ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ

ΤΟΜΕΑΣ ΓΕΝΕΤΙΚΗΣ, ΑΝΑΠΤΥΞΗΣ ΚΑΙ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

Ανάπτυξη και αξιολόγηση εφαρμογών βιοπληροφορικής με κλινική χρησιμότητα στη γενετική του καρκίνου

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΔΕΣΠΟΙΝΑ ΚΑΛΦΑΚΑΚΟΥ

Πτυχιούχος Τμήματος Πληροφορικής και Τηλεπικοινωνιών ΕΚΠΑ

ΘΕΣΣΑΛΟΝΙΚΗ 2021

ARISTOTLE UNIVERSITY OF THESSALONIKI

FACULTY OF SCIENCES

SCHOOL OF BIOLOGY

DEPARTMENT OF GENETICS, DEVELOPMENT AND MOLECULAR BIOLOGY

Development and evaluation of bioinformatics applications with clinical utility for cancer genetics

PhD Dissertation

DESPOINA KALFAKAKOU

Graduate of the Department of Informatics and Telecommunications, UoA, Greece

THESSALONIKI 2021

Εγώ η Δέσποινα Καλφακάκου βεβαιώνω ότι είμαι συγγραφέας της παρούσας εργασίας και ότι έχω αναφέρει ή παραπέμψει σε αυτήν, ρητά και συγκεκριμένα, όλες τις πηγές από τις οποίες έκανα χρήση δεδομένων, ιδεών, προτάσεων ή λέξεων, είτε αυτές μεταφέρονται επακριβώς (στο πρωτότυπο ή μεταφρασμένες) είτε παραφρασμένες.

Η έγκρισις της παρούσης διδακτορικής διατριβής υπό του Τμήματος Βιολογίας της Σχολής Θετικών Επιστημών του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης δεν υποδηλοί αποδοχήν των γνωμών του συγγραφέως (Ν.5343/1932, άρθρ. 202, παρ. 2).

ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

- Ελένη Δροσοπούλου[1,2,3], Αναπληρώτρια Καθηγήτρια Τμήματος Βιολογίας, ΑΠΘ
- Ζαχαρίας Σκούρας[2,3], Καθηγητής Τμήματος Βιολογίας, ΑΠΘ
- Ειρήνη Κωνσταντοπούλου[2,3], Ερευνήτρια Β', ΕΚΕΦΕ «Δημόκριτος»
- Δρακούλης Γιαννουκάκος[3], Διευθυντής Ερευνών, ΕΚΕΦΕ «Δημόκριτος»
- Αναστασία Κουβάτση[3], Καθηγήτρια, Τμήμα Βιολογίας, ΑΠΘ
- Αλέξανδρος Τριανταφυλλίδης[3], Καθηγητής, Τμήμα Βιολογίας, ΑΠΘ
- Ηλίας Καππάς[3], Επίκουρος Καθηγητής, Τμήμα Βιολογίας, ΑΠΘ

---

[1]Επιβλέπουσα

[2]Μέλος της τριμελούς Συμβουλευτικής Επιτροπής

[3]Μέλος της επταμελούς Εξεταστικής επιτροπής

Προτεινόμενος τρόπος αναφοράς της Διδακτορικής Διατριβής

Καλφακάκου Δέσποινα (2021) Ανάπτυξη και αξιολόγηση εφαρμογών βιοπληροφορικής με κλινική χρησιμότητα στη γενετική του καρκίνου. Διδακτορική διατριβή. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Kalfakakou Despoina (2021) Development and evaluation of bioinformatics applications with clinical utility for cancer genetics. PhD dissertation. Aristotle University of Thessaloniki, Greece (in greek with english summary)

ΕΝΙΣΧΥΣΗ ΕΡΕΥΝΑΣ

ΙΔΡΥΜΑ ΣΤΑΥΡΟΣ ΝΙΑΡΧΟΣ
STAVROS NIARCHOS
FOUNDATION

ΠΡΟΛΟΓΟΣ

Η παρούσα διδακτορική διατριβή πραγματοποιήθηκε στο Εργαστήριο Μοριακής Διαγνωστικής του ΕΚΕΦΕ «Δημόκριτος», σε συνεργασία με τον Τομέα Γενετικής, Ανάπτυξης και Μοριακής Βιολογίας του Τμήματος Βιολογίας του ΑΠΘ. Βασική προϋπόθεση για την εκπόνησή της ήταν η συνεργασία με κάποιους ανθρώπους, χωρίς τους οποίους δε θα μπορούσε να πραγματοποιηθεί.

Θα ήθελα να ευχαριστήσω θερμά την **Αναπληρώτρια Καθηγήτρια κ. Δροσοπούλου Ελένη** για την εμπιστοσύνη που μου έδειξε αναλαμβάνοντας ως Επιβλέπουσα Καθηγήτρια της διδακτορικής διατριβής μου, για το ειλικρινές ενδιαφέρον της, τη στήριξη και τις καίριες παρατηρήσεις της. Επιπλέον, θέλω να ευχαριστήσω τον **Καθηγητή κ. Σκούρα Ζαχαρία**, ο οποίος με μεγάλη προθυμία αποτέλεσε μέλος της τριμελούς συμβουλευτικής επιτροπής, γεγονός που αποτελεί εξέχουσα τιμή για μένα.

Οφείλω ένα τεράστιο ευχαριστώ στην **Ερευνήτρια Β' του Εργαστηρίου Μοριακής Διαγνωστικής κ. Κωνσταντοπούλου Ειρήνη**, για την καθοδήγηση, τη συμβουλή και την επιστημονική, αλλά και τη συναισθηματική, στήριξη που μου παρείχε όλα αυτά τα χρόνια, και ιδιαίτερα πλησιάζοντας στην ολοκλήρωση της διατριβής. Την ευχαριστώ για την εμπιστοσύνη που μου έχει επιδείξει και την αμέριστη βοήθειά της σε κάθε θέμα, όσο μικρό ή μεγάλο.

Ευχαριστώ ολόψυχα τον **Διευθυντή Ερευνών του Εργαστηρίου Μοριακής Διαγνωστικής κ. Δρακούλη Γιαννουκάκο**, που με δέχτηκε στην ερευνητική του ομάδα και μου προσέφερε τα απαραίτητα μέσα για την εκπόνηση της διατριβής μου. Επιπλέον, τον ευχαριστώ για τη στήριξη, την καθοδήγηση, τις συμβουλές του και τις απολαυστικές συζητήσεις που μου έχει χαρίσει.

Θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της επταμελούς εξεταστικής επιτροπής, την **Καθηγήτρια κ. Κουβάτση Αναστασία**, τον **Καθηγητή κ. Τριανταφυλλίδη Αλέξανδρο** και τον **Επίκουρο Καθηγητή κ. Καππά Ηλία**, για τη διάθεσή τους να συμμετάσχουν σε αυτή καθώς και για τον πολύτιμο χρόνο που διέθεσαν για τη διόρθωση της παρούσας Διδακτορικής Διατριβής.

Δε θα παραλείψω να ευχαριστήσω τα υπόλοιπα μέλη του Εργαστηρίου Μοριακής Διαγνωστικής, με τους οποίους έχουμε περάσει πολλές χαρές και δυσκολίες. Αρχικά, ευχαριστώ ολόθερμα την **Ερευνήτρια Γ' κ. Φλωρεντία Φωστήρα** για την αμέριστη συμπαράσταση και υποστήριξή της, για την ουσιαστική καθοδήγησή της σε όλα τα επίπεδα και για τη βοήθεια που μου έχει παρέχει σε όλη αυτή την πορεία. Ως η μόνη βιοπληροφορικός στο εργαστήριο στο ξεκίνημά μου, η προσαρμογή μου είχε πολλές προκλήσεις και η στενή συνεργασία μου με την κ. Φωστήρα ήταν καθοριστική για την αντιμετώπισή τους.

Ευχαριστώ θερμά τις συναδέλφους **κ. Αποστόλου Παρασκευή** και **κ. Δελλατόλα Βασιλική**, καθώς και τον έτερο βιοπληροφορικό του εργαστηρίου, **κ. Παπαθανασίου Αθανάσιο** για τη βοήθειά τους, τις ευχάριστες στιγμές αλλά και τις αγωνίες που μοιραστήκαμε. Επιπλέον, θέλω να ευχαριστήσω τα πρώην

μέλη του εργαστηρίου **κ. Βαγενά Ανδρομάχη, κ. Γιαννακοπούλου Ευγενία, κ. Γαβρά Ιωάννα** και **κ. Βλάχο Ιωάννη**, για την στήριξη τόσο σε επαγγελματικό όσο και σε προσωπικό επίπεδο.

Η διεκπεραίωση μιας διδακτορικής διατριβής είναι μια δύσκολη διαδικασία η οποία κρύβει προκλήσεις, έχει σημαντικό κόστος και απαιτεί θυσίες· όταν μάλιστα συμπίπτει με μια πανδημία φέρει επιπλέον συναισθηματικό φορτίο. Δε θα μπορούσα να τα καταφέρω χωρίς τη στήριξη των ανθρώπων που έχω την τύχη να αποκαλώ «δικούς μου». Αυτοί είναι οι γονείς μου, Ηλίας και Λευκοθέα, οι αδερφές μου, Μαρία, Γιάννα και Αθανασία, οι αγαπημένοι μου φίλοι και ο Γιώργος. Τους ευχαριστώ για όλα.

ΠΕΡΙΕΧΟΜΕΝΑ

CONTENTS

ΠΕΡΙΛΗΨΗ

Η Ιατρική Ακριβείας αποτελεί μία νέα προσέγγιση της ιατρικής που λαμβάνει υπόψη το γενετικό προφίλ του κάθε ασθενούς, με στόχο τη χορήγηση εξατομικευμένης θεραπείας, αλλά και την πρόληψη ασθενειών. Ο καρκίνος είναι μια ασθένεια του γονιδιώματος, κι έτσι η Ιατρική Ακριβείας έχει ιδιαίτερη εφαρμογή στην ογκολογία. Πιο συγκεκριμένα, ένας από τους βασικούς στόχους της Ιατρικής Ακριβείας στην ογκολογία είναι η διάγνωση των γενετικών ιδιαιτεροτήτων του όγκου κάθε ασθενούς και, κατά συνέπεια, η χορήγηση συγκεκριμένης θεραπείας που θα στοχεύει αυτές τις ιδιαιτερότητες. Παράλληλα, μια άλλη πτυχή της Ιατρικής Ακριβείας στην ογκολογία είναι η πρόληψη της νόσου, μέσω της μελέτης του κληρονομικού καρκίνου, η οποία επιτρέπει την έγκαιρη αναγνώριση των ατόμων που έχουν αυξημένο κίνδυνο να διαγνωστούν με καρκίνο και, τελικά, τη σωστή κλινική διαχείρισή τους.

Τα τελευταία χρόνια, η τεχνολογία αλληλούχησης του DNA έχει εξελιχθεί ραγδαία χάρη στις τεχνολογικές βελτιώσεις και τις νέες μεθόδους αυτοματοποίησης. Η εισαγωγή της μεθόδου Αλληλούχησης Επόμενης Γενιάς έχει ως αποτέλεσμα την παραγωγή μεγάλου όγκου γενετικών δεδομένων, τα οποία επιτρέπουν την κατανόηση των βιολογικών μηχανισμών που διέπουν την ανθρώπινη νόσο. Όλες αυτές οι εξελίξεις, έχουν συνηγορήσει στην ενίσχυση και την εφαρμογή της Ιατρικής Ακριβείας, ωστόσο, εισάγουν νέες προκλήσεις σχετικά με την ανάλυση, την καταγραφή και τη διανομή αυτού του τεράστιου όγκου δεδομένων.

Η επιστήμη της βιοπληροφορικής έχει συμβάλει δυναμικά στην αντιμετώπιση αυτών των προκλήσεων. Επιτρέποντας την ανάλυση ακατέργαστων δεδομένων μεγάλης κλίμακας, τα οποία είναι αδύνατο να αναλυθούν χειροκίνητα, την ενσωμάτωση ετερογενών βιολογικών δεδομένων και τη χρήση προηγμένων υπολογιστικών και στατιστικών μεθόδων για την κατανόησή τους, η βιοπληροφορική παρέχει τα εργαλεία για την ανίχνευση βιοδεικτών και υποψηφίων φαρμάκων. Επιπλέον, συμβάλει και στην αποθήκευση αυτού του τεράστιου όγκου δεδομένων, μέσω των βάσεων βιολογικών δεδομένων που έχουν δημιουργηθεί, οι οποίες όχι μόνο διευκολύνουν τη διαχείριση των δεδομένων, αλλά και το διαμοιρασμό τους σε ολόκληρη την επιστημονική κοινότητα.

Σκοπός της παρούσας διδακτορικής διατριβής αποτελεί η μελέτη εφαρμογών βιοπληροφορικής με κλινική χρησιμότητα στη γενετική καρκίνου και την ιατρική ακριβείας. Σε αυτό το πλαίσιο, αναπτύχθηκαν η βάση δεδομένων CanVaS που αποτελεί την πρώτη βάση δεδομένων για τον κληρονομικό καρκίνο που αφορά τον ελληνικό πληθυσμό, η οποία καταγράφει το φάσμα της γενετικής ποικιλομορφίας των Ελλήνων ασθενών με καρκίνο, καθώς και τα φαινοτυπικά και κλινικά χαρακτηριστικά αυτών των ατόμων, και η ροή διοχέτευσης εντολών διεργασιών VarTrace, με σκοπό την ανάλυση της γενετικής πληροφορίας που έχει παραχθεί από Αλληλούχηση Επόμενης Γενιάς DNA όγκου. Τέλος, αξιολογήθηκε ένα εμπορικό λογισμικό για τον εμπλουτισμό παραλλαγών που ανιχνεύονται σε ένα πείραμα Αλληλούχησης Επόμενης Γενιάς, με κλινική εφαρμογή.

Η έλευση της μεθόδου Αλληλούχηση Επόμενης Γενιάς, έχει καταστήσει αναγκαία την καταγραφή των γενετικών και γονιδιωματικών δεδομένων σε πληθυσμιακό επίπεδο, με απώτερο σκοπό την μεγαλύτερη ακρίβεια στην αποτύπωση της γενετικής ετερογένειας μεταξύ, αλλά και εντός των πληθυσμών, καθώς και την έγκυρη κατηγοριοποίηση των σπάνιων αλληλομόρφων. Η ανάπτυξη εθνικών βάσεων γενετικών δεδομένων συνεισφέρει ουσιαστικά στην κατανόηση των γενετικών ιδιαιτεροτήτων ενός πληθυσμού και

κατά συνέπεια στην καλύτερη διαχείριση των ασθενών με καρκίνο. Η βάση CanVaS αποτελείται από τη συλλογή δεδομένων που προέρχονται από τις γενετικές αναλύσεις 7.363 Ελλήνων που έχουν διαγνωσθεί με διαφόρους τύπους καρκίνου ή/και σύνδρομα που συμπεριλαμβάνουν κακοήθειες και υγιών συγγενών τους, οι οποίοι έχουν αναλυθεί για την ύπαρξη γενετικών αλλαγών σε 1 έως 97 γονίδια, παθογόνοι παραλλαγές των οποίων έχουν σχετισθεί με προδιάθεση στον καρκίνο. Οι συγκεκριμένες αναλύσεις έχουν πραγματοποιηθεί στο Εργαστήριο Μοριακής Διαγνωστικής του ΕΚΕΦΕ «Δημόκριτος». Η συλλογή δεδομένων του CanVaS αποτελείται από περίπου 24.000 λειτουργικά χαρακτηρισμένες παραλλαγές. Για κάθε παραλλαγή, περιλαμβάνεται η συχνότητα αλληλομόρφου για τον ελληνικό πληθυσμό, η αξιολόγηση, ερμηνεία και κλινική επίπτωσή της βάσει των κανόνων του American College of Clinical Geneticists (ACMG), καθώς και τα φαινοτυπικά χαρακτηριστικά των ατόμων που τη φέρουν. Επιπλέον, παρέχονται πληροφορίες σχετικά με τη γεωγραφική κατανομή των παραλλαγών σε ολόκληρη τη χώρα, επιτρέποντας τη μελέτη των απομονωμένων ελληνικών πληθυσμών. Αξίζει να σημειωθεί ότι έχει δοθεί προτεραιότητα στην ενσωμάτωση και τη διασύνδεση των δεδομένων με τις εδραιωμένες ανοιχτά προσβάσιμες βάσεις δεδομένων. Η βάση CanVaS υποστηρίζεται από το λογισμικό ανοιχτού κώδικα Leiden Open Variation Database (LOVD), επιτρέποντας έτσι την εύκολη διασύνδεση με κεντρικές πηγές δεδομένων. Κατ' αυτόν τον τρόπο, χρησιμεύει ως μέρος της λύσης στη διαιρεμένη διαθεσιμότητα των γενετικών δεδομένων σε διάφορες βάσεις δεδομένων. Η βάση δεδομένων CanVaS είναι διαθέσιμη στην ηλεκτρονική διεύθυνση: http://ithaka.rrp.demokritos.gr/CanVaS

Η ροή διοχέτευσης εντολών διεργασιών VarTrace αναπτύχθηκε με σκοπό την ακριβή ανίχνευση σωματικών παραλλαγών στο DNA καρκινικών κυττάρων. Η Αλληλούχηση Επόμενης Γενιάς του DNA όγκων χρησιμοποιείται ευρέως για την ανίχνευση βιοδεικτών που μπορούν να αποτελέσουν στόχο για εξατομικευμένη θεραπεία. Χαρακτηριστικό παράδειγμα αποτελεί η εκτίμηση της ανεπάρκειας του ομόλογου ανασυνδυασμού σε όγκους μέσω της ανίχνευσης παθογόνων παραλλαγών στα γονίδια *BRCA1* και *BRCA2*. Οι ασθενείς με ανεπαρκείς όγκους μπορούν να επωφεληθούν από τη χορήγηση αναστολέων των πολυμερασών της πολυαδενοφωσφορικής ριβόζης (polyADP-ribose polymerase, PARP). Ωστόσο, η αλληλούχηση του DNA όγκων κρύβει πολλές προκλήσεις, καθώς το γονιδίωμα του όγκου χαρακτηρίζεται από ετερογένεια, ενώ οι μέθοδοι διατήρησης του ιστού αλλοιώνουν το γενετικό του υλικό. Ως αποτέλεσμα, η αλληλούχηση του DNA των όγκων συχνά οδηγεί στην ανίχνευση αυξημένου αριθμού ψευδώς θετικών παραλλαγών. Το εργαλείο VarTrace αντιμετωπίζει αυτούς τους περιορισμούς μέσω της χρήσης πολλαπλών αλγορίθμων κλήσης παραλλαγών και της εξειδικευμένης προσαρμογής των παραμέτρων του. Προκειμένου να αποδειχθεί η απόδοση και η χρησιμότητά του, το VarTrace εφαρμόστηκε σε δεδομένα που προέρχονται από την αλληλούχηση των γονιδίων *BRCA1* και *BRCA2* σε όγκους 75 ασθενών με επιθηλιακό καρκίνωμα των ωοθηκών, ενώ τα αποτελέσματα συγκρίθηκαν με δύο εμπορικά διαθέσιμες ροές διοχέτευσης εντολών διεργασιών. Το VarTrace αποτελεί μια αυτοματοποιημένη, πλήρως διαμορφώσιμη και παράλληλη ροή διοχέτευσης εντολών διεργασιών, γραμμένη σε R και Perl και είναι διαθέσιμο στην ηλεκτρονική διεύθυνση: https://gitlab.com/dkalfakakou/TumorPipeline.

Τέλος, αξιολογήθηκε ο χαρακτηρισμός παραλλαγών ως προς την επίπτωσή τους, όπως πραγματοποιείται από ένα φιλικό προς το χρήστη κι εμπορικά διαθέσιμο λογισμικό, το Illumina® VariantStudio v3.0 (VS), το οποίο χρησιμοποιεί ένα συγκεκριμένο σύνολο μεταγράφων. Για λόγους

σύγκρισης, χρησιμοποιήθηκε το Ensembl VEP, ένα εργαλείο γραμμής εντολών ανοιχτού κώδικα, καθώς παρέχει ευελιξία όσον αφορά την επιλογή των μεταγράφων. Για την αξιολόγηση, χρησιμοποιήθηκαν δεδομένα που προέκυψαν από τον προσδιορισμό της αλληλουχίας 857 δειγμάτων DNA γαμετικής σειράς ασθενών με καρκίνο με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς. Η σύγκριση μεταξύ των δύο συνόλων μεταγράφων έδειξε αναντιστοιχία σε ποσοστό 82,82%. Επιπλέον, χρησιμοποιώντας το προεπιλεγμένο σύνολο μεταγράφων του VS, δεν χαρακτηρίστηκαν σωστά το 11,45% των ταυτοποιημένων παραλλαγών απώλειας λειτουργίας, γεγονός που στην περίπτωση που ο γονιδιακός έλεγχος έχει κλινική εφαρμογή, μπορεί να έχει άμεσο αντίκτυπο στην κλινική διαχείριση των ασθενών. Τα αποτελέσματά μας επισημαίνουν τη σημασία της προσεκτικής επιλογής λογισμικού και μεταγράφων και την ανάγκη για αξιόπιστη ανάλυση δεδομένων, με σκοπό τη μείωση του ποσοστού σφάλματος των γονιδιακών ελέγχων και κατ’ επέκταση, τη βελτίωση της φροντίδας των ασθενών.

ABSTRACT

Precision Medicine is an emerging approach to medicine that studies the genetic profile of each patient, aiming in providing personalized treatment and preventing disease. Cancer is a disease of the genome, so precision medicine has a special application in oncology. More specifically, one of the main goals of precision medicine in oncology is to identify the distinct genetic features of each patient's tumor and, consequently, to provide specific treatment that targets them. Additionally, another aspect of precision medicine in oncology is disease prevention, through the study of hereditary cancer, which allows the timely identification of individuals who are at increased risk of developing cancer and, ultimately, their proper clinical management.

In recent years, DNA sequencing technology has rapidly evolved thanks to technological improvements and new automation methods. The introduction of Next Generation Sequencing resulted in the production of a large volume of genetic data, which allows the understanding of the biological machineries that govern human disease. These developments have paved the way to precision medicine; they have also introduced new challenges regarding the analysis, the recording, and the distribution of this huge amount of data.

Bioinformatics is a science that has decidedly contributed to addressing these challenges. By allowing the analysis of large-scale raw data that is impossible to analyze manually, the integration of heterogeneous biological data, and the use of advanced computational and statistical methods to understand them, bioinformatics provides the tools for detecting biomarkers and candidate drugs. Moreover, many biological databases have been created for storing this huge amount of data, enabling easy data management, and sharing with the scientific community.

The aim of the present thesis is the study of bioinformatics applications with clinical utility in cancer genetics and precision medicine. In this context, CanVaS (Cancer Variation reSource), the first database on hereditary cancer for the Greek population, was developed, recording the genetic heterogeneity of Greek cancer patients along with their phenotypic and clinical characteristics. Moreover, VarTrace, a pipeline for the analysis of genetic data generated from Next Generation Sequencing of tumor DNA is presented. Finally, a commercial software for the annotation of germline variants identified through Next Generation Sequencing was evaluated.

The advent of Next Generation Sequencing technology has introduced the need to record population-specific genetic data. National genetic variation registries vastly increase the level of detail for the relevant population, while directly affecting patient management. CanVaS comprises a collection of data from the genetic analyses of 7,363 Greek patients that have been diagnosed with various types of cancer and/or cancer syndromes and their healthy relatives. All these individuals have been tested in the Molecular Diagnostics Laboratory of NCSR "Demokritos" for germline variants in 1 to 97 genes that have been associated with predisposition to cancer. The CanVaS data collection consists of approximately 24,000 functionally annotated variants. For each variant, the allele frequency for the Greek population, its evaluation, interpretation, and clinical impact according to the American College of Clinical Geneticists (ACMG) guidelines are recorded. Most importantly, the genetic dataset is accompanied by the phenotypic and clinical traits of the carriers. Moreover, information on the geographical distribution of the variants

throughout the country is provided, allowing the study of Greek isolates. CanVaS is developed using the Leiden Open Variation Database (LOVD) open-source software, allowing easy connection to central data sources. This way, it serves as a part of the solution to the divided availability of genetic data in many different databases. CanVaS is available at: http://ithaka.rrp.demokritos.gr/CanVaS.

The VarTrace pipeline was developed to accurately detect somatic variants in tumor DNA. Next Generation Sequencing of somatic DNA is widely used for biomarker detection that can be targeted for personalized treatment. A prominent example is the assessment of homologous recombination deficiency in tumors, through the detection of pathogenic variants in *BRCA1* and *BRCA2*. Patients with homologous recombination deficient tumors can benefit from the administration of poly ADP ribose polymerase (PARP)-inhibitors. However, tumor DNA sequencing bears many challenges, as the tumor genome is vastly heterogeneous and existing tissue preservation methods alter its genetic material. As a result, tumor DNA sequencing often leads to the detection of an increased number of false-positive variants. VarTrace addresses these limitations using multiple variant calling algorithms and fine-tuned customization of its parameters. To demonstrate its efficacy and utility, we applied VarTrace to data from the sequencing of *BRCA1* and *BRCA2* genes in 75 patients with ovarian epithelial carcinoma and compared the results with two commercially available pipelines. VarTrace is an automated, fully configurable, and parallel pipeline, written in R and Perl and is available at: https://gitlab.com/dkalfakakou/TumorPipeline.

Finally, we evaluated the variant annotation performed by a user-friendly and commercially available software, Illumina® VariantStudio v3.0 (VS), on a clinical setting. VS uses a specific set of transcripts for the annotation process. For the sake of comparison, we employed Ensembl VEP, an open source command line tool, which provides flexibility in transcript selection. Data from the germline DNA sequencing of 857 cancer patients that were analyzed using Next Generation Sequencing were used for the evaluation. The comparison between the two transcript sets showed a discordance rate of 82.82%. Moreover, using the default VS transcript set, 11.45% of the identified loss-of-function variants were characterized incorrectly. In the case of a clinical application of genetic testing, these results could directly and negatively impact patient management. Our results highlight the importance of careful software and transcript selection and the need for reliable data analysis, in order to reduce the error rate of genetic testing and, consequently, improve patient care.

Λέξεις-κλειδιά

Βιοπληροφορική, γενετική του καρκίνου, κληρονομικός καρκίνος, ιατρική ακριβείας, Αλληλούχηση Επόμενης Γενιάς, βάση δεδομένων, πληθυσμιακή βάση δεδομένων, ροή διοχέτευσης εντολών διεργασιών, κλήση παραλλαγών, χαρακτηρισμός παραλλαγών

Keywords

Bioinformatics, Cancer Genetics, Hereditary Cancer, Precision Medicine, Next Generation Sequencing, Database, Population Database, Pipeline, Variant Calling, Variant Annotation

1. ΕΙΣΑΓΩΓΗ

1.1. Ιατρική Ακριβείας στην ογκολογία

Η **Ιατρική Ακριβείας** αποτελεί μία νέα και διαφορετική προσέγγιση που λαμβάνει υπόψη το γενετικό προφίλ, το περιβάλλον και τον τρόπο ζωής κάθε ασθενούς, με στόχο τη χορήγηση εξατομικευμένης θεραπείας, αλλά και την πρόληψη ασθενειών (Ashley, 2016; Hodson, 2016). Ωστόσο, η έννοια της Ιατρικής Ακριβείας δεν είναι νέα. Ο Ιπποκράτης πρώτος επεσήμανε τη σημασία της μελέτης του κάθε ασθενούς ξεχωριστά, ενώ κατέδειξε τη διατροφή, την υγιεινή, αλλά και το περιβάλλον, όπως η γεωγραφία, το κλίμα αλλά και το υδρολογικό περιβάλλον του τόπου κατοικίας του ασθενούς, ως καθοριστικούς παράγοντες που ρυθμίζουν την εμφάνιση και την πορεία των ασθενειών. Μάλιστα, του αποδίδεται το ρητό: «*Ο θεραπευτής πρέπει πρώτα να γνωρίζει ολόκληρο τον άνθρωπο ως μια μοναδική ψυχοσωματική οντότητα σε σχέση με το κοινωνικό και φυσικό του περιβάλλον*» (Pulciani *et al.*, 2017; Marketos).

