ΥΠΟΛΟΓΙΣΤΙΚΕΣ ΑΣΚΗΣΕΙΣ
ΠΛΗΘΥΣΜΙΑΚΗΣ ΚΑΙ ΕΞΕΛΙΚΤΙΚΗΣ ΓΕΝΕΤΙΚΗΣ

ΜΑΡΙΑΝΘΗ ΓΕΩΡΓΙΤΣΗ – Επίκουρη Καθηγήτρια

Ακαδημαϊκό έτος 2024-2025

ΜΕΡΟΣ 1ο: ΒΑΣΕΙΣ ΔΕΔΟΜΕΝΩΝ

ΕΞΟΙΚΕΙΩΣΗ ΜΕ ΤΗ ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ ENSEMBL

Η ολοκλήρωση της χαρτογράφησης του ανθρώπινου γονιδιώματος το 2001 [International Human Genome Sequencing Consortium (2001), Initial sequencing and analysis of the human genome, *Nature* 409:860-921 και Venter et al. (2001), The sequence of the human genome, *Science* 291:1304-1351], καθώς και η λεπτομερής καταγραφή των παραλλαγών του [International HapMap Consortium (2007), A second generation human haplotype map of over 3.1 million SNPs, *Nature* 449:851-861], συνετέλεσε καθοριστικά στην ανάπτυξη ενός πλήθους βάσεων (γενετικών) δεδομένων που λειτουργούν ως αποθετήρια συλλογής και ως μέσα επικαιροποίησης της συνεχώς αυξανόμενης πληροφορίας από την παραγωγή νέων δεδομένων.

Ορισμένες βάσεις δεδομένων έχουν μεγάλο εύρος και μικρό βάθος (πχ έχουν δεδομένα για μακρομόρια και πολλούς οργανισμούς, αλλά χωρίς εξαντλητικές λεπτομέρειες για το κάθε ένα) και άλλες έχουν μικρό εύρος και μεγάλο βάθος (πχ locus-specific databases, βάσεις δεδομένων για συγκεκριμένα γονίδια/νοσήματα ή οργανισμούς, κα). Μία εξαιρετικά δημοφιλής βάση δεδομένων που χαρακτηρίζεται από μεγάλο εύρος και μεγάλο βάθος πληροφορίας (αν και όχι εξαντλητικό) είναι η Ensembl (https://www.ensembl.org/index.html). Εδώ, μπορούμε να αντλήσουμε πληροφορίες για το γονιδίωμα του ανθρώπου, με το οποίο και θα ασχοληθούμε στην άσκηση αυτή, αλλά και ενός μεγάλου αριθμού άλλων οργανισμών (προκαρυωτικοί, ευκαρυωτικοί), συμπεριλαμβανομένων και όλων των εργαστηριακών οργανισμών. Η βάση δεδομένων Ensembl ξεκίνησε τη λειτουργία της το 1999 και αποτελεί συνεργασία του European Bioinformatics Institute (EBI), του European Molecular Biology Laboratory (EMBL), και του Wellcome Genome Campus (WGC) στο Ηνωμένο Βασίλειο.

Περιληπτικά, η βάση δεδομένων Ensembl, στην οποία θα πλοηγηθούμε, με πρότυπο οργανισμό τον άνθρωπο και το γονιδίωμά του, μας δίνει πληροφορίες για:
(αναφέρονται τα πλέον βασικά στοιχεία)

1. Χρωμοσώματα και συντεταγμένες κάθε χρωμοσώματος
2. Γονίδια και γενετικούς τόπους (επίπεδο gDNA)
3. Μετάγραφα γονιδίων (επίπεδο mRNA, cDNA)
4. Πρωτεΐνες που κωδικοποιούνται από τα γονίδια
5. Γονίδια που δεν κωδικοποιούν πρωτεΐνες
6. Ψευδογονίδια (γονίδια που δεν εκφράζονται)
7. Ρυθμιστικά στοιχεία (σε υποκινητές, ενισχυτές, κλπ)
8. Παραλλαγές του γονιδιώματος (SNVs, CNV, etc) και καταγεγραμμένες συχνότητες
9. Ανισορροπία σύνδεσης σε διαφορετικούς πληθυσμούς

10. Εργαλεία ανάλυσης του γονιδιώματος (BLAST, BioMart, VEP, εργαλεία φυλογένεσης)

