίας που στοχεύει στο σχεδιασμό και τη σύνθεση νέων βιολογικών συστημάτων που δεν συναντώνται στη φύση, χρησιμοποιώντας την τεχνογνωσία από ένα μεγάλο εύρος διαφορετικών τομέων έρευνας όπως οι επιστήμες μηχανικών, η βιοτεχνολογία, η χημεία κ.α. Οι κύριες μέθοδοι αξιοποίησης των εφαρμογών της Συνθετικής Βιολογίας χωρίζονται σε δύο κύριες κατηγορίες [1]. Η μία περιλαμβάνει τη χρήση ειδικά σχεδιασμένων μορίων για την αναπαραγωγή φυσικών βιολογικών διεργασιών με σκοπό τη δημιουργία τεχνητής ζωής για συγκεκριμένες βιοτεχνολογικές εφαρμογές , ενώ η δεύτερη εστιάζει στην απομόνωση λειτουργικών οντοτήτων από τη φύση και χρήση τους ως δομικούς λίθους για την ανακατασκευή νέων τεχνητών βιολογικών συστημάτων. Σε κάθε μία από τις παραπάνω μεθοδολογίες σημαντική τροχοπέδη αποτελεί το κομμάτι του σχεδιασμού και της μοντελοποίησης των βιολογικών συστημάτων [2] και απαιτεί εξαιρετικά πολυσύνθετες βιοπληροφορικές προσεγγίσεις καθώς επίσης και ειδικά διαμορφωμένες υπολογιστικές υποδομές με δυνατότητες διαχείρισης και ανάλυσης δεδομένων μεγάλου όγκου [3]. Στην εργασία αυτή παρουσιάζεται η ανάπτυξη μίας διαδικτυακής πλατφόρμας διαχείρισης και βιοπληροφορικής ανάλυσης «ομικών» δεδομένων, σχεδιασμένη για την επίλυση των προβλημάτων που προκύπτουν κατά τον σχεδιασμό και την in silico προσομοίωση συστημάτων Συνθετικής Βιολογίας. Η πλατφόρμα αυτή περιλαμβάνει μία «εργαλειοθήκη» προγραμμάτων και ειδικά διαμορφωμένων αλγορίθμων που αναλαμβάνουν τη διαχείριση δεδομένων μεγάλου όγκου καθώς και ένα πλήθος διαφορετικών αναλύσεων, από το επίπεδο του γονιδιώματος έως την ανακατασκευή των μεταβολικών μονοπατιών που διέπουν τόσο τους φυσικούς όσο και τεχνητούς οργανισμούς. Τα εργαλεία αυτά διατίθενται διαδικτυακά μέσω ενός διαδραστικού γραφικού περιβάλλοντος χρήστη, καθιστώντας έτσι τη χρήση τους αρκετά εύκολη ακόμα και για χρήστες χωρίς ιδιαίτερες γνώσεις πληροφορικής. Ταυτόχρονα, η πλατφόρμα περιλαμβάνει έτοιμες αυτοματοποιημένες ροές υπολογιστικών διεργασιών με προκαθορισμένες τις βέλτιστες παραμέτρους λειτουργίας των προγραμμάτων τους, μειώνοντας ακόμα περισσότερο την πολυπλοκότητα βασικών αναλύσεων. Η υπολογιστική υποδομή στην οποία στηρίζεται η πλατφόρμα περιλαμβάνει έναν διακομιστή (server) υψηλών υπολογιστικών δυνατοτήτων (64 CPUs – 512 GB RAM – 7,2 TB αποθηκευτικός χώρος) ο οποίος είναι εγκατεστημένος στο Υπολογιστικό Κέντρο της σχολής Χημικών Μηχανικών ΕΜΠ, προσφέροντας έτσι δυνατότητες ταυτόχρονης αξιοποίησης της πλατφόρμας από πολλαπλούς χρήστες. Το εν λόγω εγχείρημα αποτελεί την καρδιά της ψηφιακής ενοποίησης της Εθνικής Ερευνητικής Υποδομής OMIC-ENGINE (MIS 5002636), η οποία έχει σκοπό την προώθηση της έρευνας στην Συνθετική Βιολογία και τη δημιουργία προϊόντων υψηλής προστιθέμενης αξίας στο ελληνικό αγροδιατροφικό σύμπλεγμα και χρηματοδοτείται από το Επιχειρησιακό Πρόγραμμα «Ανταγωνιστικότητα, Επιχειρηματικότητα και Καινοτομία» (ΕΣΠΑ 2014-2020).