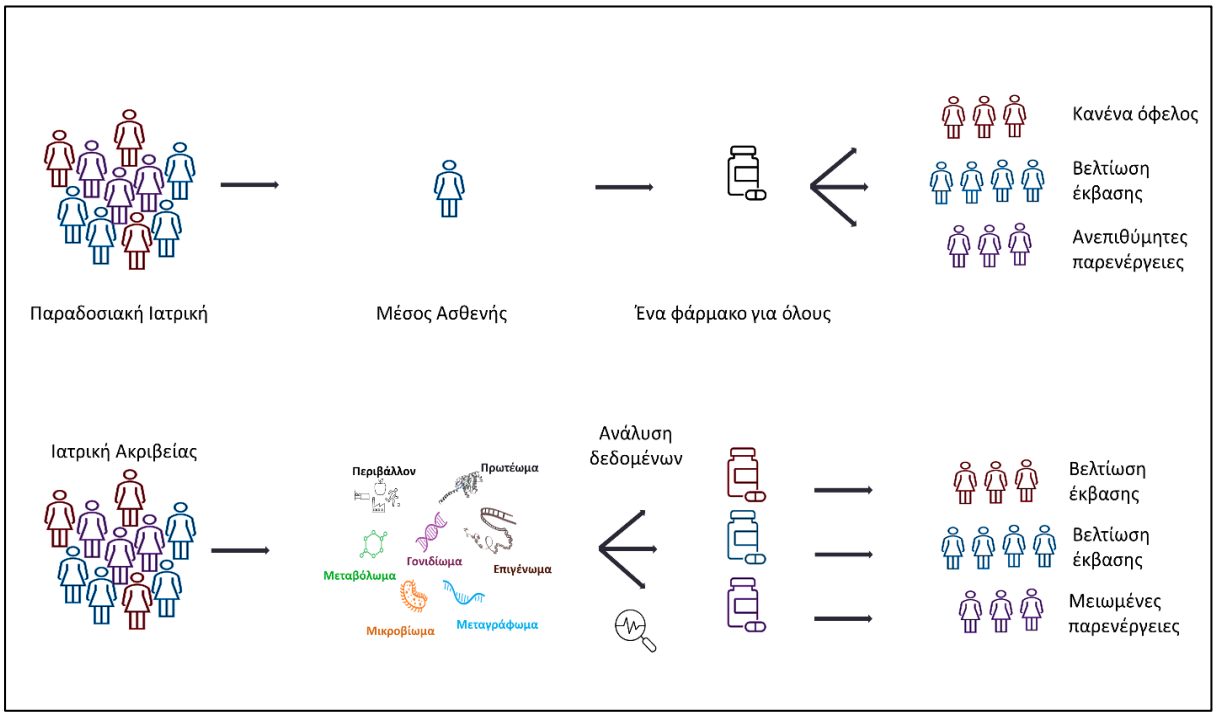
Τις τελευταίες δύο δεκαετίες γινόμαστε μάρτυρες ραγδαίων εξελίξεων στη βιοτεχνολογία, τη γενετική και την επιστήμη των δεδομένων, οι οποίες έχουν φέρει αληθινή επανάσταση στην ιατρική. Πιο συγκεκριμένα, πλέον, είμαστε σε θέση να διαβάσουμε την ακολουθία ολόκληρου του γονιδιώματος ενός ανθρώπου μέσα σε μόνο λίγες ώρες. Παράλληλα, υπάρχουν τα υπολογιστικά εργαλεία και η υπολογιστική δύναμη που χρειάζονται για την ανάλυση των παραγόμενων δεδομένων, ενώ έχει εγκαθιδρυθεί μια πληθώρα βάσεων δεδομένων που επιτρέπουν την καταγραφή και τη διανομή ενός τεράστιου όγκου βιολογικών δεδομένων. Όλες αυτές οι εξελίξεις, έχουν συνηγορήσει στην ενίσχυση και την εφαρμογή της Ιατρικής Ακριβείας.

Όσον αφορά την ογκολογία συγκεκριμένα, ένα βασικό πρόβλημα των παραδοσιακών μεθόδων θεραπείας είναι η χαμηλή ειδικότητα τους, με αποτέλεσμα τον κυτταρικό θάνατο όχι μόνο των καρκινικών κυττάρων, αλλά και των υγιών. Ένας από τους βασικούς στόχους της Ιατρικής Ακριβείας στην ογκολογία είναι η διάγνωση των γενετικών ιδιαιτεροτήτων του όγκου κάθε ασθενούς και, κατά συνέπεια, η χορήγηση συγκεκριμένης θεραπείας που θα στοχεύει αυτές τις ιδιαιτερότητες. Με αυτόν τον τρόπο, η **στοχευμένη θεραπεία** θα οδηγήσει μόνο τα καρκινικά κύτταρα σε απόπτωση, κι έτσι θα είναι λιγότερο τοξική για τον ασθενή, ενώ ιδανικά η έκβαση της νόσου θα είναι καλύτερη (Εικόνα 1).

Παράλληλα, μια άλλη πτυχή της Ιατρικής Ακριβείας στην ογκολογία είναι η πρόληψη της νόσου. Η μελέτη του **κληρονομικού καρκίνου** έχει συμβάλλει αποφασιστικά σε αυτόν το στόχο, μέσω της αναγνώρισης των ατόμων που έχουν αυξημένο κίνδυνο να διαγνωστούν με καρκίνο. Τα άτομα που ανήκουν στις ομάδες υψηλού κινδύνου έχουν διάφορες επιλογές για να αποτρέψουν την εμφάνιση της νόσου, ανάλογα με το γενετικό τους προφίλ, όπως η ένταξή τους σε εξειδικευμένα πρωτόκολλα παρακολούθησης, η προληπτική χορήγηση φαρμάκων αλλά και τα προφυλακτικά χειρουργεία (Kulkarni & Carley, 2016).

Στην παρούσα διατριβή, μελετάται η συμβολή της βιοπληροφορικής στην Ιατρική Ακριβείας και, πιο συγκεκριμένα, ο τρόπος που η βιοπληροφορική χρησιμεύει ως εργαλείο για την ακριβή αλληλούχηση του DNA, με απώτερο σκοπό τη διάγνωση της κληρονομικής προδιάθεσης στον καρκίνο, αλλά και την ανίχνευση βιοδεικτών που αποτελούν στόχο για εξατομικευμένη θεραπεία. Έτσι, στα επόμενα κεφάλαια

της εισαγωγής, θα εξηγηθούν οι μηχανισμοί της καρκινογένεσης, οι βασικές αρχές της γενετικής και οι μέθοδοι βιοπληροφορικής ανάλυσης γενετικών δεδομένων.



Εικόνα 1: Σχηματική αναπαράσταση των διαφορών μεταξύ της παραδοσιακής ιατρικής και της ιατρικής ακριβείας.

Figure 1: Schematic representation of the differences between traditional and precision medicine.

1.2. Η βιολογία του καρκίνου

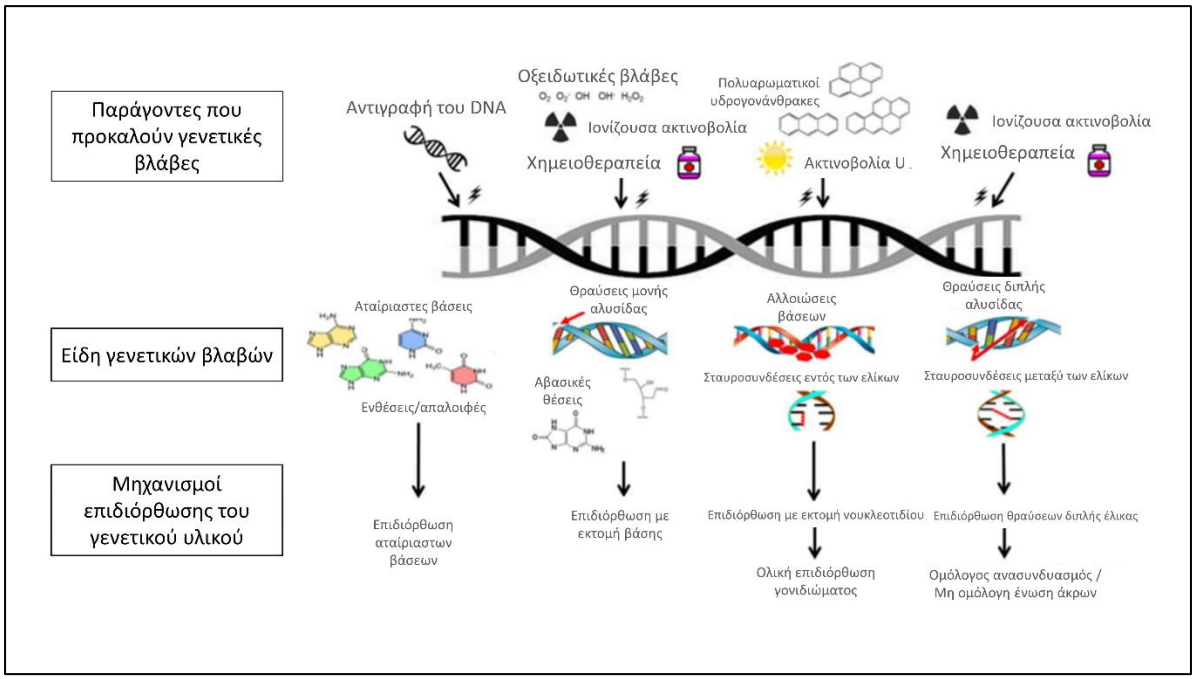
Ο **καρκίνος** είναι μία ασθένεια του γονιδιώματος. Η γενετική αυτή ασθένεια, χαρακτηρίζεται από συσσώρευση **γενετικών αλλαγών**, αλλά και **χρωμοσωμικών ανωμαλιών**, οι οποίες συνήθως συμβαίνουν στα σωματικά κύτταρα κι επομένως δεν είναι κληρονομικές. Κανονικά, τα ανθρώπινα κύτταρα αναπτύσσονται και διαιρούνται ώστε να σχηματίσουν νέα κύτταρα καθώς το σώμα τα χρειάζεται. Όταν τα κύτταρα γερνούν ή πάθουν κάποια ανεπανόρθωτη βλάβη, πεθαίνουν και αντικαθίστανται από νέα κύτταρα (κυτταρικός θάνατος ή απόπτωση) (Cooper & Hausman, 2016). Όταν αναπτύσσεται καρκίνος, αυτή η ομαλή διαδικασία διακόπτεται. Ως αποτέλεσμα αυτών των γενετικών αλλαγών του καρκίνου, τα καρκινικά κύτταρα έχουν την ιδιότητα να πολλαπλασιάζονται ανεξέλεγκτα, ενώ τα γερασμένα ή κατεστραμμένα κύτταρα επιβιώνουν όταν θα έπρεπε να οδηγηθούν σε απόπτωση. Κάθε φορά που τα καρκινικά κύτταρα πολλαπλασιάζονται, υπάρχει μεγάλη πιθανότητα να προκληθούν επιπλέον γενετικές αλλαγές, οι οποίες τελικά δημιουργούν ένα πλήρως κατακερματισμένο γονιδίωμα, με αποτέλεσμα αυτός ο φαύλος κύκλος να συνεχίζεται περαιτέρω. Επιπλέον, τα καρκινικά κύτταρα που έχουν υποστεί τις κατάλληλες γενετικές αλλαγές έχουν αποκτήσει την ιδιότητα να διεισδύουν σε παραπλήσιους υγιείς ιστούς ή και να αποσπώνται από τον κυρίως όγκο και μέσω της κυκλοφορίας του

αίματος να μεταφέρονται σε άλλους ιστούς, όπου συνεχίζοντας τον ανεξέλεγκτο πολλαπλασιασμό τους, δημιουργούν νέους όγκους (μετάσταση) (Hanahan & Weinberg, 2011).

1.2.1. Βλάβες στο DNA και μηχανισμοί επιδιόρθωσης

Καθημερινά υπολογίζεται ότι πραγματοποιούνται περίπου 70.000 **βλάβες στο γενετικό υλικό** ενός κυττάρου ενός ανθρώπου, γεγονός που οδηγεί σε έναν τεράστιο αριθμό βλαβών στο DNA σε ολόκληρο το σώμα ενός ανθρώπου, αν αναλογιστεί κανείς ότι αποτελείται από 10^{13} κύτταρα. Οι βλάβες αυτές, μπορεί να είναι αποτέλεσμα των φυσιολογικών λειτουργιών των κυττάρων, όπως οι βλάβες που πραγματοποιούνται κατά την αντιγραφή του DNA, αλλά και εξωγενών παραγόντων, όπως η ιονίζουσα ακτινοβολία των ακτινών γ και Χ και διάφορες χημικές ουσίες (χημειοθεραπευτικοί και αλκυλιωτικοί παράγοντες). Σε περίπτωση που δεν επιδιορθωθούν, οι βλάβες του γενετικού υλικού μπορεί να οδηγήσουν σε αλλαγές στο DNA ή και επιγενετικές αλλοιώσεις, που έχουν ως αποτέλεσμα την τροποποίηση ή τη διακοπή της λειτουργίας ή/και της έκφρασης του γονιδίου και, πιθανώς, τη συμβολή τους στην ανάπτυξη καρκίνου (Tubbs & Nussenzweig, 2017).

Για την αντιμετώπιση των βλαβών του DNA, σηματοδοτούνται ισχυρά βιολογικά **μονοπάτια απόκρισης σε βλάβες του DNA** (DNA Damage Response – DDR), τα οποία αναγνωρίζουν τη βλάβη και στη συνέχεια επιτρέπουν στους **μηχανισμούς επιδιόρθωσης του DNA** να την απομακρύνουν. Αν η βλάβη δεν μπορεί να επιδιορθωθεί, το κύτταρο οδηγείται σε απόπτωση. Υπάρχουν τουλάχιστον πέντε κύριοι μηχανισμοί επιδιόρθωσης βλαβών DNA - η **επιδιόρθωση με εκτομή βάσης (Base Excision Repair - BER),** η **επιδιόρθωση με εκτομή νουκλεοτιδίων (Nucleotide Excision Repair - NER),** η **επιδιόρθωση αναντιστοιχίας βάσεων (MisMatch Repair - MMR),** ο **ομόλογος ανασυνδυασμός (Homologous Recombination -HR)** και η **μη ομόλογη ένωση άκρων (Non Homologous End Joining - NHEJ)** – οι οποίοι είναι ενεργοί σε διάφορα στάδια του κυτταρικού κύκλου, επιτρέποντας στα κύτταρα να επιδιορθώσουν τη βλάβη του DNA (Εικόνα 2) (Helena *et al.*, 2018; Christmann *et al.*, 2003).



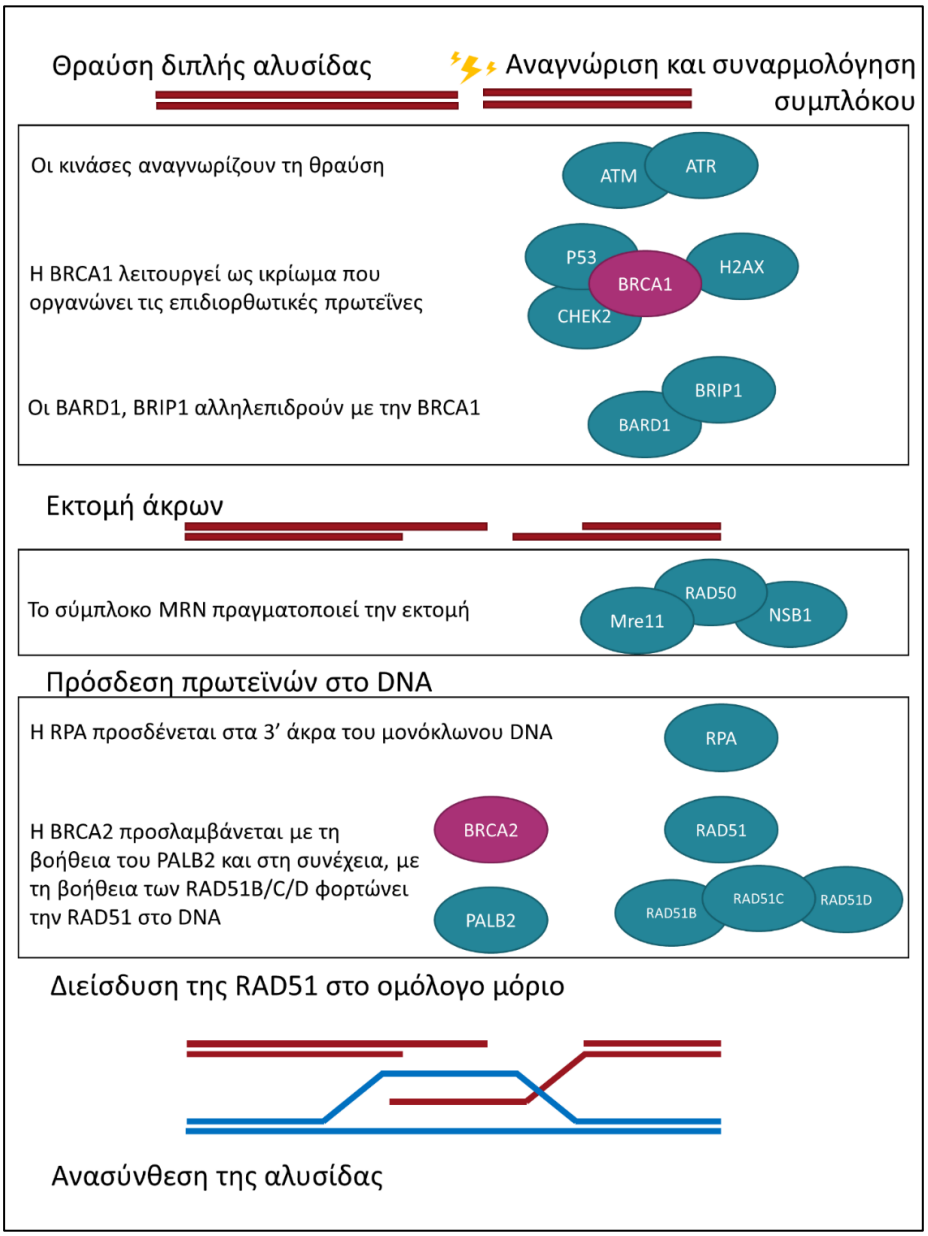
Εικόνα 2: Παράγοντες που προκαλούν βλάβη στο DNA και μηχανισμοί επιδιόρθωσης. Τροποποίηση από (Helena *et al.*, 2018).

Figure 2: DNA damaging factors and repair mechanisms. Modification from (Helena *et al.*, 2018).

1.2.2. Ομόλογος ανασυνδυασμός

Ένας από τους πιο σημαντικούς και ακριβείς μηχανισμούς επιδιόρθωσης του DNA είναι ο **ομόλογος ανασυνδυασμός (Homologous Recombination – HR)**. Ο HR είναι ενεργός κατά το στάδιο της αντιγραφής του γενετικού υλικού κι επιδιορθώνει τις θραύσεις της διπλής αλυσίδας του DNA, οι οποίες σε περίπτωση που δεν επιδιορθωθούν μπορεί να προκαλέσουν μεγάλες γονιδιωματικές αναδιατάξεις. Ο HR χρησιμοποιεί το ομόλογο χρωμόσωμα ή αδελφή χρωματίδα, η οποία λειτουργεί ως εκμαγείο για την επιδιόρθωση. Αρχικά, αναγνωρίζεται η περιοχή της βλάβης από τις κινάσες ATM και ATR και στη συνέχεια, η πρωτεΐνη BRCA1 λειτουργεί ως ικρίωμα το οποίο οργανώνει τις επιδιορθωτικές πρωτεΐνες. Στο επόμενο στάδιο, ο διπλός κλώνος του DNA διασπάται από το σύμπλοκο MRN, το οποίο αποτελείται από τις πρωτεΐνες Mre11, RAD50 και NSB1, ώστε να προκύψουν μονόκλωνα τμήματα DNA τα οποία προεξέχουν στο 3’-άκρο. Στη συνέχεια, η πρωτεΐνη RPA προσδένεται στα ελεύθερα 3’-άκρα ώστε να μπορέσει να φορτωθεί στο μονόκλωνο DNA η RAD51. Ταυτόχρονα, προσλαμβάνεται η BRCA2 με τη βοήθεια της PALB2, η οποία με τη σειρά της και με τη βοήθεια των πρωτεϊνών RAD51B, RAD51C και RAD51D φορτώνει την RAD51. Με την RAD51 προσδεμένη στο μονόκλωνο DNA, τα άκρα διεισδύουν στο ομόλογο μόριο, ενώ ακολουθεί η ανασύνθεση της αλυσίδας όπου εντοπίζεται η θραύση χρησιμοποιώντας ως πρότυπο την αδελφή χρωματίδα (Wright *et al.*, 2018; Walsh, 2015). Στην εικόνα 3,

παρουσιάζονται οι πρωτεΐνες που διαδραματίζουν τον κυριότερο ρόλο στον HR. Πολλά από τα γονίδια που κωδικοποιούν για αυτές τις πρωτεΐνες προδιαθέτουν σε μια σειρά από καρκινικά σύνδρομα.

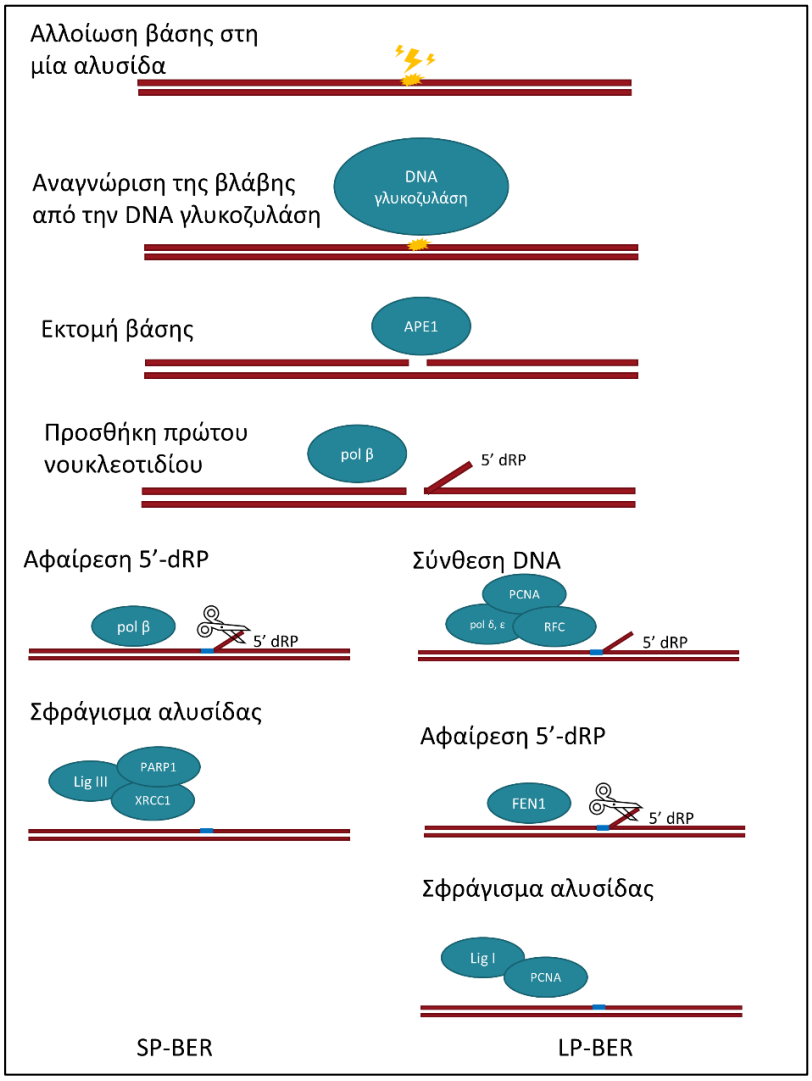


Εικόνα 3: Η διαδικασία επιδιόρθωσης βλαβών στο DNA με τον ομόλογο ανασυνδυασμό και τα μόρια που διαδραματίζουν καθοριστικό ρόλο. Αρχική εικόνα από (Walsh, 2015).

Figure 3: The process of DNA damage repair with homologous recombination and the molecules that play a key role. Original image by (Walsh, 2015)

1.2.3. Επιδιόρθωση με εκτομή βάσεων

Ο **μηχανισμός επιδιόρθωσης με εκτομή βάσεων (Base Excision Repair – BER)** είναι υπεύθυνος για την απομάκρυνση μικρών μεμονωμένων αλλοιώσεων από το γενετικό υλικό. Η αποτυχία απομάκρυνσης των κατεστραμμένων βάσεων μπορεί να προκαλέσει παραλλαγές στο DNA λόγω εσφαλμένης σύζευξης ή να οδηγήσει σε διάσπαση του DNA κατά την αντιγραφή του. Κατά το αρχικό στάδιο του μηχανισμού επιδιόρθωσης με εκτομή βάσεων, η γλυκοσυλάση του DNA αναγνωρίζει και αφαιρεί τις κατεστραμμένες βάσεις, δημιουργώντας μία θέση AP στο DNA, δηλαδή μία θέση η οποία δεν περιλαμβάνει ούτε πουρίνη ούτε πυριμιδίνη. Στη συνέχεια, η θέση AP υποβάλλεται σε περαιτέρω επεξεργασία από την ενδονουκλεάση AP1 (APE1), ώστε η DNA πολυμεράση β (Polβ) να μπορέσει να καλύψει το κενό στη μονή αλυσίδα του DNA.



Εικόνα 4: Η διαδικασία επιδιόρθωσης βλαβών στο DNA με εκτομή βάσεων και τα μόρια που διαδραματίζουν καθοριστικό ρόλο. Αρχική εικόνα από (Christmann *et al.*, 2003).

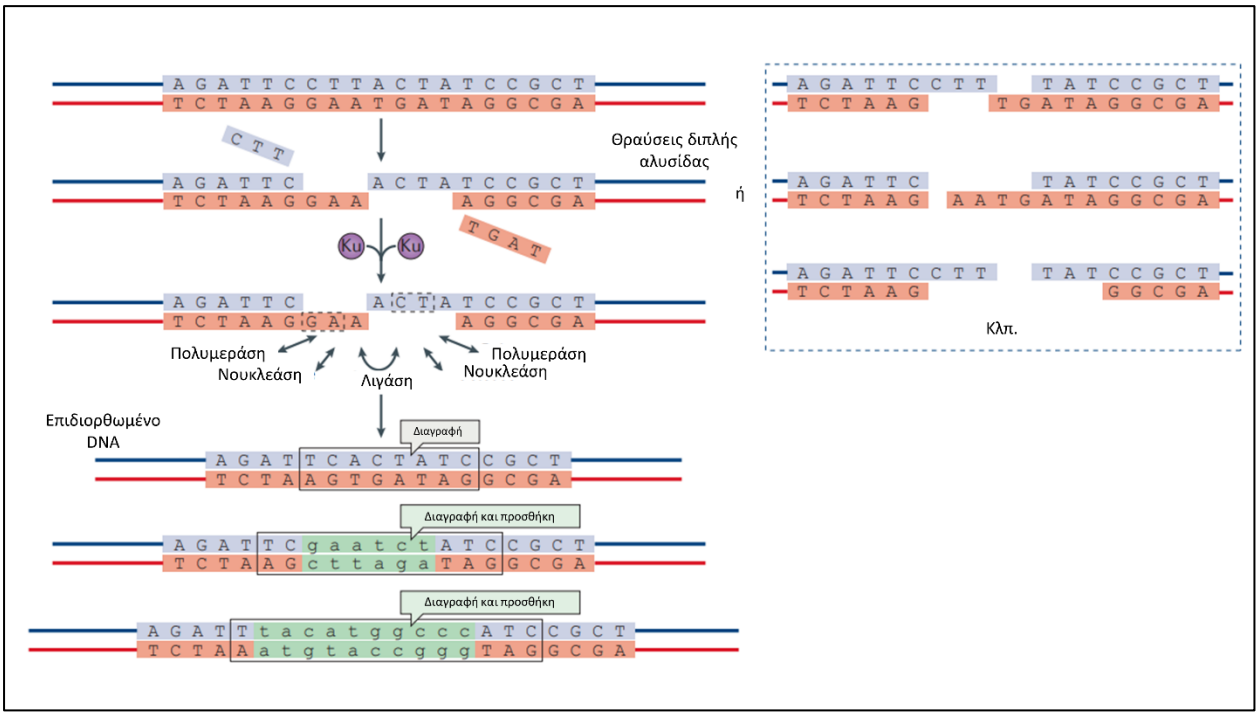
Figure 4: The process of DNA damage repair with base excision repair mechanism and the molecules that play a key role. Original image by (Christmann *et al.*, 2003).