Παραθέτουμε ένα link με εκπαιδευτικό υλικό με τη μορφή tutorial για όποιον ενδιαφέρεται να ανατρέξει και μελλοντικά για να εξοικειωθεί περαιτέρω με τη χρήση της βάσης και τις πολλαπλές δυνατότητές της: https://www.ensembl.org/info/website/tutorials/index.html

Πρακτικό Μέρος

1. Μεταβείτε στην αρχική σελίδα του Ensembl και επιλέξτε τη σελίδα: https://www.ensembl.org/info/about/species.html για να δείτε την πλήρη λίστα των διαθέσιμων γονιδιωμάτων στα ζώα. Για τις κατηγορίες άλλων οργανισμών υπάρχουν οι αντίστοιχες «αδελφές σελίδες» Ensembl Bacteria, Ensembl Fungi, Ensembl Plants και Ensembl Protists. Επιλέξτε '**Human**' για το ανθρώπινο είδος (https://www.ensembl.org/Homo_sapiens/Info/Index) και στη συνέχεια επιλέξτε τα παρακάτω links (στο μενού στα αριστερά της σελίδας):

    - **'Genome assembly: GRCh37'** και **'View Karyotype'**: Τι πληροφορίες παίρνουμε από αυτά τα links για το ανθρώπινο γονιδίωμα?

    - Επιστρέψτε στην αρχική σελίδα και επιλέξτε τώρα **'Example region'**: Τι πληροφορίες μπορούμε να αντλήσουμε από την απεικόνιση αυτή?

---

2. Επιστρέψτε στην αρχική σελίδα και πληκτρολογήστε το γονίδιο *AIP*. Επιλέξτε την πρώτη επιλογή ***AIP* (Human Gene)** – Δίνεται και ο απευθείας σύνδεσμος: (https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000110711;r=11:67468174-67491154)

    - Ποιο είναι το πλήρες όνομα του γονιδίου? Έχει συνώνυμες ονομασίες?

    - Σε ποιο χρωμόσωμα εδράζεται το γονίδιο *AIP* και πόσο μήκος (σε base pairs – bp) καταλαμβάνει στο χρωμόσωμα? Τις πληροφορίες αυτές μπορείτε εναλλακτικά να τις βρείτε από την καρτέλα **'Location 11: 67,468,174-67,491,154'**.

    - Πόσα μετάγραφα έχει? Έχει εναλλακτικά μετάγραφα (splice variants)?

    - Ποιο είναι το μήκος του(των) μετάγραφου(ων)? Αντιλαμβάνεστε τη διαφορά μεταξύ μήκους γονιδίου και μήκους μετάγραφου?

---

3. Πατήστε στο link Chromosome 11: 67,468,174-67,491,154 forward strand: Σε ποια κυτταρογενετική ζώνη του q σκέλους του χρωμοσώματος βρίσκεται το γονίδιο *AIP*? Αναφέρετε μερικά από τα γονίδια στην ίδια περιοχή που γειτονεύουν με το *AIP*. Κωδικοποιούν όλα τα γειτονικά γονίδια πρωτεΐνες? Κάνετε zoom in για να δείτε αναλυτικότερα.

---

4. Αφού επιστρέψετε στην προηγούμενη σελίδα, πατήστε το link **Transcript ID** (ENST00000279146.8): Από πόσα εξώνια και πόσα ιντρόνια αποτελείται το γονίδιο *AIP*? Ποιος κλώνος του DNA (+ ή -) κωδικοποιεί το γονίδιο?

---

5. Στην αριστερή πλευρά της σελίδας, πατήστε το link **Exons**: Εδώ βλέπετε την ακριβή ακολουθία του DNA του γονιδίου σε εξώνια και ιντρόνια (κωδικεύων κλώνος μόνο). Ποιο είναι το μεγαλύτερο εξώνιο και ποιο το μεγαλύτερο ιντρόνιο?

---

6. Στη συνέχεια, πατήστε ακριβώς από κάτω στο link **cDNA**: Για ποιο λόγο τα νουκλεοτίδια στην πρώτη σειρά παρουσιάζονται και με μπλε και με μαύρο χρώμα, ενώ πρόκειται για το ίδιο cDNA?

---

7. Να βρείτε τι βάσεις υπάρχουν στις θέσεις του cDNA 47, 112 και 403.