## ΕΙΣΑΓΩΓΗ

Η διαρκής προσπάθεια για την πλήρη αποσαφήνιση των πολύπλοκων βιολογικών λειτουργιών τόσο των προκαρυωτικών όσο και των ευκαρυωτικών κυττάρων έχει οδηγήσει στον προσδιορισμό των διακριτών βιολογικών λειτουργικών οντοτήτων (biological parts) [4, 5] από τις οποίες διέπονται. Ο κλάδος της Συνθετικής Βιολογίας πραγματεύεται τη χρήση αυτών των οντοτήτων ως δομικούς λίθους για τον σχεδιασμό και τη σύνθεση νέων τεχνητών βιολογικών συστημάτων που δεν συναντώνται στη φύση. Η δημιουργία των εν λόγω συστημάτων μπορεί να περιλαμβάνει από τον σχεδιασμό ενός απλού μεταβολικού μονοπατιού και τη σύνθεση των αντίστοιχων γονιδίων του με σκοπό τον μετασχηματισμό ενός κυττάρου ξενιστή, έως τη δημιουργία μικροοργανισμών των οποίων η λειτουργία βασίζεται αποκλειστικά σε de-novo κατασκευασμένου συνθετικού γονιδιώματος [6]. Ο κύριος στόχος τέτοιων εγχειρημάτων είναι ο επιτυχής «επαναπρογραμματισμός» κυττάρων είτε για την κατεύθυνση μίας

συγκεκριμένης κυτταρικής λειτουργίας σε επιθυμητά επίπεδα (π.χ. την υπερπαραγωγή κάποιου κυτταρικού προϊόντος), είτε για την εισαγωγή εντελώς νέων κυτταρικών λειτουργιών [7], ή για την περαιτέρω κατανόηση των βιολογικών συστημάτων [8]. Για τον σχεδιασμό αλλά και για τη μετέπειτα αξιολόγηση των ιδιοτήτων των συστημάτων αυτών μπορούν να αξιοποιηθούν πλέον οι συνεχώς εξελισσόμενες –ομικές τεχνολογίες, προσφέροντας υψηλής ποιότητας πληροφορία από το επίπεδο των γονιδιακών αλληλουχιών έως το επίπεδο των εκφραζόμενων πρωτεϊνών και των μεταβολικών μονοπατιών στα οποία αυτές εμπλέκονται. Παρ' όλα αυτά, η υψηλή ανάλυση που προσφέρουν οι –ομικές τεχνολογίες αποτελεί και την κύρια πηγή πολυπλοκότητας για την επεξεργασία των αποτελεσμάτων τους. Στη δυσκολία αυτή συντελεί το μεγάλο μέγεθος των δεδομένων που προκύπτουν από τα σύγχρονα μηχανήματα (π.χ. αλληλούχιση νέας γενιάς) καθώς και η απαίτηση ενός μεγάλου πλήθους διαφορετικών βιοπληροφορικών εργαλείων για την πλήρη ανάλυσή τους, για τη χρήση των οποίων μάλιστα είναι απαραίτητες εξειδικευμένες γνώσεις πληροφορικής. Στην εργασία αυτή περιγράφουμε μία λύση στο πρόβλημα διαχείρισης και ανάλυσης –ομικών δεδομένων για τον σχεδιασμό συνθετικών βιολογικών συστημάτων, η οποία παρουσιάζεται ως μία διαδικτυακή πλατφόρμα βιοπληροφορικών αναλύσεων.