Στο τελικό στάδιο, αναλόγως τον αριθμό των νουκλεοτιδίων που χρειάζεται να προστεθούν, ενεργοποιείται είτε ο BER βραχείας επιδιόρθωσης (για αντικατάσταση ενός νουκλεοτιδίου· short patch BER – SP-BER) είτε ο BER μακράς επιδιόρθωσης (όπου συντίθενται 2-13 νουκλεοτίδια· long patch BER – LP-BER) για την ολοκλήρωση της επιδιόρθωσης. Στον SP-BER, η πολυμεράση β, έχοντας παράλληλα και ενεργότητα λιγάσης, απομακρύνει τα 5'-2-δεοξυριβόζη-5-φωσφατάση (5'-dRP) ελεύθερα άκρα που προέκυψαν από τη δράση της APE1, και στη συνέχεια, η DNA λιγάση III, με τη συμμετοχή της πρωτεΐνης XRCC1, της πολυμεράσης β και της πολυμεράσης της πολυαδενοφωσφορικής ριβόζης 1 (polyADP-ribose polymerase-1, PARP1), σφραγίζει την αλυσίδα. Στον LP-BER, αρχικά συντίθεται η υπόλοιπη ακολουθία του DNA από την πολυμεράση δ ή ε, με τη συμμετοχή των πρωτεϊνών PCNA και RFC. Η απομάκρυνση του 5'-dRP που προεξέχει, πραγματοποιείται από την ενδονουκλεάση FEN1 (flap endonuclease 1) και στη συνέχεια ακολουθεί το σφράγισμα της αλυσίδας από την DNA λιγάση I και την πρωτεΐνη PCNA (Εικόνα 4) (Beard *et al.*, 2019; Christmann *et al.*, 2003).

1.2.4. Μη ομόλογη ένωση άκρων

Ο **μηχανισμός επιδιόρθωσης με μη ομόλογη ένωση άκρων (non-homologous end joining - NHEJ)** επισκευάζει τις θραύσεις διπλής αλυσίδας του DNA, χωρίς να απαιτεί την ύπαρξη της αδελφής χρωματίδας. Ο NHEJ μπορεί να λάβει χώρα σε όλη τη διάρκεια του κυτταρικού κύκλου και στρατολογείται όταν υπάρχει ανεπάρκεια HR. Καθώς δε χρησιμοποιεί το ομόλογο χρωμόσωμα για την επισκευή του DNA, ο NHEJ είναι πιο επιρρεπής σε λάθη.

Κατά το αρχικό στάδιο του μηχανισμού NHEJ, η θραύση διπλής αλυσίδας αναγνωρίζεται από το ετεροδιμερές σύμπλοκο Ku, το οποίο αποτελείται από τις πρωτεΐνες Ku70 και Ku80. Στη συνέχεια, το Ku προσδένεται στη βλάβη του DNA και λειτουργεί ως ικρίωμα το οποίο οργανώνει τα σύμπλοκα πολυμερασών, νουκλεασών και λιγάσης που συμμετέχουν στο μηχανισμό. Οι νουκλεάσες διασφαλίζουν ότι τα δύο άκρα είναι συμβατά, πραγματοποιώντας εκτομή μικρών περιοχών των 5' ή 3' προεξοχών ώστε να δημιουργήσουν μικρές περιοχές μικροομολογίας. Οι πολυμεράσες DNA που συμμετέχουν στο NHEJ για τη σύνθεση του DNA είναι οι Polμ και Polλ. Δεν είναι απολύτως σαφές πως λειτουργούν τα μόρια του συμπλόκου λιγάσης, αλλά η DNA λιγάση IV και η πρωτεΐνη XRCC4 είναι τα κεντρικά μόρια του συμπλόκου στο μηχανισμό NHEJ. Επιπλέον, στο σύμπλοκο συμμετέχουν οι πρωτεΐνες XLF (XRCC4-like factor) και PAXX (paralogue of XRCC4 and XLF), που αλληλεπιδρούν με την XRCC4 και το σύμπλοκο Ku, αντίστοιχα. Ανάλογα με τη μορφή των άκρων που προκύπτουν από τη θραύση της διπλής έλικας, τα ένζυμα που λαμβάνουν μέρος στο μηχανισμό NHEJ, μπορούν να δρουν στις βλάβες με οποιαδήποτε σειρά και πολλαπλές φορές για να προσθέσουν και να αφαιρέσουν νουκλεοτίδια (Εικόνα 5) (Chang *et al.*, 2017).



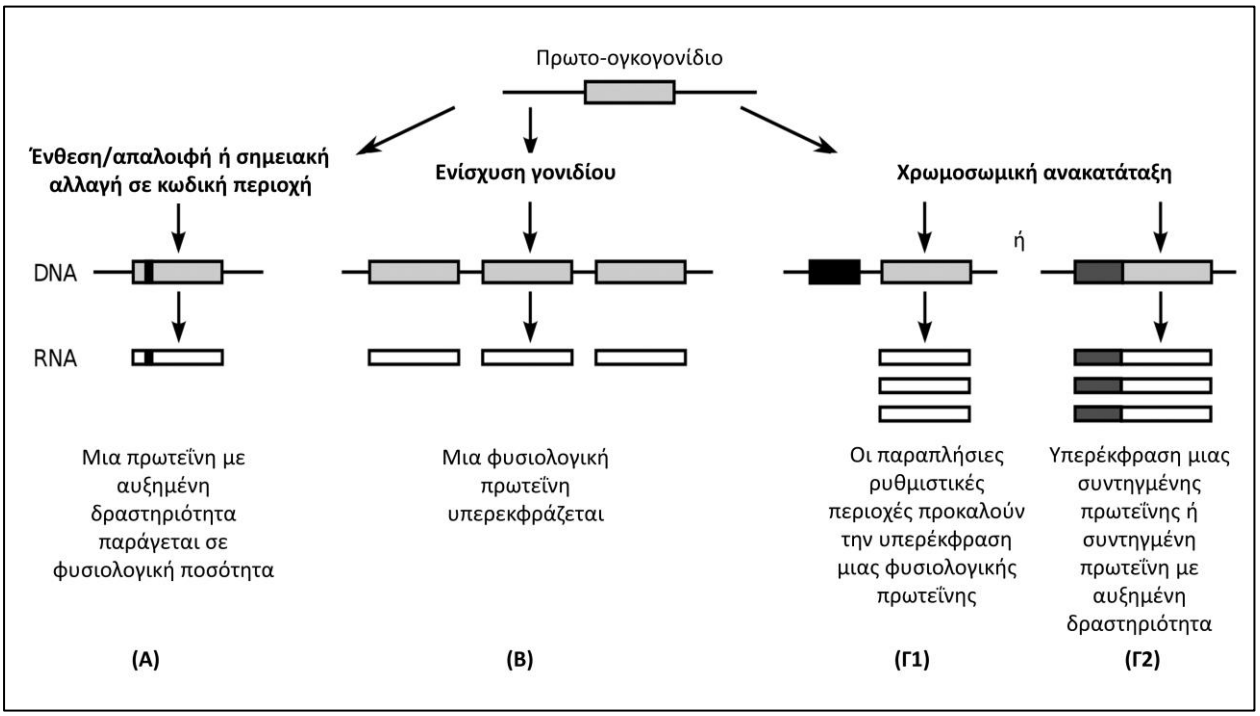
Εικόνα 5: Ο μηχανισμός επιδιόρθωσης θραύσεων διπλής αλυσίδας με μη ομόλογη ένωση άκρων. Τροποποίηση από (Chang *et al.*, 2017).

Figure 5: The double strand break repair mechanism with non-homologous end joining. Modification from (Chang *et al.*, 2017).

1.2.5. Ογκογονίδια και ογκοκατασταλτικά γονίδια

Έπειτα από πολλά χρόνια έρευνας, πλέον είμαστε σε θέση να γνωρίζουμε ότι ο καρκίνος είναι μία ασθένεια που οφείλεται σε δυναμικές αλλαγές του γονιδιώματος. Η ανακάλυψη γενετικών αλλαγών που προσδίδουν **κέρδος λειτουργίας (gain of function)** και **απώλεια λειτουργίας (loss of function)** σε **πρώτο-ογκογονίδια** και **ογκοκατασταλτικά γονίδια** αντίστοιχα, ήταν από τις πρώτες ενδείξεις που οδήγησαν σε αυτό το συμπέρασμα (Weinberg, 2013; Croce, 2008).

Μέχρι σήμερα, έχουν ανακαλυφθεί εκατοντάδες **ογκογονίδια**, γονίδια δηλαδή που συμβάλλουν στην καρκινογένεση και υπερεκφράζονται σε καρκινικούς ιστούς. Τα περισσότερα ογκογονίδια ξεκινούν ως πρωτο-ογκογονίδια. Τα πρωτο-ογκογονίδια είναι φυσιολογικά γονίδια που συνήθως μετέχουν στους μηχανισμούς της κυτταρικής ανάπτυξης και του κυτταρικού πολλαπλασιασμού ή/και στην αναστολή της κυτταρικής απόπτωσης. Υπό ορισμένες συνθήκες, έπειτα από τροποποίηση της αρχικής λειτουργίας τους, τα πρωτο-ογκογονίδια μετατρέπονται σε ογκογονίδια και συμβάλλουν στον ανεξέλεγκτο κυτταρικό πολλαπλασιασμό καθώς και στην αδυναμία κυτταρικής απόπτωσης (Hartl & Bister, 2013; Croce, 2008). Οι μηχανισμοί ενεργοποίησης των ογκογονιδίων είναι διάφοροι και ανάμεσά τους συμπεριλαμβάνονται οι εξής: α. η ύπαρξη μιας παθογόνου παραλλαγής σε ένα πρωτο-ογκογονίδιο ή εντός μιας ρυθμιστικής περιοχής του, β. η μετατόπιση τμήματος ή ολόκληρου του πρωτο-ογκογονιδίου σε διαφορετική θέση στο γονιδίωμα και γ. η ενίσχυση (amplification) του πρωτο-ογκογονιδίου (Εικόνα 6) (Jan & Chaudhry, 2019; Jang *et al.*, 2019; Abraham *et al.*, 2017; Barillot *et al.*, 2012).



Εικόνα 6: Μηχανισμοί ενεργοποίησης των ογκογονιδίων. Τροποποίηση από (Barillot *et al.*, 2012).

Figure 6: Oncogene activation mechanisms. Modification from (Barillot *et al.*, 2012).

Τα **ογκοκατασταλτικά γονίδια** κωδικοποιούν πρωτεΐνες που ρυθμίζουν την κυτταρική διαίρεση και αντιγραφή και χωρίζονται στα **γονίδια-φρουρούς (gatekeepers)** και τα **γονίδια-φροντιστές (caretakers)** (Bunz, 2016; Epstein, 2015; van Heemst *et al.*, 2007). Τα γονίδια-φρουροί ρυθμίζουν την κυτταρική ανάπτυξη, είτε αναστέλλοντας τον κυτταρικό κύκλο είτε προκαλώντας απόπτωση. Οι παθογόνοι παραλλαγές στα γονίδια-φρουρούς έχουν ως αποτέλεσμα την απώλεια λειτουργίας τους και κατά συνέπεια τον ανεξέλεγκτο πολλαπλασιασμό των κυττάρων και, τελικά, την καρκινογένεση. Τα γονίδια-φροντιστές φροντίζουν για τη διατήρηση και την προστασία της ακεραιότητας του γονιδιώματος. Η απώλεια λειτουργίας των γονιδίων-φροντιστών δεν έχει ως άμεσο αποτέλεσμα την καρκινογένεση. Στην πραγματικότητα, η απενεργοποίηση των γονιδίων-φροντιστών οδηγεί στη συσσώρευση νέων παραλλαγών συνολικά στο γονιδίωμα, γεγονός που αυξάνει την πιθανότητα της απώλειας λειτουργίας των ογκογονιδίων και γονιδίων-φρουρών (Clayton *et al.*, 2020).

Παρόλο που υπάρχουν αυτές οι δύο κατηγορίες των ογκοκατασταλτικών γονιδίων, ο διαχωρισμός τους δεν είναι απόλυτος. Ένα γονίδιο μπορεί να ανήκει και στην κατηγορία των γονιδίων-φρουρών και στην κατηγορία των γονιδίων-φροντιστών, ενώ η λειτουργία των ογκοκατασταλτικών γονιδίων δεν έχει κατανοηθεί συνολικά (Jeggo *et al.*, 2016; Pierron, 2015). Το γονίδιο *TP53* αποτελεί ένα χαρακτηριστικό παράδειγμα ενός πολύ σημαντικού γονιδίου που ανήκει και στις δύο κατηγορίες, χάρη στη συμμετοχή του στη ρύθμιση της διακοπής του κυτταρικού κύκλου ή/και της απόπτωσης, καθώς και στην έμμεση συμμετοχή του στην επιδιόρθωση του DNA (Mello & Attardi, 2018). Στον Πίνακα 1 παρουσιάζονται κάποια ογκογονίδια και ογκοκατασταλτικά γονίδια και η λειτουργία τους.

Πίνακας 1: Ογκογονίδια, ογκοκατασταλτικά γονίδια και η λειτουργία τους.

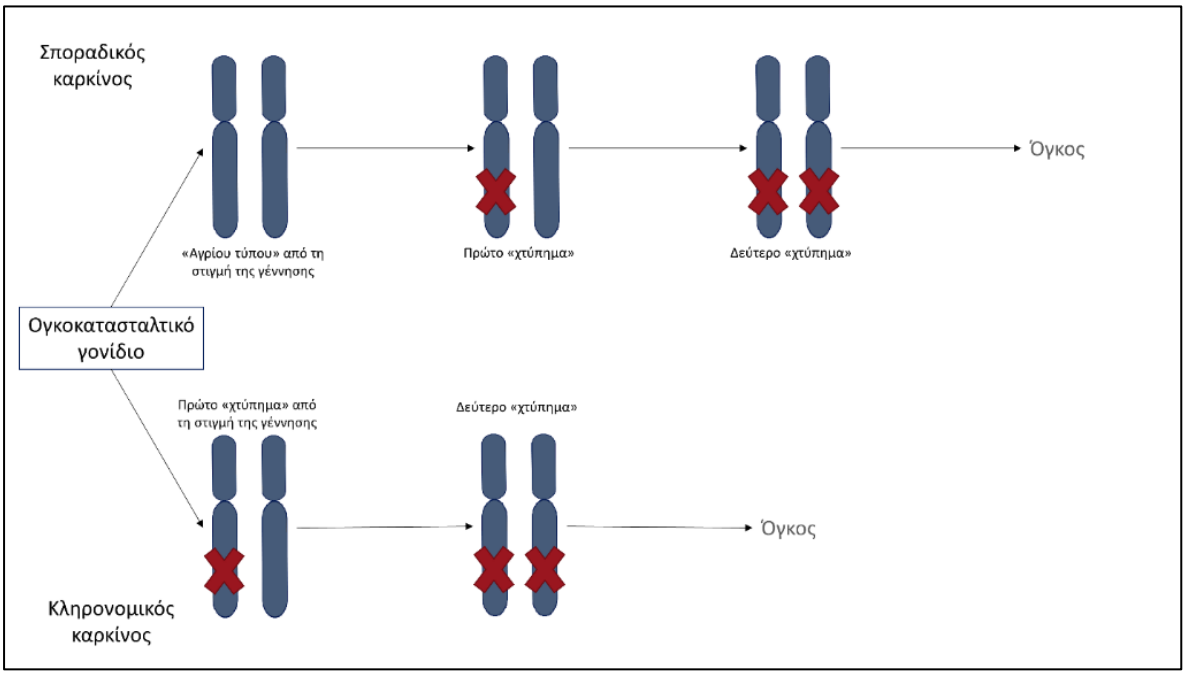
Table 1: Oncogenes, tumor suppressor genes and their function.

Ογκογονίδια		
Γονίδιο	**Θέση στο γονιδίωμα**	**Λειτουργία**
ABL1	9q34	Κυτταρική ανάπτυξη
AKT2	19q13	Κινάση σερίνης/θρεονίνης
ALK	2p23	Υποδοχέας κινάσης τυροσίνης
ALK/NPM	t(2;5)(p23:q35)	Γονίδιο σύντηξης
RUNX1 (AML1)	21q22	Μεταγραφικός παράγοντας
BCL-2	18q21	Καταστολή απόπτωσης
BCR/ABL	t(9;22)(q34;q11)	Γονίδιο σύντηξης
MYC (c-MYC)	8q24	Κυτταρικός πολλαπλασιασμός και σύνθεση του DNA
EGFR	7p11	Ενεργοποίηση κυτταρικής ανάπτυξης
ERBB2 (HER2/neu)	17q12	Ενεργοποίηση κυτταρικής ανάπτυξης
FGF4	11q13	Αυξητικός παράγοντας ινοβλαστών
KIT	4q12	Υποδοχέας κινάση τυροσίνης
MYCL	1p34	Μεταγραφικός παράγοντας
MYCN	2p24	Κυτταρικός πολλαπλασιασμός και σύνθεση του DNA
HRAS	11p15	Μεταγωγή σήματος
KRAS	12p12	Μεταγωγή σήματος
NRAS	1p13	Μεταγωγή σήματος
RET	10q11	Υποδοχέας κινάσης τυροσίνης
NOTCH1	9q34	Διαμεμβρανικός υποδοχέας μονής διέλευσης
NTRK1	1q23	Υποδοχέας κινάσης τυροσίνης
Ογκοκατασταλτικά Γονίδια		
Γονίδιο	**Θέση στο γονιδίωμα**	**Λειτουργία**
TP53	17p13	Ρύθμιση της επιδιόρθωσης του DNA, του κυτταρικού κύκλου, της απόπτωσης και της αγγειογένεσης
RB1	23q14	Αναστολέας του κυτταρικού κύκλου
PTEN	10q23	Διπλή ειδική φωσφατάση
BRCA1	17q21	Επιδιόρθωση του DNA
BRCA2	13q12	Επιδιόρθωση του DNA
ATM	11q22	Επιδιόρθωση του DNA
STK11	19p13	Κινάση σερίνης-θρεονίνης
CDKN1B	12p13	Αναστολέας του κυτταρικού κύκλου
CDKN2A	9p21	Αναστολέας του κυτταρικού κύκλου
SERPINB5	18q21	Αναστολέας σερίνης-προτεάσης
IGFII-R	6q26	Υποδοχέας αυξητικού παράγοντα
CDH1 (E-cadherin)	16q22	Μόριο κυτταρικής πρόσφυσης
RARß2	3p24	Υποδοχέας ρετινοϊκού οξέος
MLH1	3p21	Επιδιόρθωση του DNA
MSH2	2p22	Επιδιόρθωση του DNA
APC	5q21	Αναστολέας μεταγραφής της β-κατενίνης

MEN1	11q13	Αλληλεπίδραση με πρωτεΐνες επιδιόρθωσης DNA κ.ά..
NF1	17q11	Ενεργοποίηση GTPάσης
NF2	22q12	Οργάνωση της κυτταρικής μεμβράνης
VHL	3p25	Ρύθμιση κυτταρικού κύκλου
WRN	8p12	Επιδιόρθωση του DNA
WT1	11p13	Μεταγραφικός παράγοντας

1.2.6. Υπόθεση «δύο χτυπημάτων» κατά Knudson

Η **υπόθεση των «δύο χτυπημάτων»** διατυπώθηκε για πρώτη φορά το 1972 από τον Alfred Knudson. Σύμφωνα με αυτή τη θεωρία, για να προκληθεί η καρκινογένεση, χρειάζεται και τα δύο αλληλόμορφα ενός ογκοκατασταλτικού γονιδίου να έχουν απενεργοποιηθεί, είτε μέσω κάποιας παραλλαγής είτε μέσω επιγενετικής σίγασης. Επομένως, για ένα άτομο που έχει ήδη γεννηθεί με κάποια κληρονομούμενη παραλλαγή που έχει ως αποτέλεσμα την αποσιώπηση ενός εκ των δύο αλληλομόρφων ενός ογκοκατασταλτικού γονιδίου σε όλα του τα κύτταρα, η πιθανότητα να αποσιωπηθεί και το δεύτερο αλληλόμορφο, κι άρα το άτομο να εμφανίσει καρκίνο κάποια στιγμή στη ζωή του, είναι σημαντικά μεγαλύτερη. Σε αυτή την περίπτωση, η απενεργοποίηση και του δεύτερου αλληλομόρφου (‘2^{ο} χτύπημα’) οδηγεί σε **απώλεια της ετεροζυγωτίας (Loss of heterozygosity - LOH)** (Hino & Kobayashi, 2017; Berger *et al.*, 2011). Η υπόθεση των «δύο χτυπημάτων» συνέβαλε ουσιαστικά στην κατανόηση του κληρονομικού καρκίνου, καθώς παρείχε ένα ολοκληρωμένο μοντέλο για την περιγραφή της καρκινογένεσης σε άτομα που φέρουν παθογόνο παραλλαγή σε κάποιο γονίδιο προδιάθεσης (Εικόνα 7).

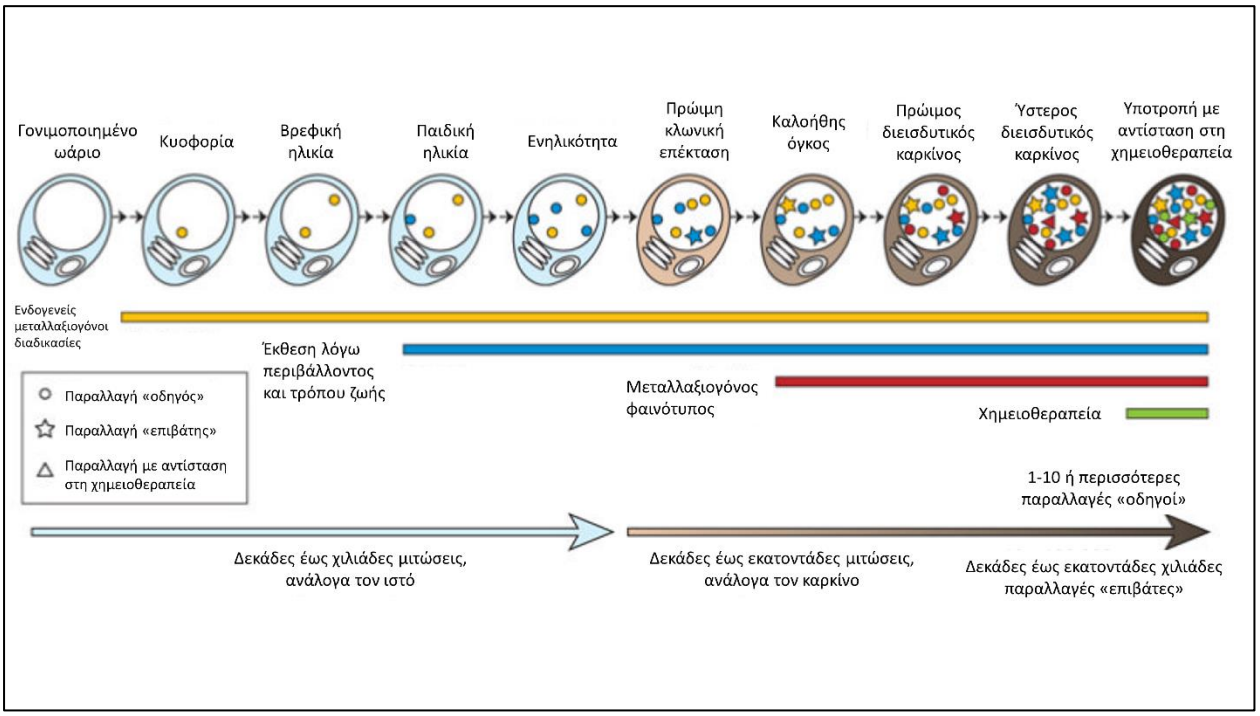


Εικόνα 7: Η υποθέση "δύο χτυπημάτων" κατά Knudson, για το σποραδικό και τον κληρονομικό καρκίνο.

Figure 7: Knudson's "two-hit" hypothesis for sporadic and hereditary cancer.

1.2.7. Το γονιδίωμα του όγκου

Ένα από τα κύρια χαρακτηριστικά του καρκίνου είναι η **γονιδιωματική αστάθεια** (Hanahan & Weinberg, 2011)**.** Όπως αναφέρθηκε ήδη, το DNA ενός ανθρώπου υπόκειται καθημερινά σε παράγοντες που προκαλούν βλάβες, οι οποίες επισκευάζονται μέσω των μηχανισμών επιδιόρθωσης. Κάποιες από αυτές τις βλάβες, δεν καταφέρνουν να επισκευαστούν και οδηγούν στην ανάπτυξη όγκου. Στα καρκινικά κύτταρα, οι βλάβες αυτές είναι δύσκολο να επιδιορθωθούν, κυρίως λόγω της απενεργοποίησης των γονιδίων που μετέχουν στους μηχανισμούς επιδιόρθωσης του DNA. Το αποτέλεσμα είναι η συσσώρευση των παραλλαγών και ο κατακερματισμός του γονιδιώματος του όγκου (Malarkey *et al.*, 2018).



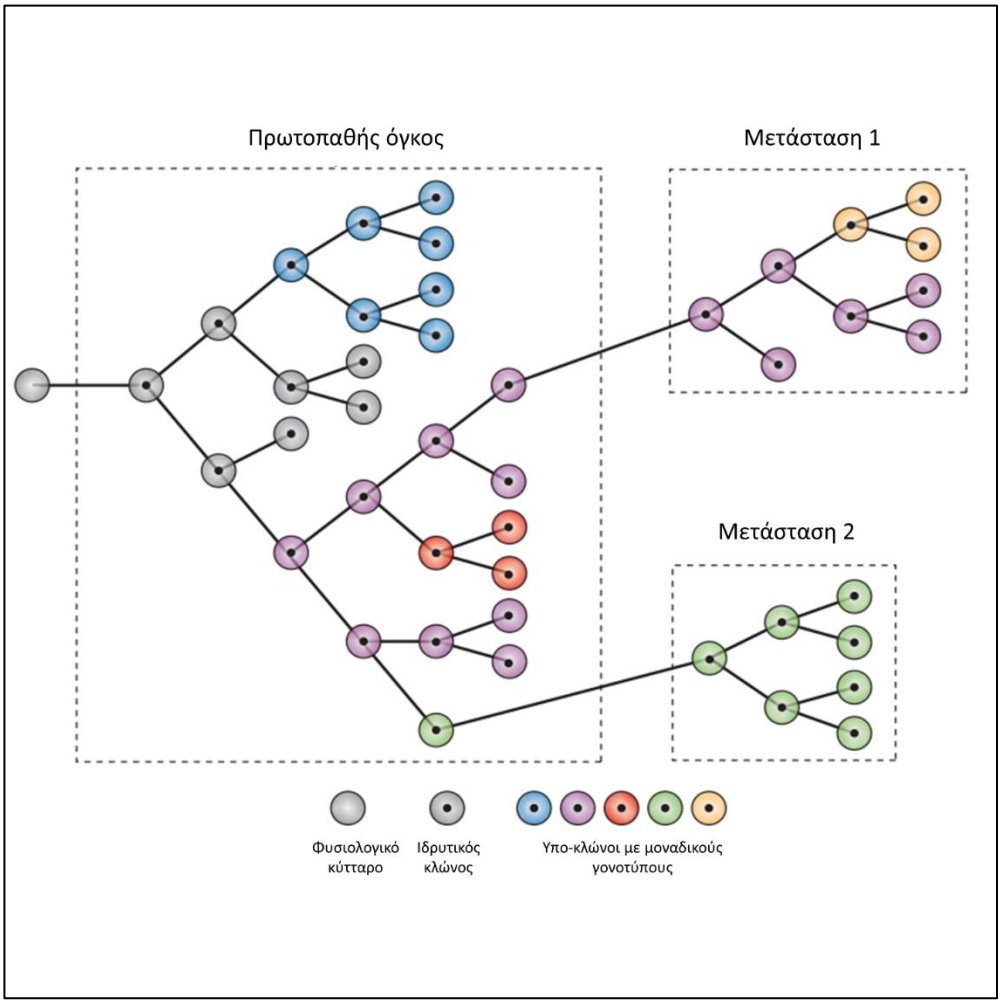
Εικόνα 8: Η γενεαλογία των μιτωτικών κυτταρικών διαιρέσεων από το γονιμοποιημένο ωάριο σε ένα μόνο κύτταρο μέσα σε έναν καρκίνο. Τροποποίηση από (Stratton *et al.*, 2009).