---

8. Για ποιο λόγο κάποιες βάσεις επισημαίνονται με κίτρινο και πράσινο χρώμα? Πόσα SNPs έχει το AIP στην κωδική του περιοχή? Η επιλογή **Variant Table** στο μενού αριστερά (κάτω από το **Genetic variation**) αποκαλύπτει πληροφορίες για όλες τις καταγεγραμμένες παραλλαγές του γονιδίου, τις οποίες μπορείτε να φιλτράρετε περαιτέρω χρησιμοποιώντας τα φίλτρα επάνω από τον συγκεντρωτικό πίνακα. Πόσες σπάνιες παραλλαγές (Global MAF=0-0.08) που είναι επιβαρυντικές για την πρωτεΐνη AIP (SIFT=0-0.3 και PolyPhen=0.7-1) ανιχνεύονται? Τι είδους παραλλαγές είναι?

---

9. Θέστε πίσω τα φίλτρα στις Default τιμές και επιλέξτε στο φίλτρο Consequences να μην εμφανίζονται οι παραλλαγές στα ιντρόνια. Στη συνέχεια κάντε κατηγοριοποίηση των δεδομένων του πίνακα (sorting) βάσει του Consequence Type με τις σοβαρότερες παραλλαγές να εμφανίζονται πρώτες. Τι είδους παραλλαγές είναι και τι πληροφορίες παίρνουμε? Παρατηρήστε πώς μεταβάλλεται η κατηγοριοποίηση των παραλλαγών καθώς «κατεβαίνουμε» προς τα κάτω στον πίνακα.

---

10. Πατήστε στο μενού αριστερά το link **Protein:** Από πόσα αμινοξέα απαρτίζεται η πρωτεΐνη? Εδώ βλέπετε την ακριβή ακολουθία της πρωτεΐνης. Για ποιο λόγο εναλλάσσονται τα αμινοξέα σε μπλε και μαύρο χρώμα? Θα μπορούσε να συμφωνεί αυτό με τις μπλε και μαύρες βάσεις του cDNA?

---

11. Στο μενού αριστερά, στην επιλογή Protein Information, επιλέξτε **Protein Summary:** Ποιο είναι το μοριακό βάρος (σε kDa) της πρωτεΐνης που εκφράζεται από το *AIP*? Στο προβλεπόμενο μοντέλο AlphaFold ποιο αμινοξύ βρίσκεται στη θέση 14 και σε ποιας μορφής δευτεροταγή δομή εντοπίζεται?

---

12. Εντοπίστε στη λίστα παραλλαγών και επιλέξτε αυτήν με το όνομα **rs104894190**. Ποιες είναι οι πληροφορίες που παίρνουμε?

---

13. Ποια είναι η συχνότητα του **rs104894190** σε διαφορετικούς πληθυσμούς του project gnomAD? Πώς ερμηνεύεται ότι μια παραλλαγή παρατηρείται σε πληθυσμούς συγκεκριμένης γεωγραφικής περιοχής, αλλά όχι σε πληθυσμούς άλλης?

---

14. Πληκτρολογήστε τώρα την παραλλαγή **rs11823597.** Ποια είναι η συχνότητά της σε διαφορετικούς πληθυσμούς, όπως έχει καταγραφεί από διαφορετικά projects ανάλυσης του ανθρώπινου γονιδιώματος? Τι παρατηρείτε?

---

15. Επιστρέψτε στην αρχική σελίδα της συγκεκριμένης παραλλαγής και επιλέξτε το εικονίδιο για το Linkage Disequilibrium. Επιλέξτε κάποιον πληθυσμό για προβολή, για παράδειγμα τους Καυκάσιους-CEU από το 1000Genomes_phase 3. Στην επιλογή 'Variants in high LD' βλέπετε το αποτέλεσμα της ισχυρής σύνδεσης (για r^2, D' >0.8). Δοκιμάστε και τους Αφρικανούς-ACB, τι παρατηρείτε?

---

16. Μπορείτε να δείτε την γραφική αναπαράσταση του LD επιλέγοντας View plot ή και σε πίνακα, επιλέγοντας View table, όπου γίνεται παρουσίαση του LD ανά ζεύγη παραλλαγών (pairwise comparison).

17. Και τώρα ξεκινήστε από την αρχή και ζητήστε το γονίδιο *BRCA1*. Επιλέξτε το πρώτο link ***BRCA1*** **(Human Gene)**. – Δίνεται και ο απευθείας σύνδεσμος: (https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000012048;r=17:43044295-43170245)
    Τί διαφορετικό παρατηρείτε σε σχέση με το *AIP*?