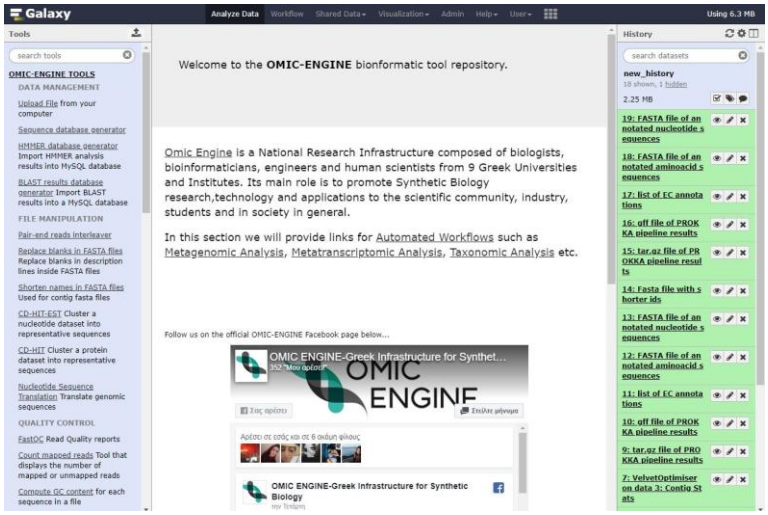
## ΣΧΕΔΙΑΣΜΟΣ ΚΑΙ ΑΝΑΠΤΥΞΗ

Ο σχεδιασμός της διαδικτυακής πλατφόρμας βασίστηκε πάνω στη διαδικτυακή εφαρμογή ανοιχτού κώδικα Galaxy [9] η οποία επιτρέπει την εύκολη αποθήκευση, διαχείριση και ανάλυση δεδομένων μεγάλου όγκου καθώς και τη δημιουργία γραφικού περιβάλλοντος για πληροφορικά εργαλεία και τη διασύνδεσή τους σε αυτοματοποιημένες ροές διεργασιών. Η εγκατάσταση της εφαρμογής Galaxy πραγματοποιήθηκε σε κατάλληλη υπολογιστική υποδομή υψηλών υπολογιστικών δυνατοτήτων (διακομιστής με 64 CPUs, 512 GB RAM και 7,2 TB αποθηκευτικό χώρο) στο Υπολογιστικό Κέντρο της σχολής Χημικών Μηχανικών ΕΜΠ και η διασύνδεσή της με το διαδίκτυο έγινε μέσω Apache HTTPD Server [10] καθιστώντας την προσβάσιμη στη διεύθυνση: http://motherbox.chemeng.ntua.gr/omic-engine/. Στον διακομιστή αυτόν, ο οποίος λειτουργεί με λογισμικό CentOS Linux εγκαταστάθηκαν τα κατάλληλα προγράμματα καθώς και οι αντίστοιχες βάσεις δεδομένων γνωστών αλληλουχιών [11, 12] που απαιτούνται για την πλήρη ανάλυση –ομικών δεδομένων. Συνοπτικά τα προγράμματα αυτά μπορούν να χωριστούν στις εξής κατηγορίες:

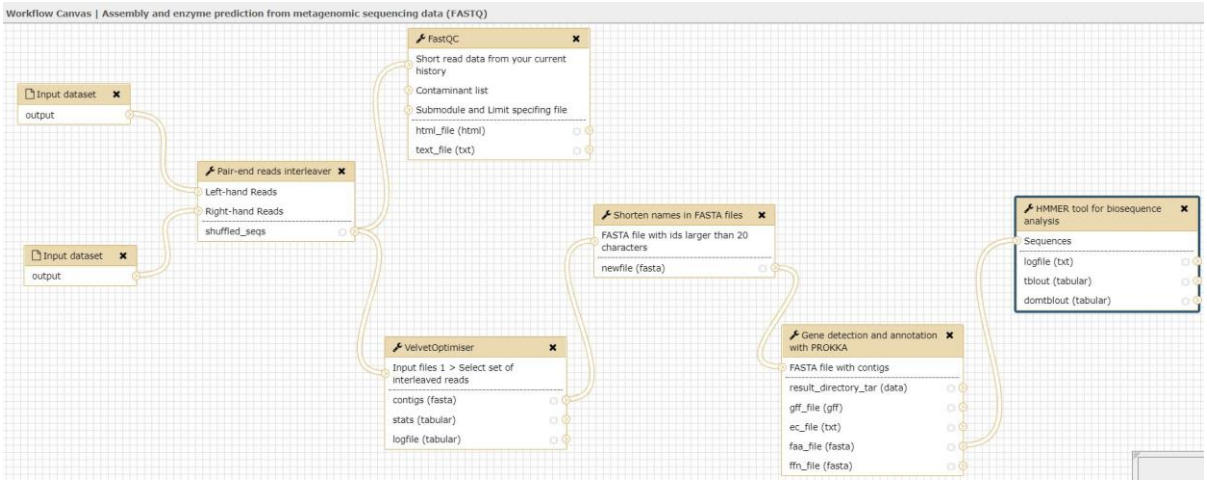
1) Διαχείριση και προ-επεξεργασία δεδομένων [13-15]
2) Ποιοτικός έλεγχος δεδομένων αλληλούχισης [16]
3) Συναρμολόγηση (assembly) δεδομένων αλληλούχισης [17-19]
4) Εντοπισμός και ταυτοποίηση γονιδιακών αλληλουχιών [20-22]
5) Ταξονομική και λειτουργική ανάλυση [23]

Η διασύνδεση κάθε εγκαταστημένου προγράμματος με τη διαδικτυακή πλατφόρμα έγινε μέσω ανάπτυξης εξειδικευμένων αλγορίθμων σε γλώσσα προγραμματισμού eXtensible Markup Language (XML) για τη δημιουργία διαδραστικού γραφικού περιβάλλοντος μέσω της πλατφόρμας και γλώσσα προγραμματισμού Python [24] για τον έλεγχο της ροής των δεδομένων εισόδου/εξόδου (parser scripts). Η βελτιστοποίηση των παραμέτρων κάθε αλγόριθμου βασίστηκε στην αντίστοιχη υπολογιστική πλατφόρμα μεταγενωμικών αναλύσεων ANASTASIA [25]. Στην τελική μορφή της πλατφόρμας που δημιουργήθηκε (Σχήμα 1.), η μεταφορά δεδομένων μπορεί να γίνει είτε διαδικτυακά με απλή μεταφόρτωση (upload) από τον χρήστη, ή μέσω File Transfer Protocol (FTP) απευθείας στον διακομιστή από όπου μπορεί να έχουν πρόσβαση τα αντίστοιχα εργαλεία ανάλυσης. Στην πλατφόρμα αυτή αναπτύχθηκαν επίσης αυτοματοποιημένες ροές υπολογιστικών διεργασιών οι οποίες επιτρέπουν τη διαδοχική λειτουργία πολλαπλών εργαλείων καθιστώντας αυτόματη τη μεταφορά δεδομένων ανάλυσης μεταξύ τους. Έτσι τα αποτελέσματα ανάλυσης του ενός εργαλείου μεταφέρονται αυτόματα σε ένα δεύτερο ως δεδομένα εισόδου, η ανάλυση του οποίου με τη σειρά της τροφοδοτεί το επόμενο εργαλείο κοκ. Κατά αυτόν τον τρόπο περιορίζεται η συνεισφορά του χρήστη μόνο στο να παρέχει τα αρχικά δεδομένα και δεν απαιτούνται ιδιαίτερες γνώσεις πληροφορικής για την λειτουργία του κάθε εργαλείου ξεχωριστά. Οι αυτοματοποιημένες ροές υπολογιστικών διεργασιών (Σχήμα 2.) που περιλαμβάνει η πλατφόρμα δέχονται ως αρχικά δεδομένα εισόδου, αρχεία σε μορφή FASTA ή FASTQ και μέσω της ανάλυσής τους προκύπτει μία λίστα από πιθανά γονίδια και τα αντίστοιχα πρωτεϊνικά μοτίβα (protein domains) που μπορούν να χρησιμοποιηθούν για την ταυτοποίηση της πιθανής ενζυμικής τους λειτουργίας. Επιπλέον, ο κάθε χρήστης έχει τη δυνατότητα να