Figure 8: The genealogy of mitotic cell divisions from the fertilized egg to a single cell within a cancer. Modification from (Stratton *et al.*, 2009).

Κάθε φορά που ένα καρκινικό κύτταρο διαιρείται, συσσωρεύονται επιπλέον παραλλαγές. Ως αποτέλεσμα, οι σωματικές παραλλαγές που υπάρχουν σε ένα καρκινικό κύτταρο μπορούν να περιγραφούν ως μία καταγραφή όλων των εξελικτικών διαδικασιών που έχει υποστεί ο όγκος από την δημιουργία του πρώτου καρκινικού κυττάρου έως τη δημιουργία του υπό εξέταση κυττάρου (Εικόνα 8) (Navin *et al.*, 2011; Stratton *et al.*, 2009). Καθώς όμως το κάθε κύτταρο έχει τους δικούς του πολλαπλούς απογόνους, δημιουργούνται πληθυσμοί κυττάρων με ίδιο γονιδίωμα, που ονομάζονται **κλώνοι**. Κατά συνέπεια, ένας όγκος χαρακτηρίζεται και από **ετερογένεια**, δηλαδή ακόμα και εντός του ίδιου όγκου μπορεί να συνυπάρχουν πολλαπλοί κλώνοι, ενώ οι μεταστάσεις συνήθως έχουν εντελώς διαφορετικό

γονιδίωμα από τον αρχικό όγκο. Η διαδικασία αυτή προσδιορίζεται ως **κλωνική εξέλιξη του όγκου (tumor clonal evolution)** (Εικόνα 9) (Turajlic *et al.*, 2019; Caldas, 2012; Navin *et al.*, 2011).

Ως αποτέλεσμα της συσσώρευσης παραλλαγών, οι όγκοι αποκτούν ιδιότητες που τους επιτρέπουν να εξελίσσονται πιο γρήγορα και να επιβιώνουν, όπως η **ανθεκτικότητα στη θεραπεία** (Dagogo-Jack & Shaw, 2018). Παράλληλα, η ετερογένεια του καρκινικού γονιδιώματος, καθιστά τον προσδιορισμό της αλληλουχίας του δύσκολο .



Εικόνα 9: Η εξέλιξη του γονιδιώματος του όγκου. Τροποποίηση από (Caldas, 2012).

Figure 9: The evolution of the tumor genome. Modification from (Caldas, 2012).

1.3. Κληρονομικός καρκίνος

Ο καρκίνος κατηγοριοποιείται σε σποραδικό, οικογενή και κληρονομικό. Ο **σποραδικός καρκίνος** αποτελεί τη μέγιστη πλειοψηφία περιπτώσεων καρκίνου, καθώς υπολογίζεται ότι αφορά περίπου το 70%-80% των διαγνώσεων. Στον σποραδικό καρκίνο, οι παραλλαγές στις οποίες οφείλεται η ανάπτυξη του όγκου είναι επίκτητες κατά τη διάρκεια ζωής του ανθρώπου (σωματικές παραλλαγές) και όχι κληρονομούμενες. Ως **οικογενής καρκίνος** θεωρείται ο καρκίνος που φαίνεται να παρουσιάζεται σε

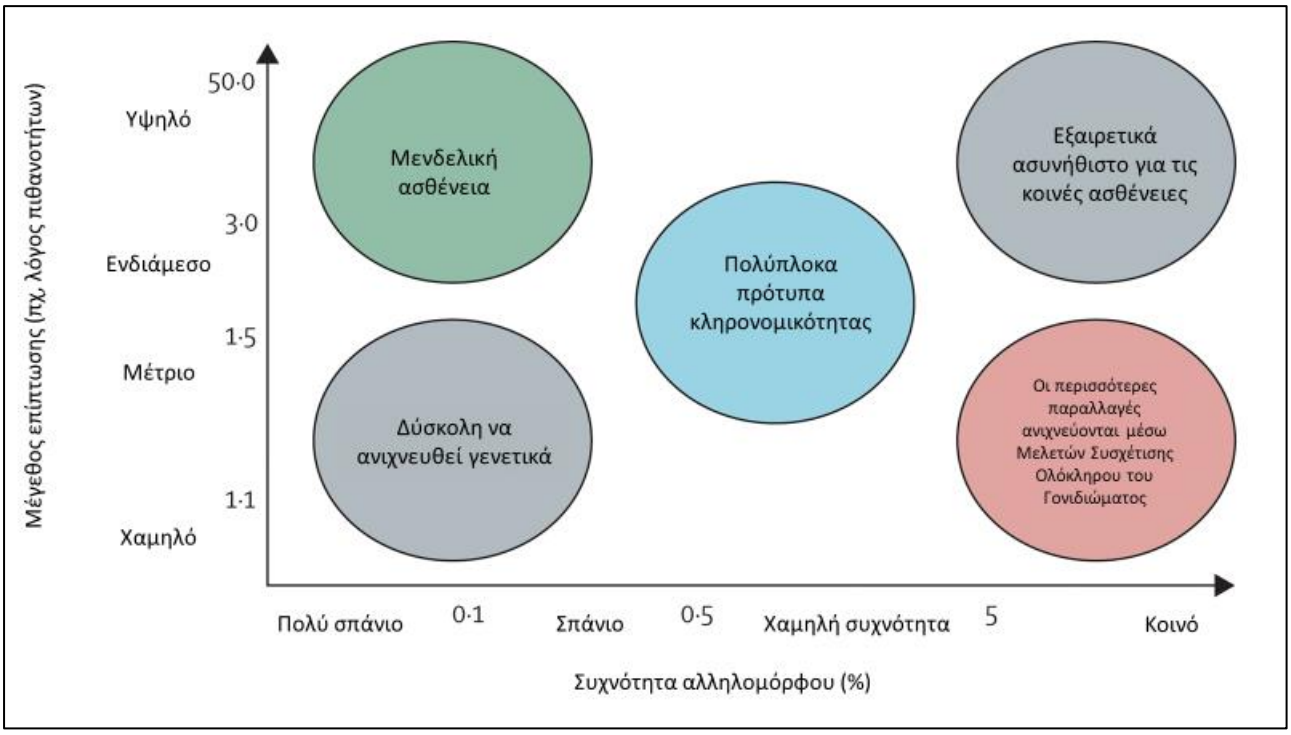
πολλά μέλη μιας οικογένειας, χωρίς ωστόσο να έχει ανιχνευθεί ένα σαφές γενετικό αίτιο. Σε αυτή την περίπτωση, ο υψηλότερος κίνδυνος εμφάνισης κακοήθειας σε μια οικογένεια σε σχέση με το γενικό πληθυσμό, μπορεί να οφείλεται σε κάποιο συνδυασμό πολλαπλών γενετικών τόπων προδιάθεσης, αλλά και σε περιβαλλοντικούς παράγοντες, όπως η διατροφή και η άσκηση. Παρόλο που δεν είναι εύκολο να ταυτοποιηθούν οι πραγματικοί παράγοντες κινδύνου, τα μέλη των οικογενειών με σοβαρό ιστορικό κακοήθειας ενδέχεται να χρειάζονται πιο στενή ιατρική παρακολούθηση (Rousset-Jablonski & Gompel, 2017; Bartsch *et al.*, 2016; Samadder *et al.*, 2015).

Ο **κληρονομικός καρκίνος** οφείλεται σε ένα σαφώς προσδιορισμένο γονιδιακό αίτιο και αφορά ένα μικρό ποσοστό όγκων, καθώς υπολογίζεται ότι μόλις το 5%-10% των περιστατικών καρκίνου οφείλονται σε κληρονομούμενες παραλλαγές (Mayer *et al.*, 2014; Apostolou & Fostira, 2013; Rahner & Steinke, 2008). Ωστόσο, το ποσοστό αυτό δεν είναι διόλου ευκαταφρόνητο, καθώς μεταφράζεται σε περίπου 1,4 εκατομμύρια νέες διαγνώσεις καρκίνου ανά έτος. Συνολικά, ο κληρονομικός καρκίνος επηρεάζει περί τα 300 εκατομμύρια άτομα παγκοσμίως. Επιπλέον, ένα ιδιαίτερο χαρακτηριστικό του κληρονομικού καρκίνου, το οποίο καθιστά την έγκαιρη διάγνωσή του απαραίτητη, είναι η ανάπτυξη κακοηθειών σε άτομα νεαρής ηλικίας, καθώς και τα πολλαπλά περιστατικά σχετιζόμενων κακοηθειών τόσο στους ίδιους τους ασθενείς όσο και σε μέλη της οικογένειάς τους (ACOG *et al.*, 2017; Apostolou *et al.*, 2015).

1.3.1. Συσχέτιση συχνότητας αλληλομόρφου και προδιάθεσης σε γενετικές ασθένειες

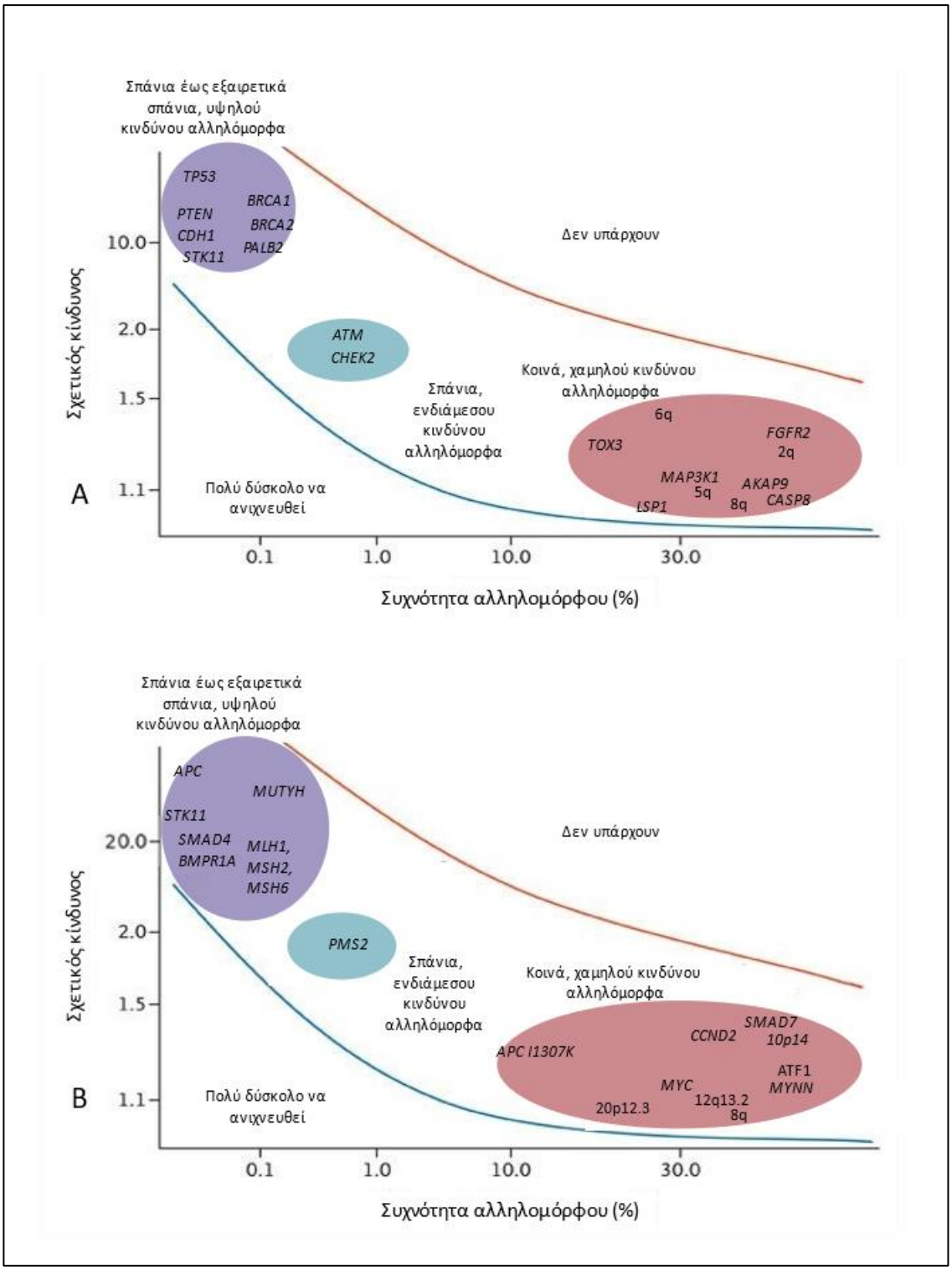
Ο κληρονομικός καρκίνος είναι **Μενδελική** ή **μονογονιδιακή ασθένεια**· η προδιάθεση σε κάποιο καρκινικό σύνδρομο οφείλεται σε γαμετικές παθογόνους παραλλαγές σε ένα γονίδιο. Τα αλληλόμορφα που προδιαθέτουν σε κάποια Μενδελική ασθένεια, συνήθως είναι **υψηλής** ή **ενδιάμεσης διεισδυτικότητας** (Speicher *et al.*, 2010). Η έννοια της διεισδυτικότητας αφορά στον κίνδυνο που το αλληλόμορφο επιφέρει για την εμφάνιση κάποιας νόσου, και στην περίπτωση του κληρονομικού καρκίνου, για την ανάπτυξη κακοήθειας. Ωστόσο, υπάρχουν κι άλλα πρότυπα κληρονόμησης, εκτός από το μοντέλο της μονογονιδιακής νόσου.

Τα τελευταία χρόνια, και έπειτα από τις ραγδαίες εξελίξεις στον τομέα της βιοτεχνολογίας, η ερευνητική κοινότητα είχε τη δυνατότητα να ταυτοποιήσει γενετικούς τόπους, οι οποίοι συνήθως βρίσκονται σε διαγονιδιακές περιοχές, παραλλαγές στους οποίους ανεβάζουν τον κίνδυνο εμφάνισης νόσου απειροελάχιστα. Για παράδειγμα, μια γαμετική παραλλαγή σε έναν τέτοιο γενετικό τόπο ενδέχεται να προσδίδει κίνδυνο στο άτομο που τη φέρει έως 1,5 φορά σε σχέση με το γενικό πληθυσμό. Ωστόσο, οι παραλλαγές αυτές λειτουργούν πολλαπλασιαστικά, δηλαδή όσες περισσότερες παραλλαγές φέρει κάποιος άνθρωπος, τόσο υψηλότερος ο κίνδυνος να νοσήσει. Οι γενετικοί αυτοί τόποι ταυτοποιούνται μέσα από πολύ μεγάλες πληθυσμιακές μελέτες, που ονομάζονται **Μελέτες Συσχέτισης Ολόκληρου του Γονιδιώματος (Genome Wide Association Studies – GWAS)** (Sud *et al.*, 2017). Μέχρι σήμερα, έχουν ανιχνευθεί τουλάχιστον 313 τέτοιοι γενετικοί τόποι, παραλλαγές στους οποίους προσδίδουν υψηλό κίνδυνο εμφάνισης καρκίνου του μαστού, ενώ υπολογίζεται πως ο διά βίου κίνδυνος ανάπτυξης όγκου για μία γυναίκα που φέρει όλες τις παραλλαγές μπορεί να φτάνει ακόμη και το 35% (Mavaddat *et al.*, 2019).



Εικόνα 10: Σχέση συχνότητας αλληλομόρφου και επίπτωσης. Τροποποίηση από (Speicher *et al.*, 2010).

Figure 10: Correlation between allele frequency and impact. Modification from (Speicher *et al.*, 2010).



Εικόνα 11: Σχέση συχνότητας αλληλομόρφου και κινδύνου εμφάνισης καρκίνου α. του μαστού και β. του παχέος εντέρου. Πληροφορίες από (Peters *et al.*, 2015; Hindorff *et al.*, 2011; Foulkes, 2008).

Figure 11: Correlation between allele frequency and risk of a. breast and b. colorectal cancer. Information from (Peters *et al.*, 2015; Hindorff *et al.*, 2011; Foulkes, 2008).

Ένα αλληλόμορφο το οποίο προσδίδει υψηλό κίνδυνο εμφάνισης μιας νόσου είναι πάρα πολύ σπάνιο στον γενικό πληθυσμό. Μάλιστα, όσο πιο σοβαρός ο φαινότυπος στον οποίο προδιαθέτει το αλληλόμορφο, τόσο πιο σπάνιο είναι. Έτσι, τα αλληλόμορφα υψηλής διεισδυτικότητας αναμένεται να βρίσκονται σε λιγότερο από το 0,5% του γενικού πληθυσμού, ενώ τα αλληλόμορφα χαμηλής διεισδυτικότητας, τα οποία ανιχνεύονται σε μελέτες συσχέτισης ολόκληρου του γονιδιώματος, αναμένεται να έχουν συχνότητα μεγαλύτερη του 5% στον γενικό πληθυσμό (Εικόνα 10) (Speicher *et al.*, 2010).

Όσον αφορά συγκεκριμένα τον κληρονομικό καρκίνο, τα γονίδια *BRCA1*, *BRCA2*, *TP53*, *CDH1*, *PTEN*, *STK11* θεωρούνται γονίδια υψηλής διεισδυτικότητας με παθογόνους παραλλαγές σε αυτά να ανιχνεύονται σπάνια στο γενικό πληθυσμό (Foulkes, 2008). Τα γονίδια αυτά προδιαθέτουν σε μια σειρά από σπάνια καρκινικά σύνδρομα, όπως το **σύνδρομο του καρκίνου μαστού/ωοθηκών (Hereditary Breast and Ovarian Cancer syndrome – HBOC),** το σύνδρομο **Li-Fraumeni,** το σύνδρομο **γαστρικού καρκίνου διάχυτου τύπου,** το σύνδρομο **Cowden** και το σύνδρομο **Peutz-Jeghers.** Πρόσφατα αποδείχθηκε ότι και το *PALB2* είναι ένα γονίδιο υψηλής διεισδυτικότητας που προσδίδει κίνδυνο ανάπτυξης καρκίνου του μαστού εφάμιλλο με αυτόν που προσδίδει το *BRCA2* (A. C. Antoniou *et al.*, 2014). Σημαντικό ρόλο στον κληρονομικό καρκίνο του μαστού διαδραματίζουν και τα γονίδια *ATM* και *CHEK2,* τα οποία θεωρούνται γονίδια ενδιάμεσης διεισδυτικότητας (Εικόνα 11) (Foulkes, 2008), ενώ τα γονίδια *RAD51C* και *RAD51D* πρόσφατα χαρακτηρίστηκαν ως γονίδια ενδιάμεσης διεισδυτικότητας για τον καρκίνο των ωοθηκών (Fostira *et al.*, 2020; Suszynska *et al.*, 2020; Castera *et al.*, 2018; Konstanta *et al.*, 2018). Από τα γονίδια που συμμετέχουν στο μηχανισμό επιδιόρθωσης του DNA με αταίριαστες βάσεις, τα οποία προδιαθέτουν στο **σύνδρομο Lynch**, το οποίο περιλαμβάνει μια σειρά κακοηθειών, όπως τον καρκίνο του παχέος εντέρου, του ενδομητρίου, των ωοθηκών, του λεπτού εντέρου κ.ά.., τα *MSH2*, *MSH6* και *MLH1* θεωρούνται γονίδια υψηλής διεισδυτικότητας, ενώ το *PMS2* θεωρείται γονίδιο ενδιάμεσης διεισδυτικότητας (Cohen *et al.*, 2019; Peters *et al.*, 2015; Hindorff *et al.*, 2011). Στην επόμενη ενότητα παρουσιάζονται κάποια γονίδια που εμπλέκονται στον κληρονομικό καρκίνο και τα σύνδρομα στα οποία προδιαθέτουν.

1.3.2. Γονίδια που εμπλέκονται σε κληρονομικά καρκινικά σύνδρομα

Ένα από τα πιο καλά μελετημένα καρκινικά σύνδρομα είναι το σύνδρομο HBOC. Η πρώτη οικογένεια με σύνδρομο HBOC περιεγράφηκε πρώτη φορά το 1866 από τον ιατρό Paul Broca (van der Groep *et al.*, 2011)· η οικογένεια, μάλιστα, ήταν αυτή της συζύγου του, η οποία είχε νοσήσει και η ίδια με καρκίνο του μαστού σε πολύ νεαρή ηλικία. Περισσότερα από εκατό χρόνια αργότερα, στις αρχές της δεκαετίας του 1970, η Mary-Claire King ξεκίνησε την αναζήτηση του γενετικού τόπου που ήταν υπεύθυνος για την εμφάνιση του συνδρόμου. Το γονίδιο *BRCA1*, το πρώτο γονίδιο που συσχετίστηκε με υψηλό κίνδυνο εμφάνισης καρκίνου μαστού/ωοθηκών, τελικά κλωνοποιήθηκε το 1994 (Miki *et al.*, 1994), ενώ λίγους μήνες αργότερα ανακαλύφθηκε κι ένα δεύτερο γονίδιο, το *BRCA2* (Wooster *et al.*, 1995).

Το σύνδρομο HBOC κληρονομείται με αυτοσωμικό επικρατή τρόπο, ενώ χαρακτηρίζεται από τη νεαρή ηλικία διάγνωσης των ασθενών και το βεβαρημένο προσωπικό ή/και οικογενειακό ιστορικό κακοήθειας. Τα **γονίδια *BRCA1* και *BRCA2***, τα οποία ευθύνονται για την εμφάνιση του συνδρόμου,

συμμετέχουν, και μάλιστα έχουν πρωταγωνιστικό ρόλο, στο μονοπάτι του HR, του μηχανισμού επιδιόρθωσης των θραύσεων στη διπλή αλυσίδα του DNA. Ο κίνδυνος ανάπτυξης καρκίνου του μαστού μίας γυναίκας που φέρει κάποια παθογόνο παραλλαγή στο γονίδιο *BRCA1* είναι 55%–72%, ενώ για το *BRCA2* κυμαίνεται ανάμεσα στο 45%–69% μέχρι την ηλικία των 80 ετών (Kuchenbaecker *et al.*, 2017; Chen & Parmigiani, 2007; A. Antoniou *et al.*, 2003), έναντι του 13% του γενικού πληθυσμού. Οι αντίστοιχοι κίνδυνοι για την κακοήθεια των ωοθηκών είναι 39%-44% για τις γυναίκες με παθογόνο παραλλαγή στο *BRCA1* και 11%-17% για τις γυναίκες με κάποια παθογόνο παραλλαγή στο γονίδιο *BRCA2* (Kuchenbaecker *et al.*, 2017; Chen & Parmigiani, 2007; A. Antoniou *et al.*, 2003), έναντι του 1,2% των γυναικών που θα αναπτύξουν καρκίνο των ωοθηκών στο γενικό πληθυσμό. Εκτός από τις κακοήθειες μαστού και ωοθηκών, τα γονίδια *BRCA1* και *BRCA2* προδιαθέτουν και σε μια σειρά από άλλες κακοήθειες, όπως κακοήθεια του παγκρέατος και καρκίνο του προστάτη, ενώ το γονίδιο *BRCA2* έχει συσχετισθεί με υψηλό ποσοστό εμφάνισης ανδρικού καρκίνου του μαστού (περίπου 6%, ενώ αποτελεί μόλις το 0.5%-1% όλων των διαγνώσεων καρκίνων μαστού) (Fostira, Saloustros*, et al.*, 2018; Abdelwahab Yousef, 2017).

Το **γονίδιο *PALB2*** θεωρείται το τρίτο σε σειρά γονίδιο επικινδυνότητας για εμφάνιση καρκίνου του μαστού (A. C. Antoniou *et al.*, 2014), ενώ τα μέχρι στιγμής δεδομένα δε φαίνεται να το συσχετίζουν με υψηλό κίνδυνο εμφάνισης καρκίνου των ωοθηκών. Το γονίδιο *PALB2* έχει επίσης καθοριστικό ρόλο στην επιδιόρθωση των θραύσεων της διπλής έλικας του DNA μέσω του HR, λειτουργώντας ως μοριακό ικρίωμα για τα γονίδια *BRCA1* και *BRCA2* ώστε να δημιουργήσουν ένα πρωτεϊνικό σύμπλοκο που είναι απαραίτητο για τον HR. Οι γυναίκες που φέρουν κάποια παθογόνο παραλλαγή στο γονίδιο *PALB2,* φαίνεται να έχουν κίνδυνο εμφάνισης από 33%-58% έως την ηλικία των 70 ετών, γεγονός που το κατατάσσει στα γονίδια υψηλής διεισδυτικότητας (X. Yang *et al.*, 2020; A. C. Antoniou *et al.*, 2014). Επιπλέον, υπάρχουν ενδείξεις ότι οι παθογόνοι παραλλαγές στο γονίδιο *PALB2* συσχετίζονται με υψηλότερο κίνδυνο εμφάνισης καρκίνου του παγκρέατος (Borecka *et al.*, 2016; Zhen *et al.*, 2015) και κακοήθειας του στομάχου διάχυτου τύπου (Fewings *et al.*, 2018).

Οι γαμετικές παθογόνοι παραλλαγές στα **γονίδια *ATM* και *CHEK2*** προδιαθέτουν σε καρκίνο του μαστού, ενώ η συσχέτισή τους με τον καρκίνο των ωοθηκών είναι ακόμα υπό διερεύνηση (Daly *et al.*, 2020). Ο διά βίου κίνδυνος ανάπτυξης κακοήθειας μαστού για τα άτομα με γαμετικές παθογόνους παραλλαγές στα δύο αυτά γονίδια ανέρχεται στο 25%-30% (Cragun *et al.*, 2020; Easton *et al.*, 2015), γεγονός που τα κατατάσσει στα γονίδια ενδιάμεσης διεισδυτικότητας. Το γονίδιο *ATM* εμπλέκεται σε μονοπάτια επιδιόρθωσης βλαβών του DNA που περιλαμβάνουν τόσο τον HR όσο και το μηχανισμό με σύνδεση μη ομόλογων άκρων. Το γονίδιο *CHEK2* διαδραματίζει επίσης σημαντικό ρόλο στην επιδιόρθωση του DNA μέσω HR αλλά και τη ρύθμιση του κυτταρικού κύκλου σε περίπτωση βλάβης του DNA.

Το **γονίδιο *TP53*,** αποτελεί ένα από τα πιο σημαντικά γονίδια στον καρκίνο, καθώς διαδραματίζει καθοριστικό ρόλο στη ρύθμιση και την εξέλιξη του κυτταρικού κύκλου, της απόπτωσης και της γονιδιωματικής σταθερότητας. Χάρη στο σημαντικό του ρόλο, το *TP53* είναι γνωστό και ως «Φύλακας του γονιδιώματος». Οι γαμετικές παθογόνοι παραλλαγές στο γονίδιο *TP53* προδιαθέτουν στο σπάνιο **γενετικό σύνδρομο Li-Fraumeni**, το οποίο περιλαμβάνει μια σειρά κακοηθειών, τόσο σε ενήλικες όσο και σε παιδιά, όπως λευχαιμίες, σαρκώματα, καρκίνο μαστού και καρκινώματα των επινεφριδίων. Το

σύνδρομο Li-Fraumeni κληρονομείται με αυτοσωμικό επικρατή τρόπο. Τα άτομα που φέρουν παθογόνους παραλλαγές στο *TP53* έχουν έως 100% διά βίου κίνδυνο εμφάνισης καρκίνου (Guha & Malkin, 2017), ενώ περίπου το 70% των γυναικών συγκεκριμένα που φέρουν παθογόνους παραλλαγές στο *TP53* θα εμφανίσει καρκίνο του μαστού σε ηλικία μικρότερη των 45 ετών (Mai *et al.*, 2016). Στους ασθενείς με σύνδρομο Li-Fraumeni δε συνίσταται η ακτινοθεραπεία λόγω αυξημένου κινδύνου ανάπτυξης δεύτερου πρωτοπαθούς όγκου (McBride *et al.*, 2014).

Το **γονίδιο *PTEN*** κωδικοποιεί μια φωσφατάση η οποία εμπλέκεται στη ρύθμιση του κυτταρικού κύκλου, εμποδίζοντας την γρήγορη ανάπτυξη και διαίρεση των κυττάρων (Lee *et al.*, 2018). Οι γαμετικές παθογόνοι παραλλαγές σε αυτό το γονίδιο ευθύνονται για την εμφάνιση του **συνδρόμου Cowden**, το οποίο κληρονομείται με αυτοσωμικό επικρατή τρόπο και χαρακτηρίζεται από την εμφάνιση αμαρτωμάτων και από υψηλό κίνδυνο εμφάνισης καλοήθων και κακοήθων όγκων του θυρεοειδούς, του μαστού και του ενδομητρίου. Τα άτομα που φέρουν παθογόνους παραλλαγές στο γονίδιο *PTEN* έχουν έως 35% διά βίου κίνδυνο ανάπτυξης καρκίνου του θυρεοειδούς και των νεφρών, ενώ οι γυναίκες με σύνδρομο Cowden έχουν έως 85% διά βίου κίνδυνο ανάπτυξης καρκίνου του μαστού και 35% διά βίου κίνδυνο ανάπτυξης καρκίνου του ενδομητρίου (Pilarski *et al.*, 2013).

Το ογκοκατασταλτικό **γονίδιο *CDH1***, που κωδικοποιεί την πρωτεΐνη Ε-καδχερίνη, προδιαθέτει στο **σύνδρομο κληρονομικού γαστρικού καρκίνου διάχυτου τύπου (Hereditary Diffuse Gastric Cancer-HDGC)**. Ο δια βίου κίνδυνος εμφάνισης γαστρικού καρκίνου σε άτομα που φέρουν γαμετικές παθογόνους παραλλαγές στο *CDH1* είναι 70% για τους άνδρες και 56% για τις γυναίκες. Επιπλέον, οι γυναίκες που φέρουν γαμετικές παθογόνους παραλλαγές στο *CDH1* έχουν δια βίου κίνδυνο ανάπτυξης καρκίνου μαστού λοβιακού τύπου περίπου 42% (Hansford *et al.*, 2015).

Οι γαμετικές παθογόνοι παραλλαγές στο **γονίδιο *STK11*** προδιαθέτουν στο **σύνδρομο Peutz-Jeghers**, το οποίο αποτελεί μια σπάνια πρώιμης έναρξης διαταραχή με αυτοσωμική επικρατούσα κληρονομικότητα. Το σύνδρομο Peutz-Jeghers περιλαμβάνει δύο βασικά χαρακτηριστικά, βάσει των οποίων, στις περισσότερες περιπτώσεις, πραγματοποιείται η κλινική διάγνωση του συνδρόμου· την ανάπτυξη πολλαπλών αμαρτωματωδών γαστρεντερικών πολυπόδων και ένα διακριτό πρότυπο στιγμάτων με εναπόθεση μελανίνης (κηλίδες χρώματος ανοικτού καφέ) στα χείλη, πέριξ και εντός της στοματικής κοιλότητας και μερικές φορές παρακείμενα των ματιών, των ρουθουνιών, πέριξ του ορθού, στα χέρια και στα πόδια. Τα άτομα που φέρουν γαμετικές παθογόνους παραλλαγές στο *STK11* έχουν διά βίου κίνδυνο ανάπτυξης γαστρεντερικού καρκίνου από 38%-66%, κίνδυνο για καρκίνο του παγκρέατος από 11%-36%, ενώ οι γυναίκες έχουν διά βίου κίνδυνο ανάπτυξης καρκίνου του μαστού και των ωοθηκών από 32%-54% και από 9%-21%, αντίστοιχα. Ο τελευταίος είναι αρκετά διακριτός και συχνά περιλαμβάνει την κακοήθεια στα κύτταρα Sertoli (Daniell *et al.*, 2018; Fostira, Mollaki*, et al.*, 2018; van Lier *et al.*, 2010).

Τα **γονίδια *RAD51C* και *RAD51D*** ανήκουν στην οικογένεια γονιδίων *RAD51.* Λόγω του σημαντικού ρόλου των RAD51 πρωτεϊνών στον HR, τα γονίδια της οικογένειας έχουν μελετηθεί εκτενώς τα τελευταία χρόνια σχετικά με το ρόλο τους στον κληρονομικό καρκίνο. Τα πιο πρόσφατα δεδομένα δείχνουν πως οι γυναίκες που φέρουν παθογόνο παραλλαγή στο *RAD51C* ή στο *RAD51D*, έχουν υψηλότερο κίνδυνο ανάπτυξης καρκίνου των ωοθηκών σε σχέση με το γενικό πληθυσμό, ενώ η σχέση τους με τον κίνδυνο

ανάπτυξης κακοήθειας του μαστού βρίσκεται ακόμα υπό διερεύνηση (Fostira *et al.*, 2020; Suszynska *et al.*, 2020; Castera *et al.*, 2018; Konstanta *et al.*, 2018).

Τα **γονίδια *MSH2, MSH6, MLH1 & PMS2***, τα οποία συμμετέχουν στο μονοπάτι επιδιόρθωσης αταίριαστων ζευγών βάσεων DNA, προδιαθέτουν στο **σύνδρομο Lynch**, το οποίο περιλαμβάνει μια σειρά κακοηθειών, όπως τον καρκίνο του παχέος εντέρου, του ενδομητρίου, των ωοθηκών, του λεπτού εντέρου κ.ά. και κληρονομείται με αυτοσωμικό επικρατή τρόπο. Οι ασθενείς με σύνδρομο Lynch έχουν διά βίου κίνδυνο ανάπτυξης κακοηθειών παχέος εντέρου και ενδομητρίου 90% και 40% αντίστοιχα. Στο σύνδρομο Lynch προδιαθέτει έμμεσα και μία παθογόνος παραλλαγή στο γονίδιο *EPCAM* -και συγκεκριμένα, η διαγραφή τμήματος του 3' άκρου του γονιδίου- η οποία προκαλεί επιγενετική απενεργοποίηση του γονιδίου *MSH2* μέσω υπερμεθυλίωσης της περιοχής του υποκινητή του (Cohen *et al.*, 2019).

Το γονίδιο *APC* είναι ένα ογκοκατασταλτικό γονίδιο του οποίου η πρωτεΐνη ρυθμίζει την κυτταρική ανάπτυξη και διαίρεση. Οι παθογόνοι παραλλαγές στο γονίδιο *APC* προδιαθέτουν στο **σύνδρομο της οικογενούς αδενωματώδους πολυποδίασης (Familial Adenomatous Polyposis – FAP)**. Τα άτομα με FAP εμφανίζουν εκατοντάδες έως χιλιάδες αδενωματώδεις πολύποδες στο παχύ και το λεπτό έντερο, στο δωδεκαδάκτυλο, στο φύμα του Vater και στο στομάχι, οι οποίοι εμφανίζονται στη δεύτερη με τρίτη δεκαετία της ζωής τους. Επιπλέον, έχουν σχεδόν 100% πιθανότητα να εμφανίσουν καρκίνο του παχέος εντέρου με μέση ηλικία διάγνωσης τα 35-45 έτη (Half *et al.*, 2009; Φωστήρα, 2009; Galiatsatos & Foulkes, 2006). Το σύνδρομο FAP κληρονομείται με αυτοσωμικό επικρατή τρόπο.

Ένα σύνδρομο που προσομοιάζει το σύνδρομο FAP είναι η ***MUTYH*-σχετιζόμενη πολυποδίαση (*MUTYH* Associated Polyposis – MAP)**. Το σύνδρομο MAP οφείλεται σε παθογόνους παραλλαγές στο γονίδιο *MUTYH* και κληρονομείται με αυτοσωμικό υπολειπόμενο τρόπο. Τα άτομα με σύνδρομο MAP έχουν επίσης σχεδόν 100% πιθανότητα να εμφανίσουν καρκίνο του παχέος εντέρου, ενώ εμφανίζουν εκατοντάδες πολύποδες στην άνω και στην κάτω γαστρεντερική οδό (Half *et al.*, 2009). Ωστόσο, τόσο η πολυποδίαση όσο και ο καρκίνος του παχέος εντέρου εμφανίζονται αργότερα στη ζωή των ατόμων με σύνδρομο MAP, με το μέσο όρο ηλικίας να είναι στα 46 και 48 έτη αντίστοιχα (Φωστήρα, 2009).

Το **σύνδρομο von Hippel-Lindau (VHL)** είναι ένα σπάνιο γενετικό σύνδρομο με αυτοσωμική επικρατούσα κληρονομικότητα που οφείλεται σε παθογόνους γαμετικές παραλλαγές στο **ογκοκατασταλτικό γονίδιο *VHL***. Τα άτομα με VHL έχουν σχεδόν 100% κίνδυνο ανάπτυξης καλοήθων ή/και κακοήθων όγκων σε διάφορα όργανα. Στο φάσμα των όγκων που αναπτύσσονται σε άτομα με σύνδρομο VHL ανήκουν τα αιμαγγειοβλαστώματα της παρεγκεφαλίδας, του αμφιβληστροειδούς και του νωτιαίου μυελού (διά βίου κίνδυνος έως 72%, 60% και 50% αντίστοιχα), το διαυγοκυτταρικό νεφροκυτταρικό καρκίνωμα (διά βίου κίνδυνος από 25%-60%) και οι παγκρεατικοί νευροενδοκρινείς όγκοι (διά βίου κίνδυνος έως 17%). Επιπλέον, τα άτομα με VHL έχουν έως 20% διά βίου κίνδυνο ανάπτυξης φαιοχρωμοκυττωμάτων ή παραγαγγλιωμάτων, έναντι του 0.0008% αντίστοιχου διά βίου κινδύνου στο γενικό πληθυσμό (Fishbein & Nathanson, 2012; Lonser *et al.*, 2003).

Εκτός από τα κληρονομικά σύνδρομα και τα γονίδια προδιάθεσης εμφάνισης καρκίνου που αναφέρθηκαν, υπάρχουν και άλλα γονίδια που σχετίζονται με κληρονομούμενα σύνδρομα που προδιαθέτουν στην ανάπτυξη όγκων. Τα σύνδρομα αυτά περιλαμβάνουν διαφόρων τύπων κακοήθειες, όπως κακοήθεια του παχέος εντέρου, του λεπτού εντέρου, των νεφρών, του θυρεοειδούς κ.ά., αλλά και

καλοήθεις όγκους. Επιπλέον, αναλόγως το σύνδρομο, τα άτομα που έχουν υψηλή προδιάθεση μπορεί να εμφανίζουν επιπρόσθετα διακριτά φαινοτυπικά χαρακτηριστικά, όπως πολύποδες, αδενώματα, λειομυώματα, δερματικές κηλίδες κ.ά.. Κάθε κληρονομούμενο καρκινικό σύνδρομο περιλαμβάνει διαφορετικά φαινοτυπικά χαρακτηριστικά κι επομένως είναι εξαιρετικά σημαντικό, η αξιολόγηση του κάθε ασθενούς και της οικογένειάς του, να πραγματοποιείται από ειδικά καταρτισμένους επαγγελματίες στον τομέα της γενετικής.

Στον πίνακα 2 παρουσιάζονται ενδεικτικά κάποια κληρονομικά σύνδρομα, οι τύποι καρκίνου και καλοηθών όγκων στους οποίους προδιαθέτουν καθώς και τα εμπλεκόμενα γονίδια.

Πίνακας 2: Κληρονομικά καρκινικά σύνδρομα, κύρια φαινοτυπικά χαρακτηριστικά και εμπλεκόμενα γονίδια.

Table 2: Hereditary cancer syndromes, main phenotypic traits and genes involved.

Σύνδρομο	Κακοήθειες	Άλλα φαινοτυπικά χαρακτηριστικά	Γονίδια
Κληρονομικού Καρκίνου μαστού/ωοθηκών	Μαστού, ωοθηκών, παγκρέατος		*BRCA1, BRCA2*
Μεσαίου κινδύνου κληρονομικός καρκίνος μαστού	Μαστού		*ATM, CHEK2*
***PALB2*-σχετιζόμενο καρκινικό σύνδρομο**	Μαστού		*PALB2*
Μεσαίου κινδύνου κληρονομικός καρκίνος ωοθηκών	Ωοθηκών		*RAD51C, RAD51D, BRIP1*
Li-Fraumeni	Σάρκωμα μαλακών μορίων, μαστού, οστεοσάρκωμα, εγκεφάλου, κεντρικού νευρικού συστήματος, λευχαιμίες		*TP53*
Cowden	Μαστού, θυρεοειδούς, ενδομητρίου, νεφρών κ.ά.	Αμαρτώματα, μεγαλοκεφαλία	*PTEN*
Κληρονομικός γαστρικός καρκίνος διάχυτου τύπου	Στομάχου (διάχυτου τύπου), μαστού (λοβιακού τύπου)		*CDH1*
Peutz-Jeghers	Παχέος εντέρου, λεπτού εντέρου, μαστού, ωοθηκών, παγκρέατος	Αρματωματώδεις πολύποδες, στίγματα με εναπόθεση μελανίνης	*STK11*
Lynch	Παχέος εντέρου, ενδομητρίου, ωοθηκών ουροδόχου κύστεως, παγκρέατος, στομάχου κ.ά.		*MLH1, MSH2, MSH6, PMS2, EPCAM (έμμεσα)*

Οικογενής Αδενωματώδης Πολυποδίαση	Παχέος εντέρου	Πολυποδίαση εντέρου	*APC*
***MUTYH*-σχετιζόμενη αδενωματώδης πολυποδίαση‡**	Παχέος εντέρου	Πολυποδίαση εντέρου	*MUTYH*
***NTHL1*-σχετιζόμενο καρκινικό σύνδρομο‡**	Παχέος εντέρου, μαστού, ενδομητρίου	Πολυποδίαση εντέρου	*NTHL1*
Νεανική πολυποδίαση	Παχέος εντέρου, λεπτού εντέρου, στομάχου	Πολυποδίαση εντέρου	*SMAD4, BMPR1A*
Συγγενής ανεπάρκεια επιδιόρθωσης λανθασμένων ζευγών βάσεων‡	Εγκεφάλου και κεντρικού νευρικού συστήματος, αιματολογικές κακοήθειες, σχετιζόμενοι με σύνδρομο Lynch		*MLH1, MSH2, MSH6, PMS2*
von Hippel-Lindau	Νεφρών	Αιμαγγειοβλαστώματα αμφιβληστροειδούς, νεφρών, φαιοχρωμοκύττωμα, παραγαγγλίωμα, παγκρεατικοί νευροενδοκρινείς όγκοι	*VHL*
Πολλαπλή ενδοκρινής νεοπλασία I	Παραθυρεοειδών αδένων, ενδοκρινούς γαστρεντερο-παγκρεατικής οδού, πρόσθιας υπόφυσης	Υπερπαραθυρεοειδισμός, πολυοζώδης βρογχοκήλη	*MEN1*
Πολλαπλή ενδοκρινής νεοπλασία II	Θυρεοειδούς (μυελοειδούς τύπου)	Φαιοχρωμοκύττωμα, παραγαγγλίωμα, αδένωμα του παραθυρεοειδούς	*RET*
Birt-Hogg-Dube	Νεφρών	Αυτόματος πνευμονοθώρακας, ινοθυλακιοσώματα	*FLCN*
Κληρονομικής λειομυομάτωσης και νεφροκυτταρικού καρκινώματος	Νεφρών	Λειομυώματα	*FH*
Αταξία τηλαγγειεκτασία‡	Λευχαιμίες, λεμφώματα, μαστού, ωοθηκών, στομάχου, ήπατος, σάρκωμα	Εκφύλιση νευρικών κυττάρων, τηλαγγειεκτασία, ανοσοανεπάρκεια	*ATM*
***CDKN2A*-σχετιζόμενο καρκινικό σύνδρομο**	Παγκρέατος, μελάνωμα		*CDKN2A*
Αναιμία Fanconi‡	Λευχαιμία, ακανθοκυτταρικό καρκίνωμα δέρματος	Ανεπάρκεια μυελού των οστών, συγγενείς ανεπάρκειες, αναπτυξιακές διαταραχές, καφεγαλακτόχροες κηλίδες	*FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL, FANCM, BRCA1, BRCA2, PALB2, RAD51C*

Bloom‡	Λευχαιμίες, λεμφώματα, πλακώδες καρκίνωμα, όγκοι Wilms, γαστρεντερικού σωλήνα	Μικρό ανάστημα, φωτοευαισθησία, ήπια ανοσοανεπάρκεια, ινσουλινοαντίσταση	*BLM*
Νευροϊνωμάτωση τύπου 1	Λευχαιμίες, εγκεφάλου, νευροϊνοσάρκωμα, φαιοχρωμοκύττωμα, οπτικά γλοιώματα, μηνιγγίωμα	Καφεγαλακτόχροες κηλίδες, νευροϊνώματα	*NF1*
Νευροϊνωμάτωση τύπου 2	Κεντρικό νευρικό σύστημα, μηνιγγίωμα	Σβαννώματα, καφεγαλακτόχροες κηλίδες	*NF2*
Οζώδης σκλήρυνση		Δερματικές δυσμορφίες, επιληπτικά επεισόδια, συμπεριφορικές διαταραχές, καλοήθεις όγκοι νεφρών, καρδιάς, πνευμόνων, επενδυματικά οζίδια εγκεφάλου, κ.ά.	*TSC1, TSC2*
Συμπλέγματος Carney		Καρδιακά μυξώματα, καλοήθεις ενδοκρινείς όγκοι, σβαννώματα	*PRKAR1A*
Gorlin	Βασικών κυττάρων, μυελοβλαστώματα	Κερατοκυστικοί οδοντογενείς όγκοι	*PTCH1, SUFU*
Shwachman-Diamond	Οξεία μυελογενής λευχαιμία	Εξωκρινής παγκρεατική ανεπάρκεια, δυσλειτουργία του μυελού των οστών, σκελετικές ανωμαλίες	*SBDS*
Ρετινοβλάστωμα	Ρετινοβλάστωμα, οστεοσάρκωμα		*RB1*

‡ Υπολειπόμενος τρόπος κληρονόμησης

1.4. Προληπτικές και θεραπευτικές προσεγγίσεις στην Ιατρική Ακριβείας στην Ογκολογία

1.4.1. Αναστολείς PARP

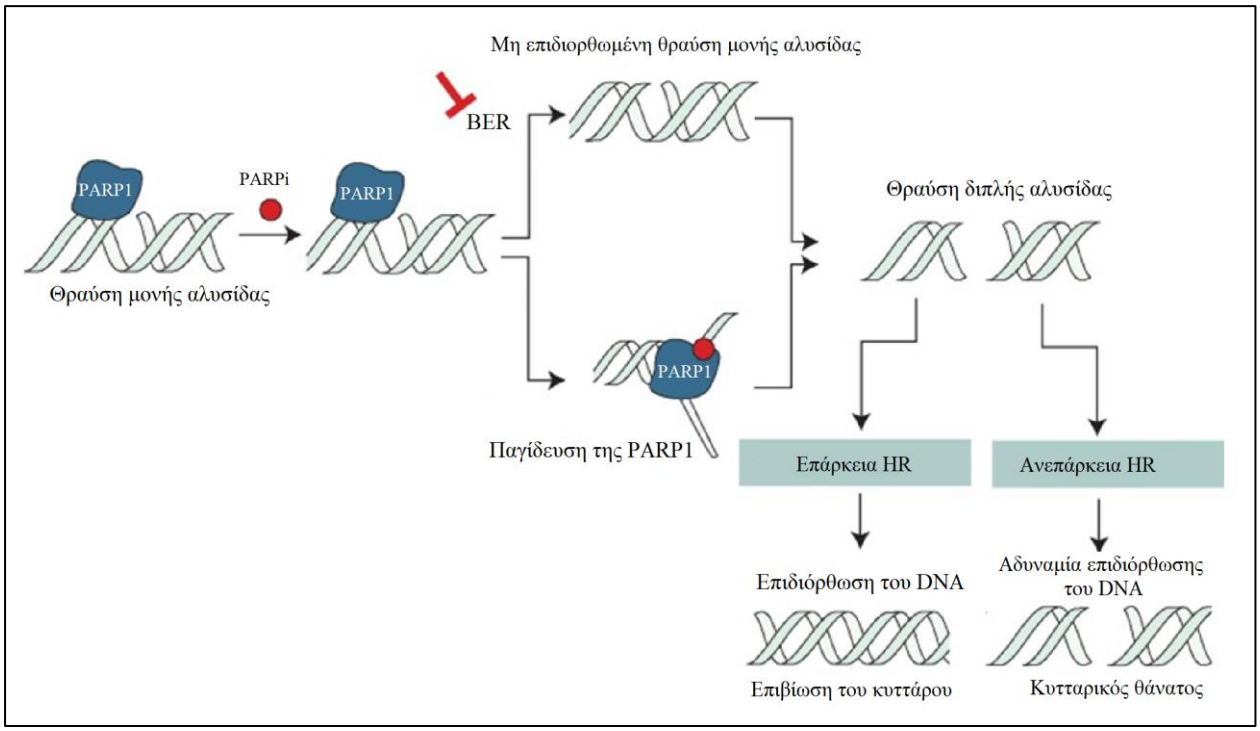
Ένα από τα κύρια χαρακτηριστικά των καρκινικών κυττάρων είναι η **αντίσταση στην απόπτωση** (Hanahan & Weinberg, 2011). Για να μπορέσουν να αποφύγουν τον **κυτταρικό θάνατο**, τα καρκινικά κύτταρα ενεργοποιούν κάποιο μηχανισμό που προκαλεί υπερέκφραση των αντι-αποπτωτικών πρωτεϊνών ή/και μειωμένη έκφραση ή δυσλειτουργία των μορίων που επάγουν την απόπτωση, μέσω της συσσώρευσης παραλλαγών και γονιδιωματικών αναδιατάξεων (Fulda, 2010). Οι μηχανισμοί που ευνοούν την επιβίωση των καρκινικών κυττάρων μπορούν να αποτελέσουν **θεραπευτικούς στόχους**.

Ένα χαρακτηριστικό παράδειγμα, είναι οι όγκοι με ανεπάρκεια HR. Οι όγκοι αυτοί δεν έχουν τη δυνατότητα να επιδιορθώσουν τις βλάβες της διπλής αλυσίδας του DNA μέσω του HR, είτε εξαιτίας παθογόνων παραλλαγών σε κάποιο από τα γονίδια που μετέχουν στο μηχανισμό, λόγου χάρη σε κάποιο εκ των *BRCA1* και *BRCA2*, είτε εξαιτίας επιγενετικών τροποποιήσεων. Τα κύτταρα αυτά, για την αποφυγή της απόπτωσης, ενεργοποιούν άλλους μηχανισμούς επιδιόρθωσης του γενετικού υλικού.

Μια πολλά υποσχόμενη θεραπεία για τους όγκους με ανεπάρκεια του ομόλογου ανασυνδυασμού είναι η **θεραπεία με αναστολείς PARP** (Πολυμεράσες της πολυαδενοφωσφορικής ριβόζης - Poly-ADP ribose polymerases). Ο τρόπος δράσης αυτής της ομάδας φαρμάκων βασίζεται στην έννοια της **συνθετικής κυτταρικής θνησιμότητας** (synthetic lethality). Η συνθετική θνησιμότητα είναι ένα φαινόμενο κατά το οποίο η απενεργοποίηση δύο ή περισσότερων γονιδίων μέσω διαφορετικών γενετικών συμβάντων οδηγεί σε κυτταρικό θάνατο, ενώ η απενεργοποίηση καθενός από αυτά τα γονίδια ξεχωριστά είναι ανεκτή από το κύτταρο (Kaelin, 2005).

Η **πρωτεΐνη PARP1** διαδραματίζει καθοριστικό ρόλο στην επισκευή του DNA, καθώς σηματοδοτεί την έναρξη των μηχανισμών επιδιόρθωσης. Η **θεραπεία με αναστολείς PARP** βασίστηκε στη θεωρία ότι τα κύτταρα με ανεπαρκή HR βασίζονται στο μηχανισμό επιδιόρθωσης BER για τη διατήρηση της γονιδιωματικής ακεραιότητας (Saleh-Gohari *et al.*, 2005). Πιο συγκεκριμένα, στον BER η πρωτεΐνη PARP1 στρατολογεί την XRCC1, την πρωτεΐνη που λειτουργεί ως ικρίωμα για τη συναρμολόγηση του συμπλόκου πρωτεϊνών που πραγματοποιούν την επιδιόρθωση. Η αναστολή της PARP1 έχει ως αποτέλεσμα την αδυναμία επιδιόρθωσης των θραύσεων μονής αλυσίδας μέσω του μηχανισμού BER, οι οποίες στη συνέχεια είναι εύκολο να εξελιχθούν σε θραύσεις διπλής αλυσίδας κατά την αντιγραφή του DNA. Κατ' αυτόν τον τρόπο, στα καρκινικά κύτταρα τα οποία παρουσιάζουν ανεπάρκεια του HR, οι θραύσεις της διπλής αλυσίδας DNA συσσωρεύονται και, συνεπώς, τα κύτταρα αυτά οδηγούνται σε απόπτωση (Mateo *et al.*, 2019; Rouleau *et al.*, 2010).

Τα τελευταία χρόνια, έχει προταθεί μία επιπλέον θεωρία για τον μηχανισμό λειτουργίας των αναστολέων PARP, μιας κι έχει αποδειχθεί πως η αναστολή της πρωτεΐνης XRCC1, που διαδραματίζει σημαντικό ρόλο στον BER, σε κύτταρα με ανεπάρκεια HR, δεν οδηγεί σε απόπτωση των κυττάρων (Helleday, 2011). Η θεωρία αυτή υποστηρίζει πως οι αναστολείς της PARP1 οδηγούν σε παγίδευση της PARP1 στο κατεστραμμένο DNA, με αποτέλεσμα την καταστολή της δημιουργίας της διχάλας αντιγραφής. Στα φυσιολογικά κύτταρα, η στάσιμη διχάλα αντιγραφής επιδιορθώνεται μέσω του HR. Ωστόσο, στα κύτταρα με ανεπάρκεια HR στρατολογείται ο NHEJ, ο οποίος είναι επιρρεπής σε λάθη, με αποτέλεσμα τη συσσώρευση παραλλαγών και τελικά τη γονιδιωματική αστάθεια η οποία οδηγεί το κύτταρο σε απόπτωση (Mateo *et al.*, 2019) (Εικόνα 12).



Εικόνα 12: Ο ρόλος των αναστολέων PARP στη συνθετική θνησιμότητα. Τροποποίηση από (Mateo *et al.*, 2019).

Figure 12: The role of PARP inhibitors in synthetic mortality. Modification from από (Mateo *et al.*, 2019).

Η θεραπεία με αναστολείς PARP πρόκειται για μια πολλά υποσχόμενη θεραπεία, καθώς στοχεύει μόνο στα καρκινικά κύτταρα και, κατά συνέπεια, είναι περισσότερο ειδική και λιγότερο τοξική από τις παραδοσιακές χημειοθεραπευτικές μεθόδους. Μέχρι στιγμής, οι βιοδείκτες που υπάρχουν για τη χορήγηση αναστολέων PARP είναι οι γαμετικές και σωματικές παθογόνοι παραλλαγές στα γονίδια *BRCA1* και *BRCA2.* Ωστόσο, φαίνεται πως υπάρχουν κι επιπλέον όγκοι που παρουσιάζουν ευαισθησία στους αναστολείς PARP. Για το λόγο αυτό, μελετάται η χορήγησή τους σε όγκους που φέρουν παθογόνους παραλλαγές και σε άλλα γονίδια που διαδραματίζουν σημαντικό ρόλο στην ορθή λειτουργία του HR.

1.4.2. Προληπτικά μέτρα για άτομα υψηλού κινδύνου

Η ανίχνευση της **κληρονομικής προδιάθεσης** στον καρκίνο είναι εξαιρετικά σημαντική όχι μόνο για την θεραπεία του ατόμου που έχει νοσήσει, αλλά πρωτίστως για την **πρόληψη της νόσου**. Οι ασθενείς με καρκίνο στους οποίους έχει ανιχνευθεί μία γαμετική παθογόνος παραλλαγή σε κάποιο γονίδιο προδιάθεσης στον καρκίνο έχουν υψηλό κίνδυνο ανάπτυξης δεύτερου πρωτοπαθούς όγκου. Επιπλέον, μέσω τις ανίχνευσης αυτών των ασθενών μπορούν να ταυτοποιηθούν και οι υγιείς συγγενείς τους που φέρουν την ίδια παθογόνο παραλλαγή κι επομένως έχουν υψηλό κίνδυνο εμφάνισης καρκίνου.