κατασκευάσει, με βάση τις αναλύσεις του, τις δικές του αυτοματοποιημένες ροές υπολογιστικών διεργασιών έτσι ώστε να προσαρμόσει τις παραμέτρους της κάθε ανάλυσης στις ανάγκες των αντίστοιχων δεδομένων. Τέλος, καθώς η πλατφόρμα είναι βασισμένη στην εφαρμογή Galaxy, η χρήση της από κάθε χρήστη πραγματοποιείται μέσω ενός συστήματος εισόδου με όνομα χρήστη και κωδικό πρόσβασης, έτσι ώστε να διασφαλίζεται η ασφαλής πρόσβαση προς αναλύσεις και τα δεδομένα του.




***Σχήμα 1.*** *Η αρχική σελίδα της διαδικτυακής πλατφόρμας βιοπληροφορικών αναλύσεων. Το αριστερό πλαίσιο (γκρι χρώμα) περιλαμβάνει τα εργαλεία ανάλυσης που είναι συνδεδεμένα με την πλατφόρμα ενώ το δεξιό πλαίσιο (πράσινο χρώμα) περιλαμβάνει τα δεδομένα που είτε έχουν μεταφορτωθεί από τον χρήστη είτε έχουν προκύψει από τη λειτουργία ανάλυσης κάποιου εργαλείου.*




***Σχήμα 2.*** *Σχηματική αναπαράσταση αυτοματοποιημένης ροής διεργασιών όπως αυτή φαίνεται από το σύστημα διαχείρισης της διαδικτυακής πλατφόρμας (workflow canvas). Τα ορθογώνια πλαίσια υποδηλώνουν την λειτουργία κάποιου εργαλείου ανάλυσης ενώ οι γραμμές υποδηλώνουν τη μεταφορά αρχείων (από αριστερά προς τα δεξιά) μεταξύ των διαφορετικών εργαλείων.*

## ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή παρουσιάστηκε ο σχεδιασμός και η ανάπτυξη μίας διαδικτυακής πλατφόρμας αυτοματοποιημένης ανάλυσης –ομικών δεδομένων για τις ανάγκες που προκύπτουν στον κλάδο της Συνθετικής Βιολογίας. Η πλατφόρμα έχει διασυνδεθεί με κατάλληλα εργαλεία βιοπληροφορικών αναλύσεων και είναι εγκατεστημένη σε μία υπολογιστική υποδομή υψηλών υπολογιστικών δυνατοτήτων που καθιστά ευκολότερη τη διαχείριση και ανάλυση δεδομένων μεγάλου όγκου. Ταυτόχρονα το διαδραστικό γραφικό της περιβάλλον και οι αυτοματοποιημένες ροές υπολογιστικών διεργασιών της καθιστούν την ανάλυση δεδομένων αρκετά εύχρηστη ακόμα και σε χρήστες με μικρή ή καθόλου εμπειρία στον κλάδο της Βιοπληροφορικής. Η πλατφόρμα αυτή ενώ είναι εξοπλισμένη με έναν μεγάλο αριθμό

διασυνδεμένων εργαλείων έχει τη δυνατότητα να εξελιχθεί περαιτέρω ανάλογα με τις ανάγκες των χρηστών. Καινούρια εργαλεία, καινούριες βάσεις δεδομένων αλλά και καινούριες αυτοματοποιημένες ροές υπολογιστικών διεργασιών μπορούν (και έχει ήδη προγραμματιστεί) να προστεθούν στο μέλλον είτε για να αντικαταστήσουν ή για να εμπλουτίσουν τις ήδη υπάρχουσες.