Στη συνέχεια, τα άτομα αυτά μπορούν να μπουν σε εξειδικευμένα πρωτόκολλα παρακολούθησης με σκοπό την έγκαιρη διάγνωσή τους ή/και να προβούν σε προφυλακτικά χειρουργεία για τη μείωση του κινδύνου εμφάνισης καρκίνου (Kulkarni & Carley, 2016). Επιπλέον, μπορεί να λάβουν κάποιο φάρμακο

προληπτικά, ανάλογα το γονίδιο στο οποίο φέρουν παθογόνο παραλλαγή. Για παράδειγμα, υπάρχουν δεδομένα που υποστηρίζουν πως η χορήγηση ασπιρίνης σε ημερήσια βάση σε άτομα με γενετική προδιάθεση στο σύνδρομο Lynch, μειώνει τον κίνδυνο εμφάνισης του καρκίνου παχέος εντέρου (Gupta *et al.*, 2020). Το National Comprehensive Cancer Network (NCCN) έχει θεσπίσει οδηγίες, τις οποίες ενημερώνει σε ετήσια βάση, για τη διαχείριση ατόμων με κληρονομική προδιάθεση σε καρκινικά σύνδρομα (Daly *et al.*, 2020; Gupta *et al.*, 2020).

1.5. Μέθοδος Αλληλούχησης Επόμενης Γενιάς

Η **μέθοδος Αλληλούχησης Επόμενης Γενιάς (Next Generation Sequencing - NGS)** άρχισε να αναπτύσσεται στα μέσα έως τα τέλη της δεκαετίας του 1990, αλλά είχε εμπορική εφαρμογή για πρώτη φορά μετά την έλευση του 21ου αιώνα. Η νέα αυτή μέθοδος αλληλούχησης ονομάστηκε «επόμενης γενιάς» ή «δεύτερης γενιάς» ώστε να μπορεί να ξεχωρίζει από τις έως τότε επικρατούσες μεθόδους, συμπεριλαμβανομένης της αλληλούχησης κατά Sanger, οι οποίες παράγουν γενετικά δεδομένα σε πολύ μικρότερη κλίμακα. Σε αντίθεση με την πρώτη γενιά αλληλούχησης, η τεχνολογία Αλληλούχησης Επόμενης Γενιάς χαρακτηρίζεται τυπικά από την υψηλή απόδοσή της, η οποία επιτρέπει τη μαζική και παράλληλη αλληλούχηση ολόκληρων γονιδιωμάτων. Γι' αυτόν το λόγο, άλλες ονομασίες με τις οποίες είναι γνωστή είναι **«μαζική παράλληλη αλληλούχηση»** και **«αλληλούχηση υψηλής απόδοσης»**. Συνήθως, αυτό επιτυγχάνεται με τον **κατακερματισμό του γονιδιώματος σε μικρά κομμάτια (fragmentation)**· στη συνέχεια, επιλέγονται τυχαία θραύσματα γενετικού υλικού τα οποία διαβάζονται με διαφορετική τεχνολογία ανά πλατφόρμα αλληλούχησης επόμενη γενιάς. Αυτά τα τυχαία θραύσματα μπορούν να διαβαστούν παράλληλα και με αυτόν τον τρόπο γίνεται εφικτή η αλληλούχηση ολόκληρου του γονιδιώματος (Slatko *et al.*, 2018; Behjati & Tarpey, 2013) .

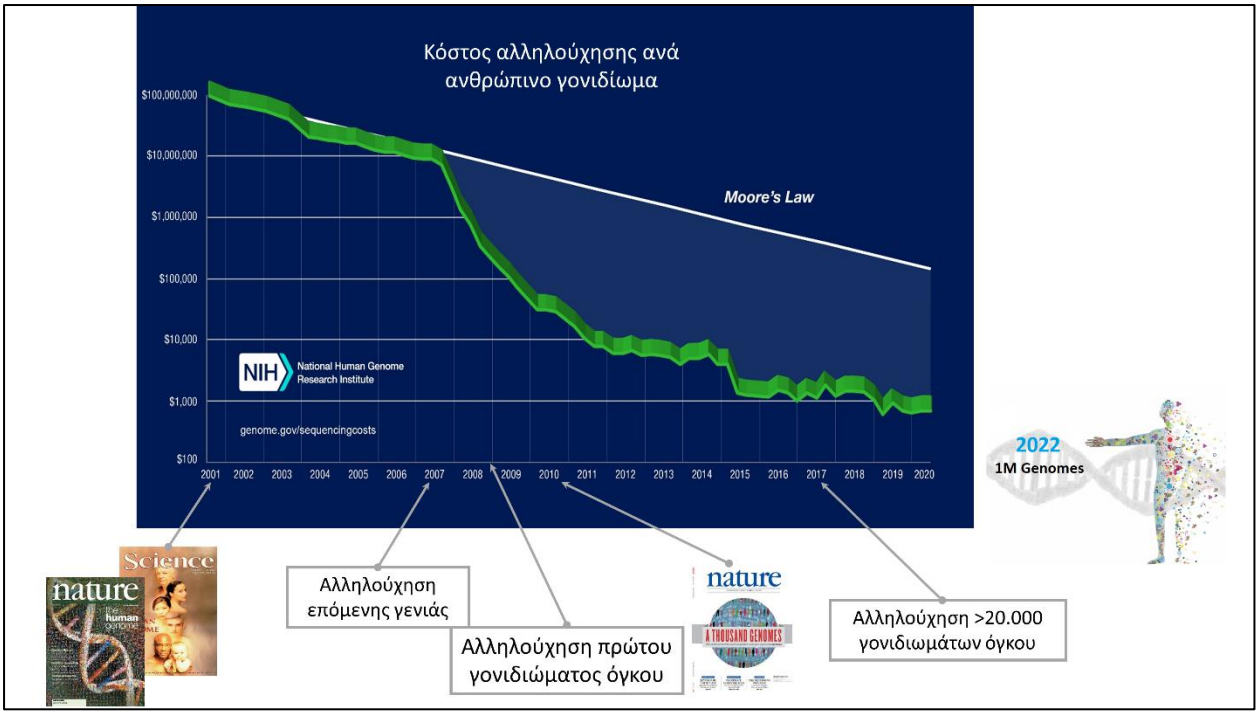
Η τεχνολογία Αλληλούχησης Επόμενης Γενιάς, συνέβαλε ριζικά στην επανάσταση στον τομέα της γενετικής που λαμβάνει χώρα τα τελευταία χρόνια. Ως αποτέλεσμα, χάρη σε αυτή την προηγμένη τεχνολογία τόσο ο χρόνος όσο και το κόστος αλληλούχησης έχει μειωθεί ραγδαία, γεγονός που έχει οδηγήσει στην παραγωγή ενός τεράστιου όγκου γενετικών δεδομένων σε καθημερινή βάση (Hu *et al.*, 2021; Mardis, 2013). Παράλληλα, η τεχνολογία αλληλούχησης επόμενη γενιάς βελτίωσε σημαντικά την κατανόησή μας ως προς την ποικιλομορφία του ανθρώπινου γονιδιώματος, ενώ βοήθησε τους ερευνητές και τους κλινικούς ιατρούς να κατανοήσουν καλύτερα τους μοριακούς μηχανισμούς του ανθρώπινου σώματος και, ιδιαίτερα, των γενετικών διαταραχών (Dhawan, 2017; Hood, 2008).

1.5.1. Ιστορική αναδρομή

Από το 1977, οπότε ο Frederick Sanger εισήγαγε την ομώνυμη μέθοδο αλληλούχησης (Sanger *et al.*, 1977), η **αλληλούχηση κατά Sanger** ήταν η κύρια μέθοδος αλληλούχησης για τουλάχιστον τρεις δεκαετίες. Η ανακάλυψη της μεθόδου αλληλούχησης κατά Sanger αποτέλεσε μία τεράστια επανάσταση για τη γενετική και τη βιοτεχνολογία (Mardis, 2013), ωστόσο η μέθοδος είναι αρκετά χρονοβόρα και κοστοβόρα αν και βελτιώθηκε σαφώς με την πάροδο του χρόνου. Είναι αξιοσημείωτο ότι το **Πρόγραμμα Αλληλούχησης του Ανθρώπινου Γονιδιώματος (Human Genome Project – HGP)** (Lander *et al.*, 2001;

Venter *et al.*, 2001) ολοκληρώθηκε με τη χρήση πολλαπλών αναλυτών αλληλούχησης κατά Sanger εντός δεκατριών ετών - με τον αρχικό χρόνο να υπολογίζεται στα δεκαπέντε έτη – και κόστισε 2,7 δισεκατομμύρια δολάρια (οικονομικού έτους 1991), έναντι των τριών δισεκατομμυρίων δολαρίων που είχε αρχικά είχε εκτιμηθεί ότι θα κοστίσει (Hood & Rowen, 2013).

Στην Εικόνα 13 παρουσιάζεται το κόστος αλληλούχησης ενός ανθρώπινου γονιδιώματος σε συνάρτηση με το χρόνο. Το 2001, που ήταν και η χρονιά που ολοκληρώθηκε το Πρόγραμμα Αλληλούχησης του Ανθρώπινου Γονιδιώματος, χρειάζονταν 100.000.000 δολάρια για να αλληλουχηθεί ένα μόνο γονιδίωμα. Το κόστος αλληλούχησης στη συνέχεια έπεφτε ακολουθώντας το νόμο του Μουρ, που περιγράφει μια μακροπρόθεσμη τάση στη βιομηχανία υλικού υπολογιστών που περιλαμβάνει τον διπλασιασμό της υπολογιστικής ισχύος κάθε δύο χρόνια. Το 2007 όμως, παρατηρείται μία κάθετη πτώση στο κόστος αλληλούχησης ενός γονιδιώματος, η οποία σηματοδοτείται από την εισαγωγή της μεθόδου Αλληλούχησης Επόμενης Γενιάς. Φτάνοντας στη σημερινή εποχή, το κόστος αλληλούχησης ενός ανθρώπινου γονιδιώματος ανέρχεται μόλις στα χίλια δολάρια (Wetterstand, 2020).



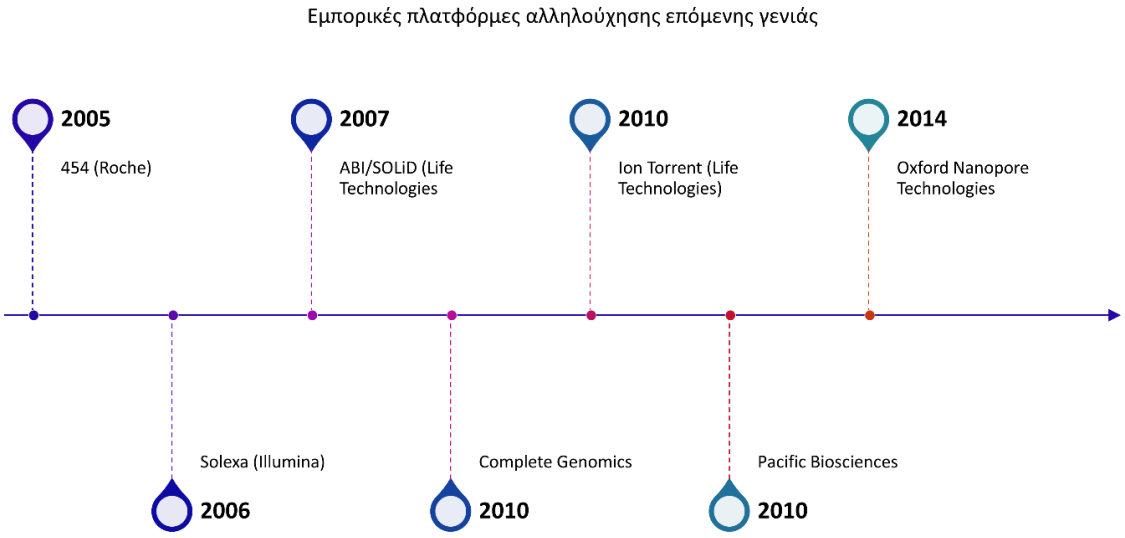
Εικόνα 13: Κόστος αλληλούχησης ανά ανθρώπινο γονιδίωμα. Αρχική εικόνα από (Wetterstand, 2020).

Figure 13: Sequencing costs per human genome. Original image by (Wetterstand, 2020).

Από τα μέσα της πρώτης δεκαετίας του 21ού αιώνα, οι τεχνολογίες Αλληλούχησης Επόμενης Γενιάς (NGS) αναδύονται ως μία από τις πιο οικονομικές, γρήγορες και υψηλής απόδοσης μεθόδους προσδιορισμού αλληλουχιών DNA. Η μέθοδος Αλληλούχησης Επόμενης Γενιάς στην πραγματικότητα περιγράφηκε το 2005 από τη Roche 454, αλλά εμπορευματοποιήθηκε το 2007 (Mardis, 2013). Στη

συνέχεια, άρχισαν να αναδεικνύονται και νέες πλατφόρμες, όπως η Solexa το 2006, η οποία εξαγοράστηκε από την εταιρία Illumina ένα χρόνο αργότερα, η ABI/SOLiD το 2007 και η IonTorrent το 2010 (Meera Krishna *et al.*, 2019; Barba *et al.*, 2014).

Οι διάφορες πλατφόρμες Αλληλούχησης Επόμενης Γενιάς έχουν κοινά χαρακτηριστικά, όπως για παράδειγμα η υψηλή απόδοση, η παράλληλη δημιουργία πολλαπλών σύντομων **αναγνώσεων (reads) DNA**, η υψηλή ταχύτητα, το χαμηλό κόστος και η υψηλή ακρίβεια. Ωστόσο, χωρίζονται σε δύο κύριες κατηγορίες βάσει της τεχνολογίας τους: την **αλληλούχηση με συρραφή (Sequencing by Ligation – SBL)** και την **αλληλούχηση με σύνθεση (Sequencing by Synthesis – SBS)** (Heather & Chain, 2016). Η μέθοδος αλληλούχησης με συρραφή χρησιμοποιεί την ευαισθησία της λιγάσης DNA στην αναντιστοιχία βάσεων ώστε να καθορίσει την υποκείμενη ακολουθία των νουκλεοτιδίων σε μια αλληλουχία DNA (McKernan *et al.*, 2009). Η μέθοδος αλληλούχησης με σύνθεση χρησιμοποιεί επισημασμένα νουκλεοτίδια τα οποία εισάγονται μεμονωμένα ώστε να αναγνωρίζονται σε πραγματικό χρόνο καθώς πραγματοποιείται η επέκταση. Παράλληλα, χρησιμοποιεί DNA πολυμεράση ή λιγάση ώστε να περικλείει πολλούς κλώνους DNA ταυτόχρονα, γεγονός που ενισχύει το παραγόμενο σήμα που εκλύεται από τα επισημασμένα νουκλεοτίδια (Turcatti *et al.*, 2008).



Εικόνα 14: Χρονοδιάγραμμα της εξέλιξης της τεχνολογίας Αλληλούχησης Επόμενης Γενιάς.

Figure 14: Evolution of next generation sequencing technology.

Οι τεχνολογίες Αλληλούχησης Επόμενης Γενιάς, έχουν ως έξοδο μικρές αναγνώσεις DNA, που συνήθως έχουν μήκος έως 300 ζεύγη βάσεων. Τα τελευταία χρόνια εισήχθησαν τεχνολογίες αλληλούχησης με τη δυνατότητα ανάγνωσης μονών μεγάλων μορίων DNA (έως και 10.000 ζεύγη βάσεων), οι οποίες είναι γνωστές με την ονομασία **«αλληλούχηση τρίτης γενιάς».** Οι πιο σημαντικές

πλατφόρμες αλληλούχησης τρίτης γενιάς είναι η πλατφόρμα Pacific Bioscience, η οποία διατέθηκε στην αγορά το 2010 (van Dijk *et al.*, 2014) και η πλατφόρμα Oxford Nanopore που εισήχθη για πρώτη φορά το 2014 (Heather & Chain, 2016; J. Clarke *et al.*, 2009) (Εικόνα 14). Η τεχνολογία αλληλούχησης τρίτης γενιάς έχει ορισμένα πλεονεκτήματα έναντι της Αλληλούχησης Επόμενης Γενιάς, όπως η ευκολία εντοπισμού μεγάλων γονιδιωματικών αναδιατάξεων, αλλά εν γένει είναι λιγότερο ακριβής. Στην παρούσα διατριβή θα μελετηθεί μόνο η μέθοδος Αλληλούχησης Επόμενης Γενιάς.

1.5.2. Εφαρμογές αλληλούχησης DNA επόμενης γενιάς

Η τεχνολογία αλληλούχησης DNA επόμενης γενιάς έχει τρεις βασικές εφαρμογές στη γενετική, ανάλογα το αντικείμενο μελέτης· την αλληλούχηση ολόκληρου του γονιδιώματος (Whole Genome Sequencing-WGS), την αλληλούχηση μόνο των κωδικών περιοχών του γονιδιώματος (Whole Exome Sequencing-WES) και την στοχευμένη αλληλούχηση συγκεκριμένης ομάδας γονιδίων (γονιδιακό πάνελ, gene panel sequencing) (Εικόνα 15).

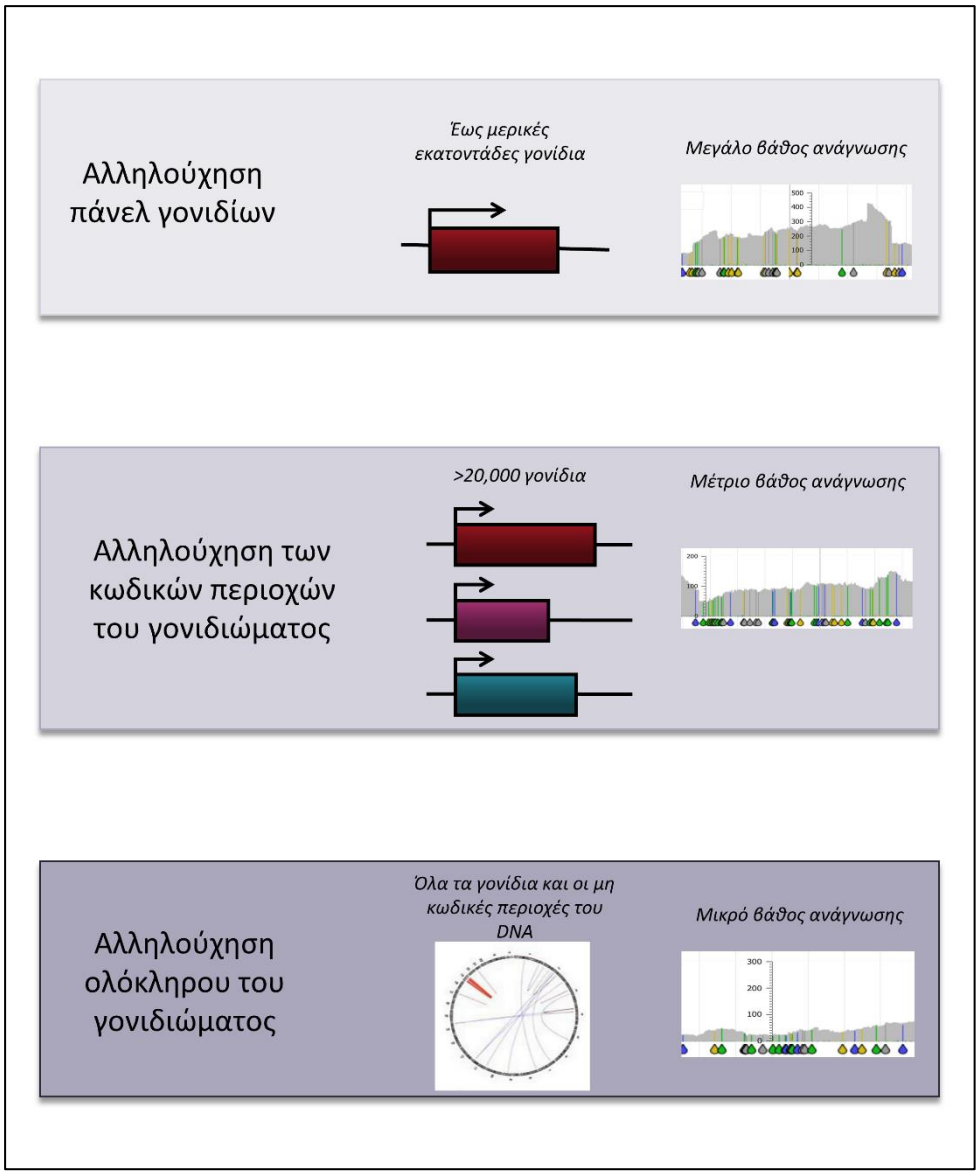
Η **μέθοδος αλληλούχησης ολόκληρου του γονιδιώματος** είναι η διαδικασία με την οποία δύναται να αναγνωστούν όλες ή σχεδόν όλες οι περιοχές του γονιδιώματος ενός ανθρώπου, σε όλα τα χρωμοσώματα, από την αλληλουχία των γονιδίων έως την αλληλουχία των περιοχών μεταξύ των γονιδίων και το μιτοχονδριακό DNA. Καθώς το ανθρώπινο γονιδίωμα είναι αρκετά μεγάλο – μιας και αποτελείται από περίπου τρία δισεκατομμύρια ζεύγη βάσεων – ένα τυπικό πείραμα αλληλούχησης ολόκληρου του γονιδιώματος ενός ανθρώπου δεν μπορεί να έχει μεγάλο **βάθος ανάγνωσης (read depth)**, δηλαδή δεν μπορεί να έχει μεγάλο αριθμό αναγνώσεων DNA σε κάθε θέση του γονιδιώματος**.** Το προτεινόμενο μέσο βάθος ανάγνωσης είναι 30x-50x (30-50 αναγνώσεις σε κάθε θέση του γονιδιώματος) (Barbitoff *et al.*, 2020; Sun *et al.*, 2015), ενώ πολλές εφαρμογές παράγουν βάθος ανάγνωσης έως 10x. Για το λόγο αυτό, μέχρι και σήμερα η αλληλούχηση ολόκληρου του γονιδιώματος χρησιμοποιείται κυρίως για ερευνητικούς σκοπούς. Ωστόσο, με την πάροδο του χρόνου η ακρίβεια της μεθόδου αυξάνεται σταθερά έχοντας ως αποτέλεσμα τη σταδιακή είσοδό της στην κλινική διάγνωση (Lionel *et al.*, 2018).

Η **μέθοδος αλληλούχησης των κωδικών περιοχών του γονιδιώματος** επιτρέπει την αλληλούχηση των εξονίων των γονιδίων που κωδικοποιούν για πρωτεΐνες. Οι κωδικές περιοχές του γονιδιώματος αποτελούν ένα μικρό ποσοστό ολόκληρου του γονιδιώματος (μόλις 1%-2%), ωστόσο υπολογίζεται πως περίπου το 85% των παραλλαγών που ευθύνονται για την εμφάνιση γενετικών νοσημάτων βρίσκονται εντός των εξονίων. Το μέσο βάθος ανάγνωσης σε ένα πείραμα αλληλούχησης των κωδικών περιοχών των γονιδίων είναι 60-100x, γεγονός που αυξάνει την ακρίβεια του (Barbitoff *et al.*, 2020; Sun *et al.*, 2015). Όλες αυτές οι παράμετροι καθιστούν τη μέθοδο αλληλούχησης των κωδικών περιοχών του γονιδιώματος πολύ πιο αποδοτική από άποψη χρόνου και χρημάτων σε σχέση με τη μέθοδο αλληλούχησης ολόκληρου του γονιδιώματος. Έτσι, η μέθοδος αλληλούχησης μόνο των κωδικών περιοχών βρίσκει όλο και συχνότερα κλινική εφαρμογή.

Η τελευταία εφαρμογή της μεθόδου Αλληλούχησης Επόμενης Γενιάς είναι η **στοχευμένη αλληλούχηση μιας ομάδας γονιδίων** ενδιαφέροντος. Η μέθοδος αυτή στοχεύει στον καθορισμό της

αλληλουχίας συγκεκριμένων γονιδίων τα οποία είναι γνωστό ότι σχετίζονται με τη νόσο που μελετάται, είτε επειδή έχει αποδειχθεί είτε επειδή υπάρχει ισχυρή υποψία ότι παθογόνοι παραλλαγές σε αυτά ευθύνονται για την εμφάνιση της νόσου (Xue *et al.*, 2015). Ένα γονιδιακό πάνελ συνήθως περιλαμβάνει από πέντε έως και μερικές εκατοντάδες γονίδια ενδιαφέροντος. Επιπλέον, συνήθως στα γονιδιακά πάνελ περιλαμβάνονται και σημειακές κοινές παραλλαγές που έχουν ανιχνευθεί από μελέτες συσχέτισης ολόκληρου του γονιδιώματος. Αυτές οι κοινές παραλλαγές εντοπίζονται σε συχνότητα μεγαλύτερη του 1% στον γενικό πληθυσμό και από μόνες τους προσδίδουν μόλις έως μιάμιση φορά μεγαλύτερο κίνδυνο εμφάνισης της νόσου, ωστόσο λειτουργούν πολλαπλασιαστικά (Kurian *et al.*, 2016). Στην παρούσα διατριβή, οι κοινές παραλλαγές που έχουν ανιχνευθεί από μελέτες συσχέτισης ολόκληρου του γονιδιώματος δε θα μελετηθούν περαιτέρω. .

Η στοχευμένη αλληλούχηση ομάδας γονιδίων είναι η μέθοδος που έχει τη μεγαλύτερη ακρίβεια και αυτό οφείλεται στο ότι το μέσο βάθος ανάγνωσης κυμαίνεται από 150x-400x. Για το λόγο αυτό, θεωρείται ότι είναι μια μέθοδος αλληλούχησης «μεγάλου βάθους» (high depth). Το μεγάλο βάθος ανάγνωσης αυτής της μεθόδου επιτρέπει την ακριβή ανάγνωση ενθέσεων και απαλοιφών μεγέθους μερικών δεκάδων νουκλεοτιδίων. Χάρη στην υψηλή της ακρίβεια και το πολύ χαμηλό κόστος, η στοχευμένη αλληλούχηση των περιοχών ενδιαφέροντος χρησιμοποιείται κατά κόρον στην κλινική διάγνωση (Kurian *et al.*, 2016). Ωστόσο, η ερευνητική εφαρμογή της είναι περιορισμένη, καθώς τα γονίδια που συμπεριλαμβάνονται στα γονιδιακά πάνελ είναι συνήθως ήδη γνωστά για το ρόλο τους στην προδιάθεση στη νόσο που μελετάται.



Εικόνα 15: Εφαρμογές της τεχνολογίας Αλληλούχησης Επόμενης Γενιάς για την αλληλούχηση DNA.

Figure 15: Applications of DNA next generation sequencing.

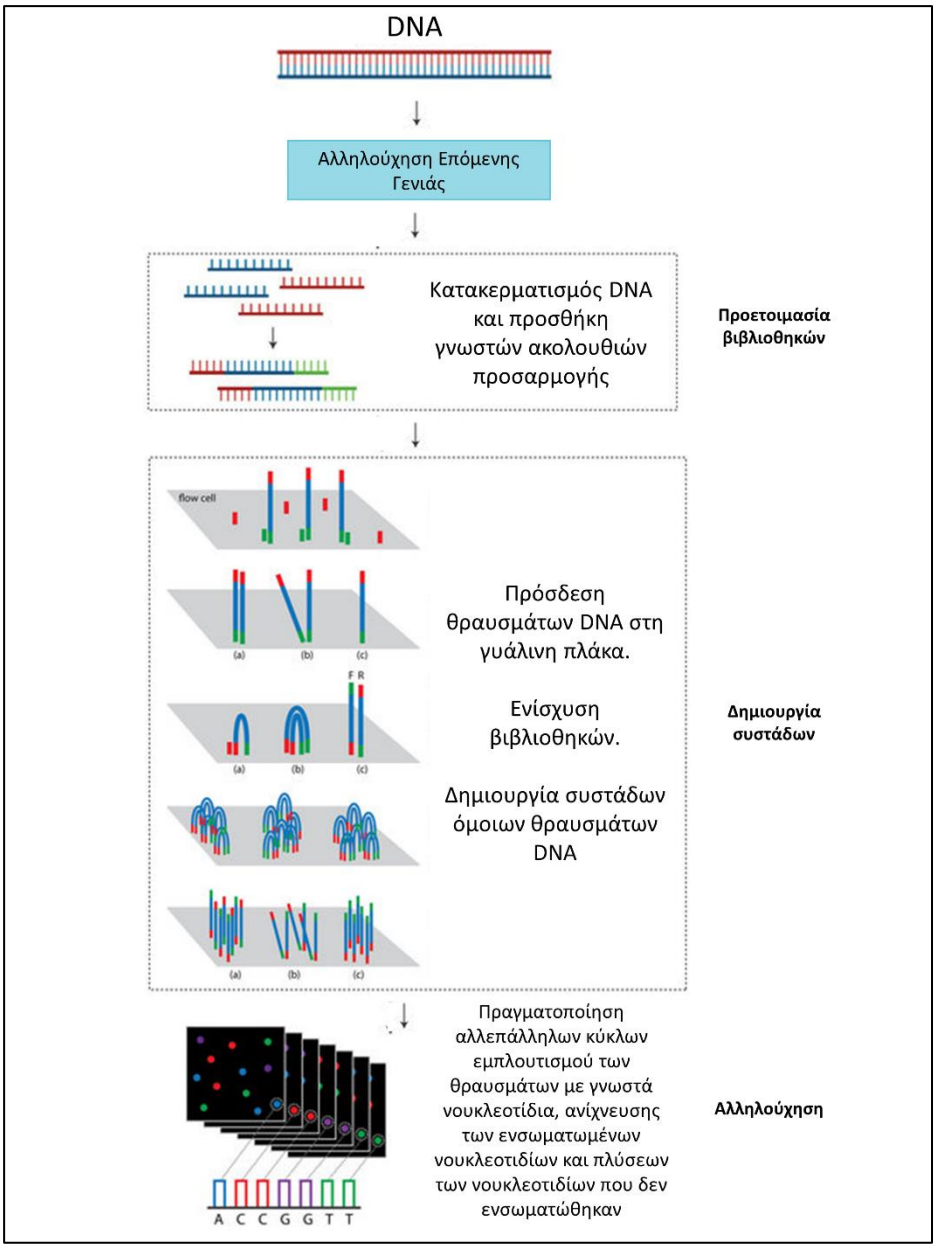
1.5.3. Βασικές αρχές λειτουργίας της μεθόδου Αλληλούχησης Επόμενης Γενιάς

Παρόλο που η μέθοδος Αλληλούχησης Επόμενης Γενιάς μπορεί να έχει πολλαπλές εφαρμογές, υπάρχουν κάποιες βασικές αρχές που ακολουθούνται σε κάθε πείραμα. Αρχικά, πραγματοποιείται προετοιμασία των δειγμάτων για την αλληλούχηση. Το βήμα αυτό περιλαμβάνει τον **κατακερματισμό (fragmentation)** του μορίου-στόχου και στη συνέχεια την προσθήκη γνωστών **αλληλουχιών προσαρμογής (adapter sequences)** στα άκρα των **θραυσμάτων (fragment)** που έχουν προκύψει. Στην περίπτωση που το πείραμα περιλαμβάνει περισσότερα από ένα δείγματα, προστίθεται επιπλέον σε κάθε θραύσμα DNA και ένας μοναδικός **μοριακός κωδικός (molecular barcode)**, ο οποίος επιτρέπει τη μετέπειτα αναγνώριση του δείγματος. Ένα πείραμα με πολλά δείγματα ονομάζεται **πείραμα με πολυπλεξία (multiplexed experiment)**. Ως αποτέλεσμα της προετοιμασίας των δειγμάτων,

κατασκευάζονται οι **βιβλιοθήκες των νουκλεϊκών οξέων**, οι οποίες στη συνέχεια είναι έτοιμες να πολλαπλασιαστούν ώστε να ακολουθήσει η ανάγνωση των αλληλουχιών τους.

Στη συνέχεια, πραγματοποιείται η ενίσχυση των βιβλιοθηκών, η οποία στις πλατφόρμες της Illumina πραγματοποιείται σε μια **γυάλινη πλάκα (flow cell)**. Τα θραύσματα DNA, έχοντας προσαρμοσμένα τα γνωστά ολιγονουκλεοτίδια στα άκρα τους, προσδένονται στα συμπληρωματικά ολιγονουκλεοτίδια που είναι προσαρμοσμένα στην επιφάνεια της γυάλινης πλάκας και στη συνέχεια πολλαπλασιάζονται με σκοπό τη **δημιουργία συστάδων από όμοια θραύσματα DNA (cluster generation)**. Η ανάγνωση της ακολουθίας των θραυσμάτων DNA πραγματοποιείται με αλληλούχηση με σύνθεση· κατά τη διάρκεια αυτής της διαδικασίας, τα θραύσματα DNA αλληλουχούνται παράλληλα. Η αλληλούχηση περιλαμβάνει αλλεπάλληλους κύκλους εμπλουτισμού των θραυσμάτων με γνωστά νουκλεοτίδια, ανίχνευσης των ενσωματωμένων νουκλεοτιδίων και πλύσεων των νουκλεοτιδίων που δεν ενσωματώθηκαν. Η ανίχνευση της ακολουθίας στις πλατφόρμες της Illumina στηρίζεται στην ανίχνευση του φθορισμού που δημιουργείται από την ενσωμάτωση των σημασμένων νουκλεοτιδίων με φθορίζουσα ουσία στην αυξανόμενη σε μήκος αλληλουχία του DNA (Εικόνα 16). Συχνά, η διαδικασία πραγματοποιείται πρώτα για το ένα άκρο του θραύσματος και στη συνέχεια για το εταίρο άκρο, με σκοπό την αύξηση της ακρίβειας της παραγόμενων δεδομένων. Σε αυτή την περίπτωση, η διαδικασία λέγεται **αλληλούχηση ζεύγους άκρων (paired end sequencing)**, έναντι της **αλληλούχησης μονού άκρου (single end sequencing)**.

Σε διαφορετικές πλατφόρμες, οι βασικές αρχές παραμένουν ίδιες, ενώ μπορεί να διαφέρουν οι τεχνολογίες που χρησιμοποιούνται. Για παράδειγμα, στην πλατφόρμα IonTorrent η ενίσχυση των θραυσμάτων DNA πραγματοποιείται στην **επιφάνεια ενός σφαιριδίου (emulsion PCR)**, ενώ η ανίχνευση την νουκλεοτιδικής ακολουθίας στηρίζεται στην ανίχνευση της αλλαγής του pH. Στο τελευταίο στάδιο ενός πειράματος Αλληλούχησης Επόμενης Γενιάς πραγματοποιείται η βιοπληροφορική ανάλυση των ακατέργαστων δεδομένων (Slatko *et al.*, 2018; Grada & Weinbrecht, 2013). Η διαδικασία αυτή παρουσιάζεται λεπτομερώς στα επόμενα κεφάλαια.



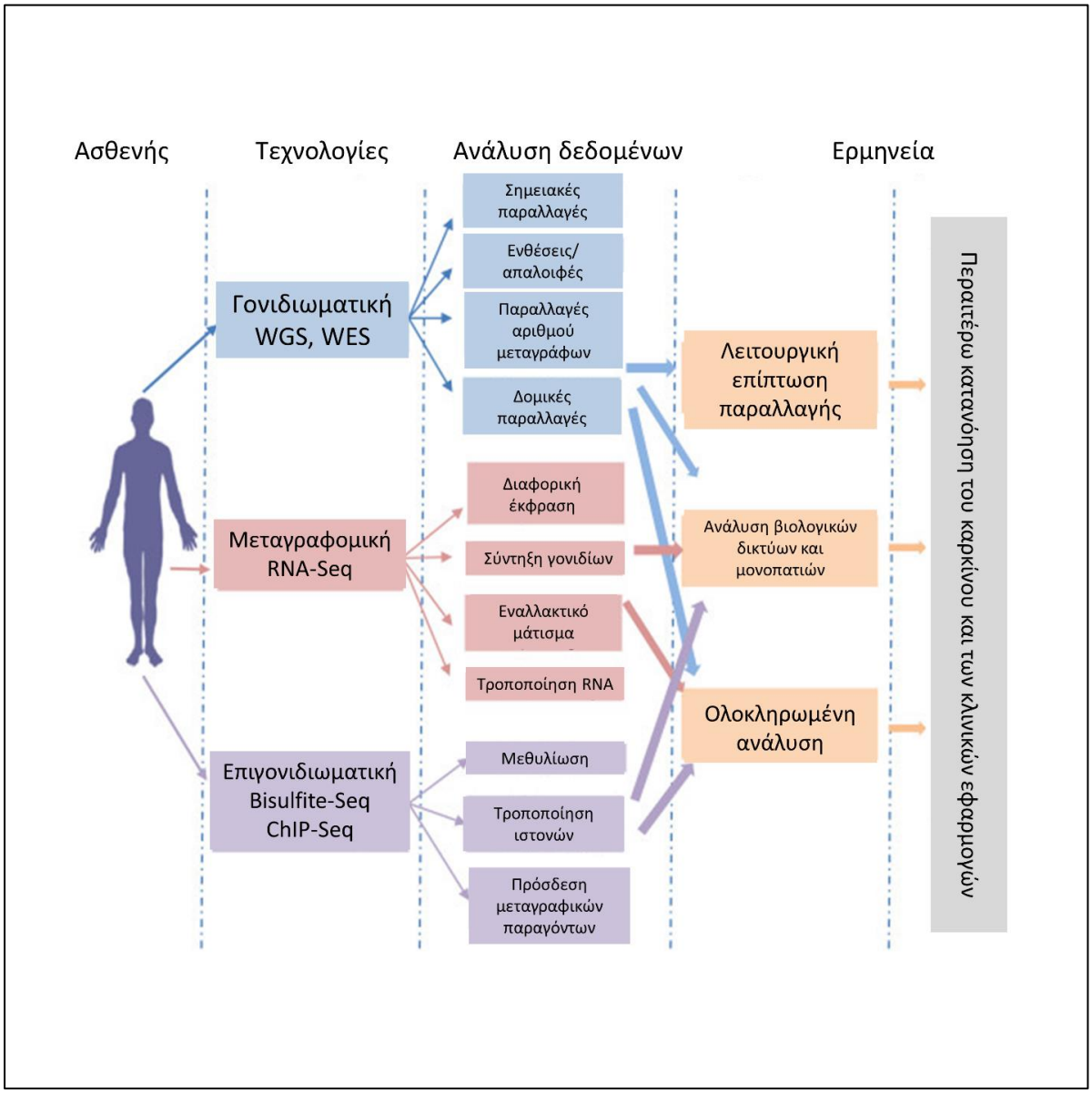
Εικόνα 16: Οι βασικές αρχές της μεθόδου Αλληλούχησης Επόμενης Γενιάς. Τροποποίηση από (Young & Gillung, 2019)

Figure 16: Basic principles of Next Generation Sequencing method. Modification from (Young & Gillung, 2019)

1.6. Εφαρμογή της βιοπληροφορικής στην Ιατρική Ακριβείας

Ως **βιοπληροφορική** ορίζεται η επιστήμη που συνδυάζει τη βιολογία, την επιστήμη υπολογιστών και τη στατιστική για την κατανόηση βιολογικών δεδομένων, με απώτερο στόχο την απάντηση βιολογικών ερωτημάτων. Η βιοπληροφορική είναι μια σχετικά νέα επιστήμη· οι απαρχές της καταγράφονται στις αρχές της δεκαετίας του 1960, οπότε εφαρμόστηκαν οι πρώτες υπολογιστικές μέθοδοι για τον προσδιορισμό της ακολουθίας πρωτεϊνών (Gauthier *et al.*, 2019). Με την έλευση της μεθόδου

Αλληλούχησης Επόμενης Γενιάς, εμφανίστηκε και η ανάγκη για ακόμα πιο αποδοτικές μεθόδους ανάλυσης αλλά και αποθήκευσης βιολογικών δεδομένων, καθώς ο όγκος των παραγόμενων δεδομένων αυξήθηκε κατακόρυφα. Είναι χαρακτηριστικό ότι το μέγεθος των ακατέργαστων αρχείων που παράγονται από την αλληλούχηση ενός μόνο γονιδιώματος ενός ανθρώπου είναι περίπου 240Gb, ενώ υπολογίζεται ότι μέχρι σήμερα έχουν αλληλουχηθεί περισσότερα από 1,5 εκατομμύρια ανθρώπινα γονιδιώματα.



Εικόνα 17: Η ροή εργασιών για την ενσωμάτωση δεδομένων στην έρευνα για τον καρκίνο και την κλινική της εφαρμογή. Τροποποίηση από (Shyr & Liu, 2013).

Figure 17: Workflow of data integration into cancer research and its clinical application. Modification from (Shyr & Liu, 2013).

Η βιοπληροφορική έχει διάφορες εφαρμογές στην Ιατρική Ακριβείας (Εικόνα 17). Σε πρώτο επίπεδο, η βιοπληροφορική επιτρέπει την ανάλυση ακατέργαστων δεδομένων μεγάλης κλίμακας, τα οποία είναι αδύνατο να αναλυθούν χειροκίνητα, με απώτερο σκοπό την αποσαφήνισή τους και την ανακάλυψη σημαντικών πληροφοριών που ενδεχομένως να αποτελούν την απάντηση σε ένα βιολογικό ερώτημα. Μέσω της ενσωμάτωσης ετερογενών δεδομένων προερχομένων από τη μελέτη διαφόρων βιολογικών αντικειμένων, όπως γονιδιωματικών, πρωτεωμικών και μεταγραφικών δεδομένων και με τη χρήση προηγμένων υπολογιστικών και στατιστικών μεθόδων, η βιοπληροφορική παρέχει τα εργαλεία για την ανίχνευση βιοδεικτών και υποψηφίων φαρμάκων. Μεγάλη είναι και η συμβολή της βιοπληροφορικής στην αποθήκευση αυτού του τεράστιου όγκου δεδομένων, μέσω των βάσεων βιολογικών δεδομένων που έχουν δημιουργηθεί, οι οποίες όχι μόνο διευκολύνουν τη διαχείριση των δεδομένων, αλλά και το διαμοιρασμό τους σε ολόκληρη την επιστημονική κοινότητα (Gomez-Lopez *et al.*, 2019).

Η παρούσα διατριβή επικεντρώνεται στην ανάλυση **γενετικών δεδομένων**, δηλαδή δεδομένων από αλληλούχηση DNA με στόχο την ανίχνευση γαμετικών παραλλαγών που ευθύνονται για την προδιάθεση σε κάποιο κληρονομικό καρκινικό σύνδρομο και σωματικών παραλλαγών που δύναται να αποτελέσουν βιοδείκτες για τη χορήγηση στοχευμένης θεραπείας.

1.6.1. Παραλλαγές στο γενετικό υλικό

Η ανάλυση δεδομένων από αλληλούχηση DNA έχει ως στόχο την ανίχνευση παραλλαγών οι οποίες εξηγούν τον φαινότυπο του ασθενούς ή μπορούν να αποτελέσουν στόχο για εξατομικευμένη θεραπεία. Ως **παραλλαγή** ορίζεται **οποιαδήποτε αλλαγή στην ακολουθία του DNA σε σχέση με την ακολουθία κάποιου γονιδιώματος αναφοράς**. Υπάρχουν διάφοροι τρόποι κατηγοριοποίησης των παραλλαγών.

Ένας πρώτος τρόπος είναι η κατηγοριοποίησή τους **σε σχέση με τη συχνότητά τους σε διάφορους πληθυσμούς**. Έτσι, υπάρχουν οι σπάνιες παραλλαγές που ανιχνεύονται σε ποσοστό μικρότερο του 1% στον πληθυσμό, και οι κοινές παραλλαγές, στις οποίες οφείλεται το 90% της διαφορετικότητας μεταξύ των ανθρώπων. Όπως αναφέρθηκε και νωρίτερα, βάσει της εξελικτικής θεωρίας, οι παραλλαγές που ευθύνονται για την προδιάθεση σε κάποιο γενετικό νόσημα, ανιχνεύονται σε πολύ μικρό ποσοστό του πληθυσμού, χωρίς αυτό να συνεπάγεται πως όλες οι σπάνιες παραλλαγές αυξάνουν τον κίνδυνο για την εμφάνιση νόσου (Karki *et al.*, 2015).

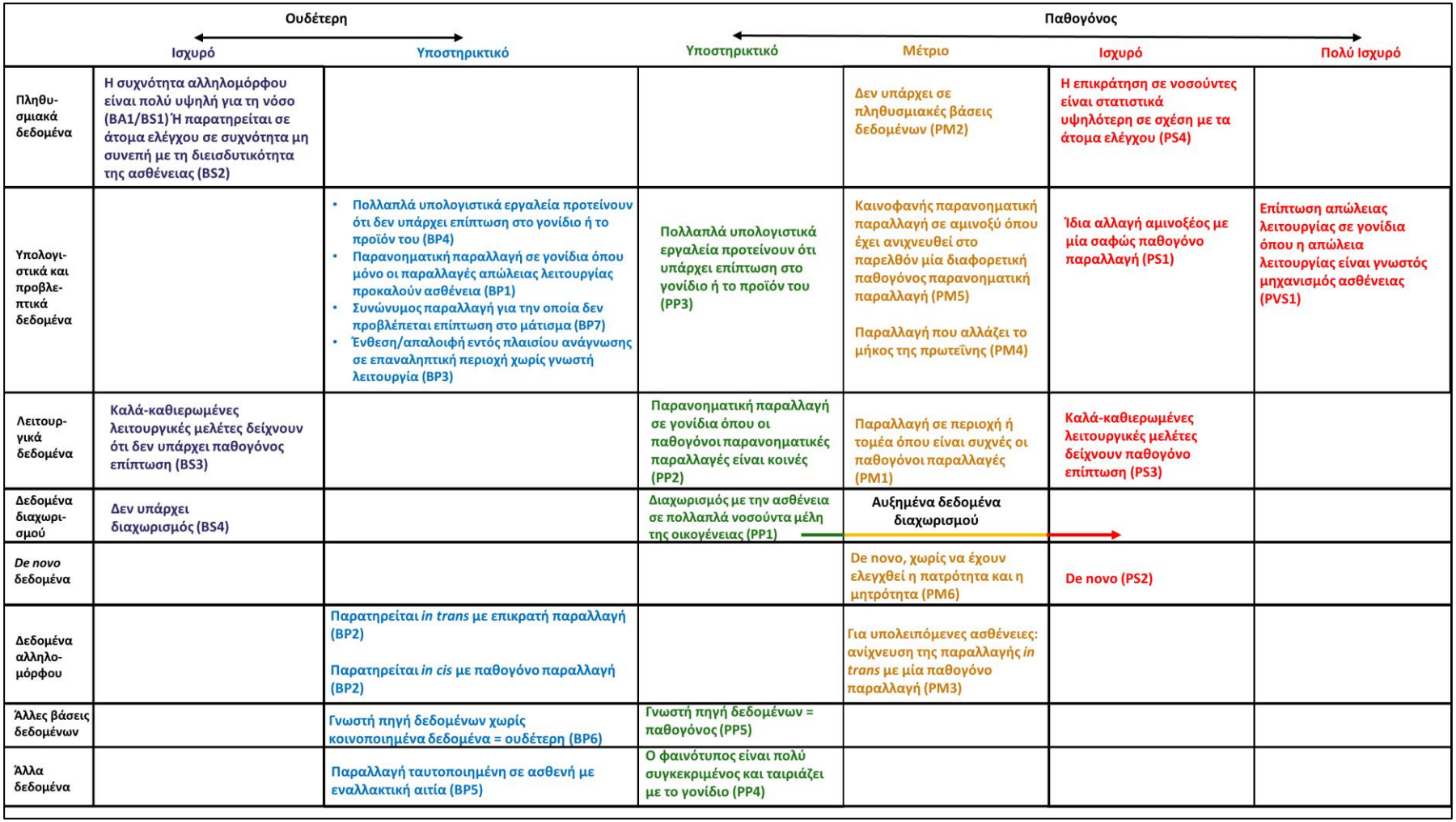
Συνεπώς, οι παραλλαγές κατηγοριοποιούνται και **ως προς την κλινική τους σημασία**. Οι οργανισμοί American College of Medical Genetics (ACMG) και Association of Molecular Pathology (AMP) έχουν θεσπίσει σαφείς κανόνες για την κατάταξη των γαμετικών παραλλαγών στις εξής πέντε κατηγορίες (Εικόνα 18) (Richards *et al.*, 2015):

1. Τις σαφώς παθογόνους παραλλαγές, για τις οποίες έχει αποδειχθεί ότι προδιαθέτουν στην εμφάνιση του υπό μελέτη φαινοτύπου.
2. Τις πιθανώς παθογόνους παραλλαγές, για τις οποίες υπάρχει ισχυρή ένδειξη ότι προδιαθέτουν στην εμφάνιση του υπό μελέτη φαινοτύπου.

3. Τις παραλλαγές αγνώστου σημασίας (Variants of Uncertain Significance – VUS), για τις οποίες δεν υπάρχουν αρκετά διαθέσιμα στοιχεία για την κατάταξή τους ως προς την κλινική σημασία τους.
4. Τις πιθανώς ουδέτερες παραλλαγές, για τις οποίες υπάρχει ισχυρή ένδειξη ότι δεν ευθύνονται για την εμφάνιση του υπό μελέτη φαινοτύπου.
5. Τις ουδέτερες παραλλαγές, για τις οποίες έχει αποδειχθεί ότι δεν ευθύνονται για την εμφάνιση του υπό μελέτη φαινοτύπου.

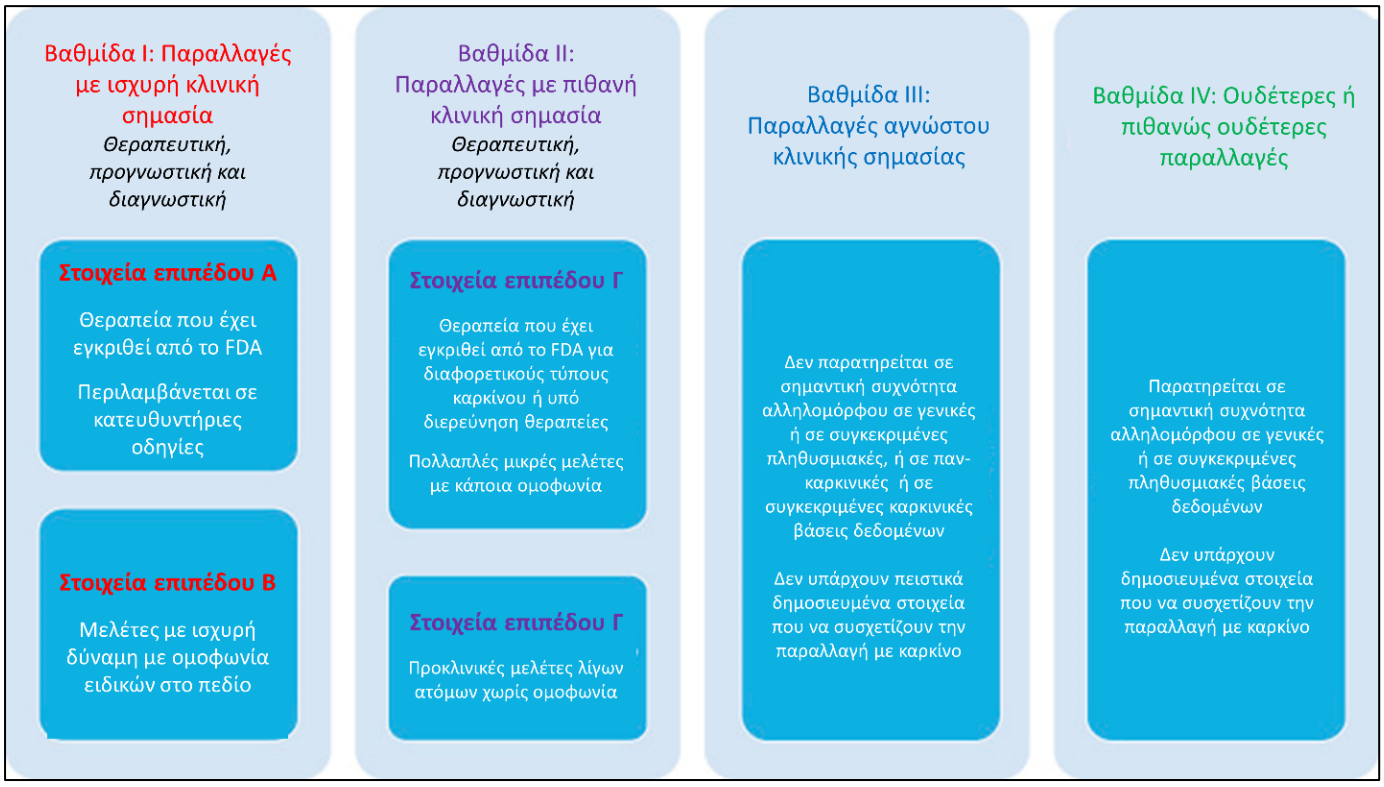
Παράλληλα, η AMP έχει θεσπίσει οδηγίες για την κατηγοριοποίηση των παραλλαγών του DNA σωματικής σειράς, **σε σχέση με την κλινική δυνατότητα δράσης (clinical actionability)** που επιφέρει η ταυτοποίησή τους, ως εξής (Εικόνα 19) (M. M. Li *et al.*, 2017):

- Βαθμίδα I: παραλλαγές για τις οποίες έχει αποδειχθεί ότι η ταυτοποίησή τους επιτρέπει τη λήψη θεραπευτικών, προγνωστικών και διαγνωστικών αποφάσεων.
- Βαθμίδα II: παραλλαγές για τις οποίες υπάρχουν ισχυρές ενδείξεις ότι η ταυτοποίησή τους επιτρέπει τη λήψη θεραπευτικών, προγνωστικών και διαγνωστικών αποφάσεων.
- Βαθμίδα III: παραλλαγές για τις οποίες δεν είναι γνωστό αν η ταυτοποίησή τους μπορεί να επηρεάσει τη λήψη θεραπευτικών, προγνωστικών και διαγνωστικών αποφάσεων.
- Βαθμίδα IV: παραλλαγές των οποίων η ταυτοποίηση δεν μπορεί να επηρεάσει τη λήψη θεραπευτικών, προγνωστικών και διαγνωστικών αποφάσεων.



Εικόνα 18: Τα κριτήρια για την κατηγοριοποίηση των γαμετικών παραλλαγών από το Αμερικανικό Κολλέγιο Ιατρικής Γενετικής και την Ένωση Μοριακής Παθολογίας. Τροποποίηση από (Richards *et al.*, 2015).

Figure 18: Criteria for the classification of germline variants as were suggested by the American College of Medical Genetics and the Molecular Pathology Association. Modification from από (Richards *et al.*, 2015).