## ΕΥΧΑΡΙΣΤΙΕΣ

Η εργασία αυτή υλοποιήθηκε στο πλαίσιο της Πράξης «Synthetic Biology: From omics technologies to genomic engineering (OMIC-ENGINE)» (MIS 5002636) που εντάσσεται στη Δράση «Ενίσχυση των Υποδομών Έρευνας και Καινοτομίας» και χρηματοδοτείται από το Επιχειρησιακό Πρόγραμμα «Ανταγωνιστικότητα, Επιχειρηματικότητα και Καινοτομία» στο πλαίσιο του ΕΣΠΑ 2014-2020, με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης (Ευρωπαϊκό Ταμείο Περιφερειακής Ανάπτυξης). Οι συγγραφείς εκφράζουν θερμές ευχαριστίες προς τη Σχολή Χημικών Μηχανικών του ΕΜΠ για τη διάθεση του Υπολογιστικού Κέντρου.


## ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Benner, S.A. and A.M. Sismour, *Synthetic biology.* Nat Rev Genet, 2005. **6**(7): p. 533-43.
2. Cheng, A.A. and T.K. Lu, *Synthetic Biology: An Emerging Engineering Discipline.* Annual Review of Biomedical Engineering, 2012. **14**(1): p. 155-178.
3. Altaf-Ul-Amin, M., et al., *Systems biology in the context of big data and networks.* Biomed Res Int, 2014. **2014**: p. 428570.
4. Shetty, R.P., D. Endy, and T.F. Knight, Jr., *Engineering BioBrick vectors from BioBrick parts.* J Biol Eng, 2008. **2**: p. 5.
5. Voigt, C.A., *Genetic parts to program bacteria.* Curr Opin Biotechnol, 2006. **17**(5): p. 548-57.
6. Gibson, D.G., et al., *Creation of a bacterial cell controlled by a chemically synthesized genome.* Science, 2010. **329**(5987): p. 52-6.
7. Heidorn, T., et al., *Synthetic biology in cyanobacteria engineering and analyzing novel functions.* Methods Enzymol, 2011. **497**: p. 539-79.
8. Drubin, D.A., J.C. Way, and P.A. Silver, *Designing biological systems.* Genes Dev, 2007. **21**(3): p. 242-54.
9. Giardine, B., et al., *Galaxy: a platform for interactive large-scale genome analysis.* Genome Res, 2005. **15**(10): p. 1451-5.
10. *The Apache HTTP Server Project*. Available from: https://httpd.apache.org/.
11. Jenuth, J.P., *The NCBI. Publicly available tools and resources on the Web.* Methods Mol Biol, 2000. **132**: p. 301-12.
12. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2000. **28**(1): p. 263-6.
13. *The MariaDB Foundation – Supporting continuity and open collaboration in the MariaDB ecosystem*. Available from: https://mariadb.org/.
14. Rice, P., I. Longden, and A. Bleasby, *EMBOSS: the European Molecular Biology Open Software Suite*. Trends Genet, 2000. **16**(6): p. 276-7.
15. Li, W. and A. Godzik, *Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.* Bioinformatics, 2006. **22**(13): p. 1658-9.
16. *FastQC - A quality control tool for high throughput sequence data*. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
17. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
18. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p. 1674-6.

19. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.
20. Eddy, S.R., *Profile hidden Markov models.* Bioinformatics, 1998. **14**(9): p. 755-63.
21. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
22. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* Bioinformatics, 2014. **30**(14): p. 2068-9.
23. Huson, D.H., et al., *MEGAN analysis of metagenomic data.* Genome Res, 2007. **17**(3): p. 377-86.
24. *Python Software Foundation.*
25. Koutsandreas, T., Ladoukakis, E., et al., *(in press), ANASTASIA: An Automated Metagenomic Analysis Pipeline for Novel Enzyme Discovery Exploiting Next Generation Sequencing Data.* Frontiers in Genetics, 2019.