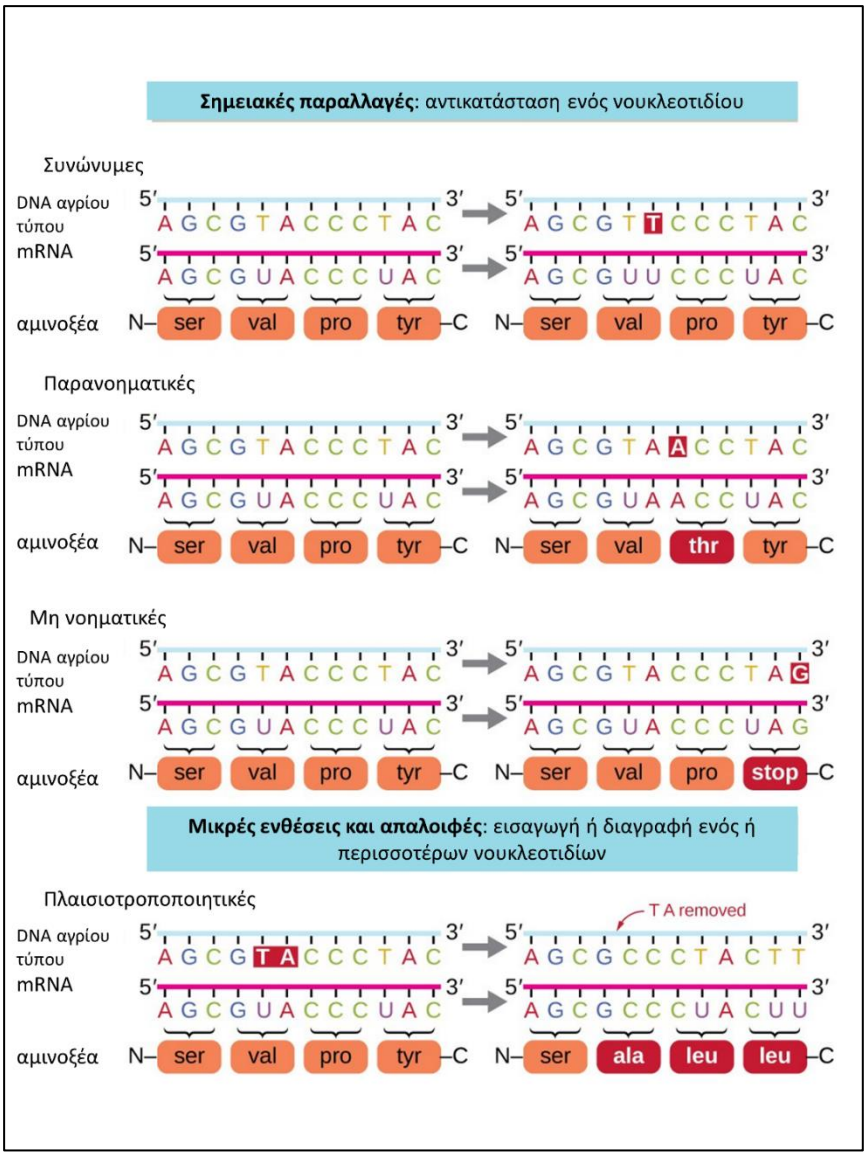


Εικόνα 19: Η κατηγοριοποίηση των σωματικών παραλλαγών από την Ένωση Μοριακής Παθολογίας. Τροποποίηση από (M. M. Li *et al.*, 2017).

Figure 19: Classification of somatic variants by the Association of Molecular Pathology. Modification from (M. M. Li *et al.*, 2017).

Οι παραλλαγές μπορούν επίσης να κατηγοριοποιηθούν **βάσει του μεγέθους τους**. Έτσι, υπάρχουν οι σημειακές παραλλαγές, οι οποίες αφορούν μερικές δεκάδες νουκλεοτιδίων και οι μεγάλες γονιδιακές αναδιατάξεις, οι οποίες συνήθως αφορούν παραλλαγές που επηρεάζουν περιοχές με μήκος μεγαλύτερο των χιλίων ζευγών βάσεων. Στην παρούσα διατριβή, μελετώνται οι σημειακές παραλλαγές.

Ανάλογα τη θέση τους στο γονιδίωμα, οι παραλλαγές μπορούν να χαρακτηριστούν ως εξονικές, ιντρονικές και διαγονιδιακές, δηλαδή παραλλαγές που εδράζονται στις κωδικές περιοχές των γονιδίων, παραλλαγές που εδράζονται εντός κάποιου γονιδίου, αλλά μεταξύ των κωδικών περιοχών, και παραλλαγές που εδράζονται σε περιοχή μεταξύ δύο γονιδίων. Παρόλο που οι κωδικές περιοχές αποτελούν μόλις το 1%-2% του γονιδιώματος, υπολογίζεται ότι περίπου το ~85% των παραλλαγών που ευθύνονται για την εμφάνιση κάποιας γενετικής ασθένειας βρίσκονται σε αυτές τις περιοχές (Gilissen *et al.*, 2012).



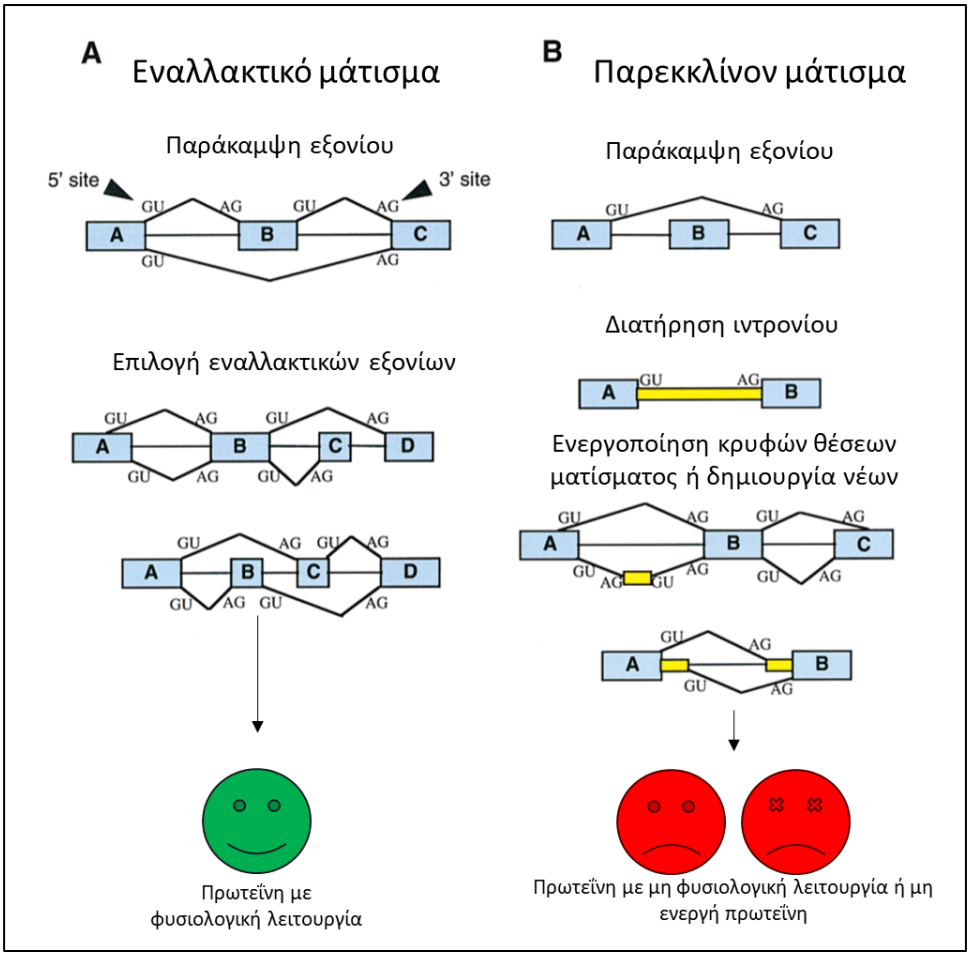
Εικόνα 20: Κατηγορίες παραλλαγών βάσει της επίπτωσης που έχουν στην πρωτεϊνική ακολουθία. Τροποποίηση από (Parker *et al.*, 2019).

Figure 20: Variant categories based on their effect on the protein sequence. Modification from (Parker *et al.*, 2019).

Οι παραλλαγές που εδράζονται σε κωδικές περιοχές μπορεί να είναι συνώνυμες, παρανοηματικές, μη νοηματικές ή μικρές ενθέσεις και απαλοιφές. Οι συνώνυμες παραλλαγές δεν επηρεάζουν την πρωτεϊνική ακολουθία, δηλαδή όταν μεταφράζεται σε πρωτεΐνη η νέα ακολουθία RNA που προκύπτει, η αμινοξική ακολουθία της παραγόμενης πρωτεΐνης παραμένει αναλλοίωτη. Οι παρανοηματικές παραλλαγές έχουν ως αποτέλεσμα την αλλαγή ενός αμινοξέος στην παραγόμενη πρωτεϊνική ακολουθία. Οι μη νοηματικές παραλλαγές εισάγουν ένα πρόωρο κωδικόνιο τερματισμού στην νουκλεοτιδική ακολουθία, με αποτέλεσμα την παραγωγή μιας κολοβωμένης πρωτεΐνης. Τέλος, οι μικρές ενθέσεις και απαλοιφές βάσεων εισάγουν ή διαγράφουν μία ή περισσότερες βάσεις από την νουκλεοτιδική ακολουθία κι έχουν δύο πιθανές επιπτώσεις στην πρωτεϊνική. Στην περίπτωση που ο αριθμός των

βάσεων που εισάγονται ή διαγράφονται είναι πολλαπλάσιος του τρία, τότε οι παραλλαγές αυτές είναι εντός πλαισίου ανάγνωσης και απλά προστίθεται ή αφαιρείται ο αντίστοιχος αριθμός αμινοξέων από την πρωτεΐνη. Στην περίπτωση που ο αριθμός των βάσεων που εισάγονται ή διαγράφονται δεν είναι πολλαπλάσιος του τρία, τότε οι παραλλαγές ονομάζονται πλαισιοτροποποιητικές, και έχουν ως αποτέλεσμα ένα διαφορετικό προϊόν από την αρχική πρωτεΐνη, του οποίου η ακολουθία και η λειτουργία δεν είναι δυνατό να είναι γνωστές (Εικόνα 20).

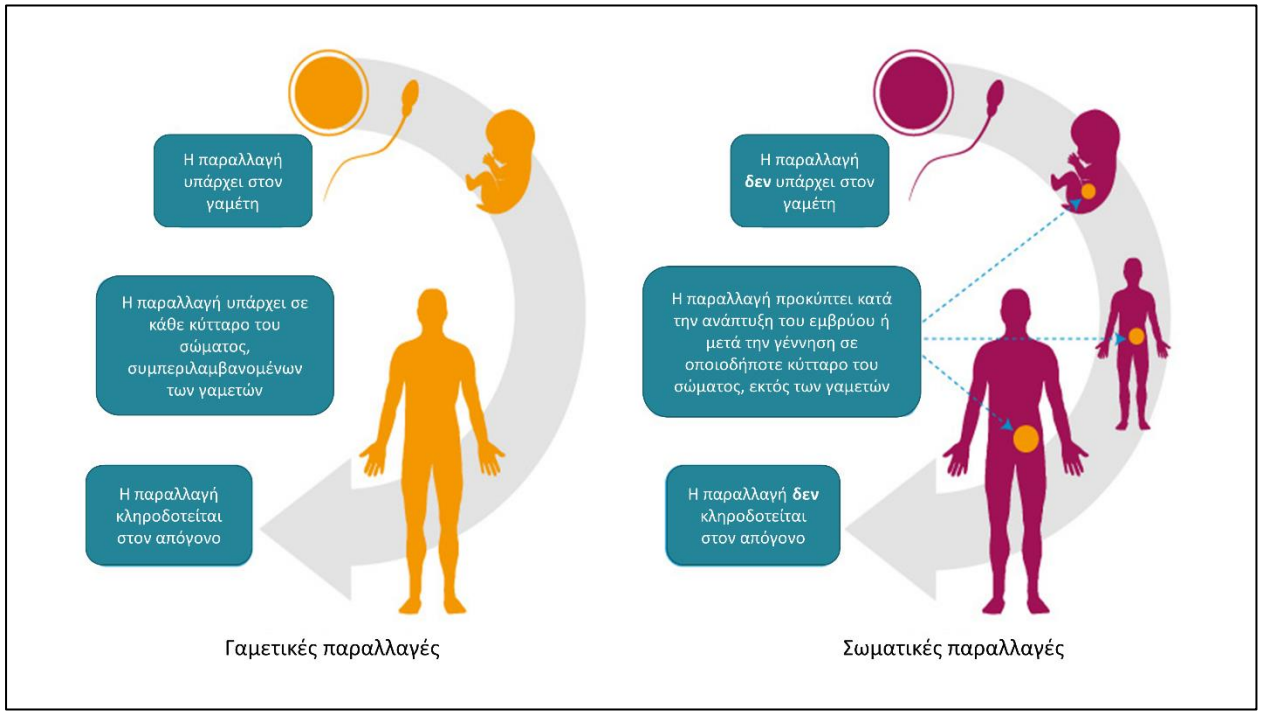
Από τις παραλλαγές που εδράζονται σε ιντρόνια, κλινικό ενδιαφέρον έχουν συνήθως αυτές που βρίσκονται κοντά ή πάνω στις θέσεις ματίσματος, γιατί είναι πιθανό να έχουν ως αποτέλεσμα παρεκκλίνον μάτισμα, δηλαδή τη διατήρηση κάποιου ιντρονίου ή την παράκαμψη κάποιου εξονίου κατά το μάτισμα, και τελικά την παραγωγή πρωτεΐνης με μη φυσιολογική λειτουργία ή ακόμα και μη ενεργής πρωτεΐνης. Το παρεκκλίνον μάτισμα δε θα πρέπει να συγχέεται με το εναλλακτικό μάτισμα, που αποτελεί μια φυσιολογική διαδικασία, η οποία επιτρέπει στο mRNA να κατευθύνει τη σύνθεση διαφορετικών πρωτεϊνικών ισομορφών που μπορεί να έχουν διαφορετικές κυτταρικές λειτουργίες (Εικόνα 21) .



Εικόνα 21: Σχηματική αναπαράσταση φυσιολογικού και παρεκκλίνοντος εναλλακτικού ματίσματος mRNA. Αρχική εικόνα από (Bai & Lipton, 1998).

Figure 21: Normal and aberrant alternative mRNA splicing. Modification from (Bai & Lipton, 1998).

Τέλος, οι παραλλαγές διακρίνονται σε **γαμετικές** και **σωματικές**. Οι γαμετικές παραλλαγές είναι οι παραλλαγές που βρίσκονται στους γαμέτες των γονέων και με αυτόν τον τρόπο κληροδοτούνται στους απογόνους και ενσωματώνονται σε όλα τα κύτταρά τους. Οι γενετικές ασθένειες οφείλονται σε γαμετικές παραλλαγές. Οι σωματικές παραλλαγές είναι επίκτητες παραλλαγές που εμφανίζονται κατά τη διάρκεια της ζωής ενός ατόμου. Οι σωματικές παραλλαγές δε βρίσκονται σε όλα τα κύτταρα του σώματος ενός ανθρώπου, παρά μόνο στο αρχικό κύτταρο όπου προκλήθηκαν και στους απογόνους του, και δεν κληροδοτούνται στους απογόνους του ατόμου που τις φέρει (Griffiths *et al.*, 2000). Το γονιδίωμα του όγκου χαρακτηρίζεται από συσσώρευση σωματικών παραλλαγών και η μελέτη του αποσκοπεί στην ανίχνευση παραλλαγών που ευθύνονται για την εξέλιξη του όγκου ή/και μπορούν να αποτελέσουν στόχους για εξατομικευμένη θεραπεία (Εικόνα 22).



Εικόνα 22: Παραλλαγές DNA γαμετικής και σωματικής σειράς. Τροποποίηση από (Genomics Education Program, 2021).

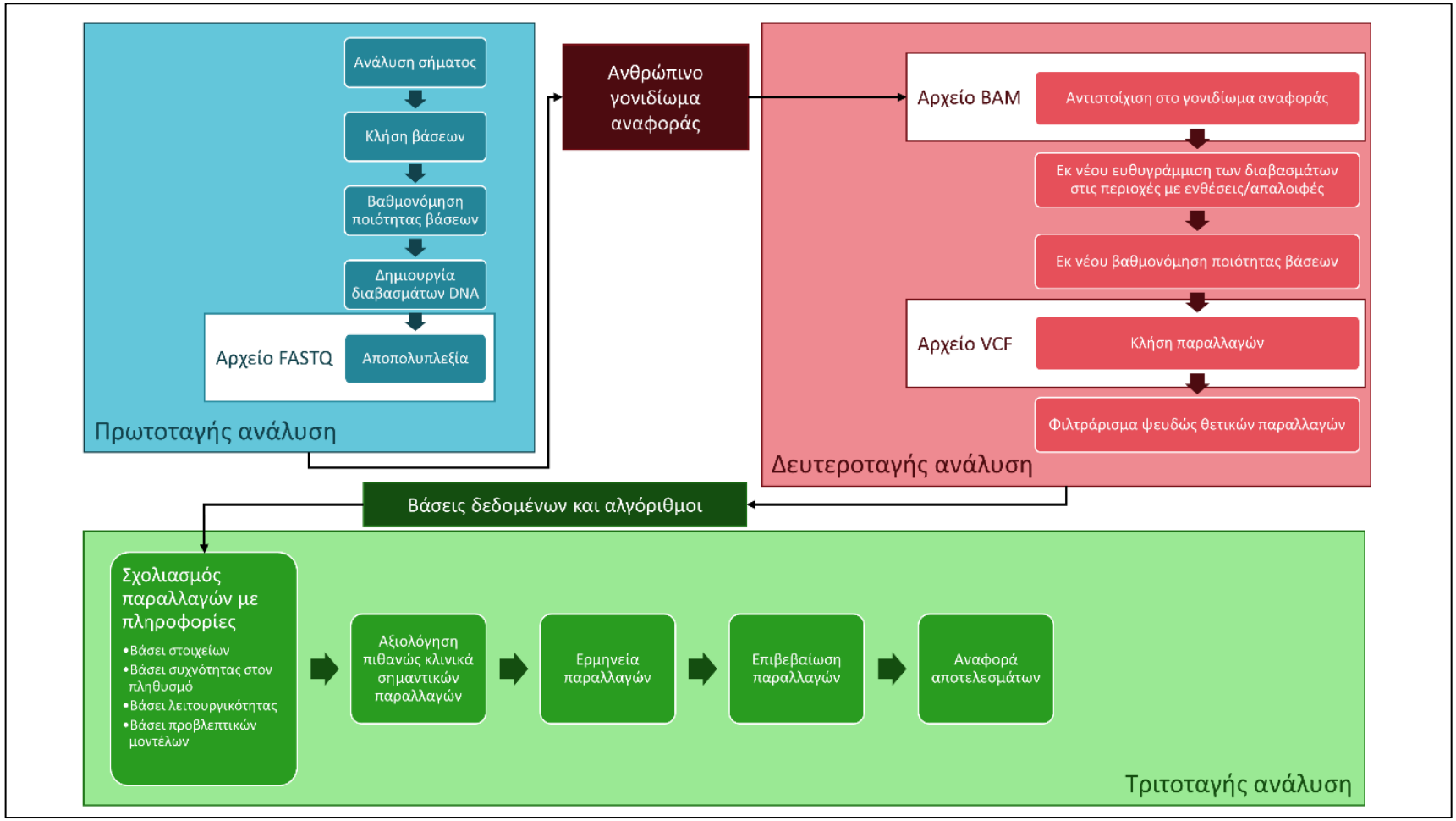
Figure 22: Germline and somatic DNA variants. Modification from (Genomics Education Program, 2021).

1.6.2. Ανάλυση δεδομένων που προέρχονται από μαζική παράλληλη αλληλούχηση DNA

Η ανάλυση των δεδομένων που προέρχονται από μαζική παράλληλη αλληλούχηση ανθρώπινου DNA αποσκοπεί στην ανίχνευση αλλαγών στο DNA του ατόμου σε σχέση με το ανθρώπινο γονιδίωμα αναφοράς. Ειδικότερα, στην περίπτωση που μελετάται κάποια Μενδελική νόσος, όπως ο κληρονομικός

καρκίνος, οι παραλλαγές που μελετώνται είναι εξαιρετικά σπάνιες και συνήθως έχουν συχνότητα αλληλομόρφου στον γενικό πληθυσμό μικρότερη του 1% (Karki *et al.*, 2015).

Η ανάλυση των δεδομένων που έχουν παραχθεί με την τεχνολογία Αλληλούχησης Επόμενης Γενιάς είναι μια μεταβλητή διαδικασία που προσαρμόζεται κάθε φορά στη φύση του πειράματος. Τα επιμέρους βήματα της ανάλυσης δεδομένων εκτελούνται από διαφορετικές διεργασίες. Κατά τη διάρκεια αυτής της διαδικασίας, η έξοδος μιας διεργασίας συνήθως συνδέεται με την είσοδο της επόμενης σε σειρά διεργασίας, δημιουργώντας έτσι μια **ροή διοχέτευσης εντολών διεργασιών (pipeline).** Κάθε ροή διοχέτευσης εντολών διεργασιών μπορεί να είναι μοναδική, αναλόγως τις ανάγκες της ανάλυσης, ωστόσο, υπάρχουν κάποιες διεργασίες οι οποίες χρειάζεται να πραγματοποιηθούν κάθε φορά και οι οποίες αναλύονται περαιτέρω στα επόμενα κεφάλαια (Εικόνα 23).



Εικόνα 23: Ροή εργασιών για την ανάλυση δεδομένων από αλληλούχηση DNA με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς με κλινική εφαρμογή. Αρχική εικόνα από (Oliver *et al.*, 2015).

Figure 23: Workflow for Next Generation Sequencing of DNA data analysis with clinical application. Original image by (Oliver *et al.*, 2015).

1.6.2.1. Πρωτοταγής ανάλυση

Η **πρωτοταγής ανάλυση** των δεδομένων πραγματοποιείται κατά κύριο λόγο στον γενετικό αναλυτή επόμενης γενιάς από το ενσωματωμένο λογισμικό, το οποίο παρέχεται από όλες τις μεγάλες εταιρείες που εμπορεύονται συστήματα και αντιδραστήρια αλληλούχησης. Ωστόσο, το ειδικό αυτό λογισμικό μπορεί επίσης να εγκατασταθεί σε υπολογιστικά συστήματα υψηλής απόδοσης ή αρχιτεκτονικές που βασίζονται σε σύννεφο για βελτιωμένη απόδοση ή αναλύσεις κατ' εξακολούθηση (Oliver *et al.*, 2015).

Εικόνα 24: Παράδειγμα αρχείου FASTQ. Σε κάθε αναγνωσμένη ακολουθία αντιστοιχούν τέσσερις σειρές. Στην πρώτη σειρά καταγράφεται το αναγνωριστικό της ακολουθίας, στη δεύτερη σειρά καταγράφεται η αναγνωσμένη ακολουθία, στην τρίτη σειρά καταγράφονται προαιρετικά μεταδεδομένα και στην τέταρτη σειρά καταγράφονται τα σύμβολα ASCII που αντιστοιχούν στο βαθμό ποιότητας κάθε βάσης.

Figure 24: Example of a FASTQ file. Four rows are reserved for each read. The first row records the identifier of the read, the second row records the read sequence, the third row records metadata optionally and the fourth row records the ASCII symbols corresponding to the quality score for each nucleotide read.

Κάθε φορά που διαβάζεται μία βάση από τον αναλυτή, εκπέμπεται ένα σήμα. Κατά την πρωτοταγή ανάλυση, το λογισμικό του αναλυτή μεταφράζει αυτό το σήμα στο νουκλεοτίδιο στο οποίο αντιστοιχεί, υπολογίζοντας παράλληλα την πιθανότητα αυτό το σήμα να έχει διαβαστεί και μεταφραστεί σωστά. Στη συνέχεια, σε κάθε νουκλεοτίδιο δίνεται ένας **βαθμός ποιότητας (quality score)** βάσει αυτής της πιθανότητας. Η βαθμολογία ποιότητας βασίζεται στην **κλίμακα Phred** και υπολογίζεται από τον εξής τύπο:

$$Q = -10\, log_{10}\, P_{err}$$

Όπου Q είναι ο βαθμός ποιότητας και P_{err} είναι η πιθανότητα η βάση στην οποία αντιστοιχεί ο βαθμός ποιότητας να έχει διαβαστεί λάθος. Αυτό πρακτικά σημαίνει πως αν ο βαθμός ποιότητας Q είναι 0, τότε το P_{err} θα είναι 1, δηλαδή η βάση θα έχει σίγουρα διαβαστεί λάθος, ενώ αν ο βαθμός ποιότητας Q είναι 40, τότε το P_{err} θα είναι 0,0001, δηλαδή η πιθανότητα να έχει διαβαστεί λάθος η βάση είναι μία στις 10.000.

Πίνακας 3: Αντιστοίχιση ASCII χαρακτήρων και κλίμακα βαθμολογίας Phred για τις πλατφόρμες αλληλούχησης της Illumina.

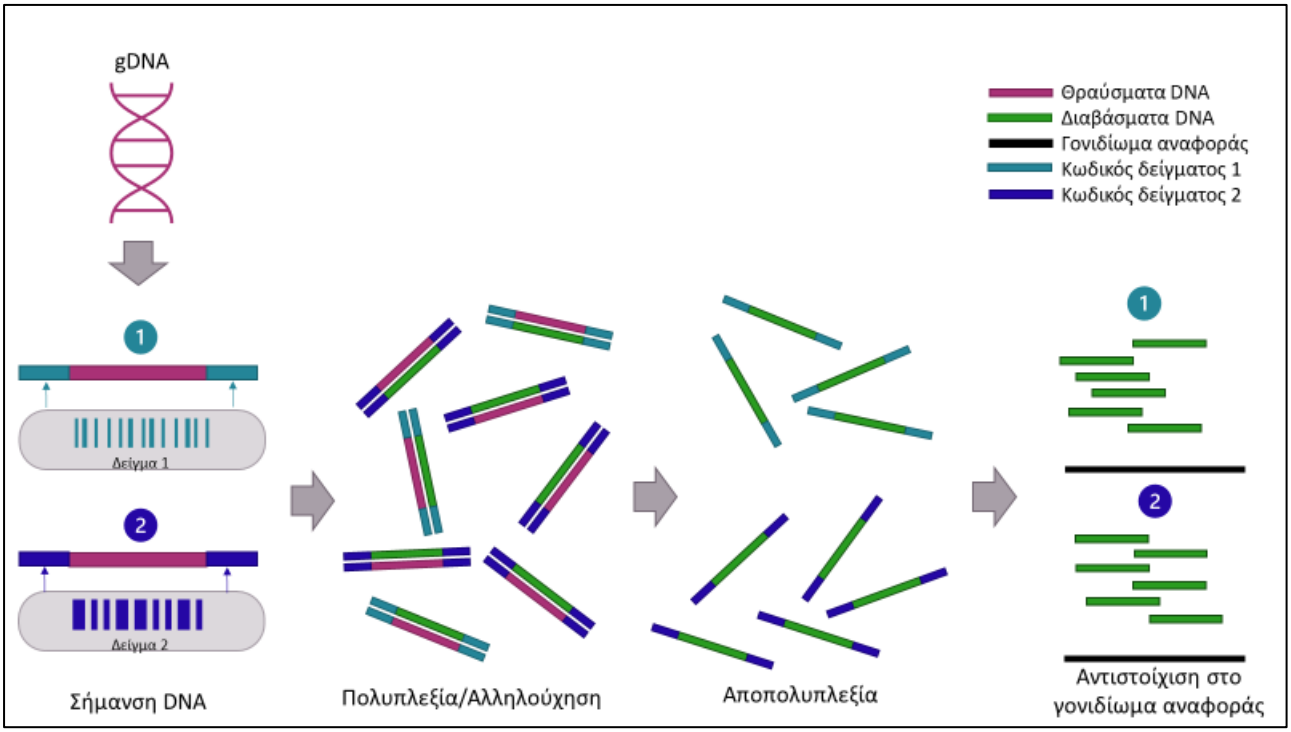
Table 3: Phred scoring system and their corresponding ASCII characters for Illumina sequencing platforms.

Σύμβολο	Χαρακτήρας ASCII	Βαθμός ποιότητας	Σύμβολο	Χαρακτήρας ASCII	Βαθμός ποιότητας
!	33	0	6	54	21
"	34	1	7	55	22
#	35	2	8	56	23
$	36	3	9	57	24
%	37	4	:	58	25
&	38	5	;	59	26
'	39	6	<	60	27
(	40	7	=	61	28
)	41	8	>	62	29
*	42	9	?	63	30
+	43	10	@	64	31
’	44	11	A	65	32
-	45	12	B	66	33
.	46	13	C	67	34
/	47	14	D	68	35
0	48	15	E	69	36
1	49	16	F	70	37
2	50	17	G	71	38
3	51	18	H	72	39
4	52	19	I	73	40
5	53	20			

Κατά το επόμενο στάδιο της ανάλυσης, θα πρέπει οι επιμέρους βάσεις, να συνενωθούν μεταξύ τους με τη σειρά που έχουν διαβαστεί, ώστε να δημιουργηθούν οι αναγνώσεις DNA. Στη συνέχεια, οι αναγνώσεις, μαζί με τους βαθμούς ποιότητας που αντιστοιχούν σε κάθε νουκλεοτίδιο, αποθηκεύονται στο **αρχείο FASTQ**. Παράλληλα, οι βαθμολογίες Phred κωδικοποιούνται στο αρχείο FASTQ ως σύμβολα ASCII, ώστε να είναι μονοί χαρακτήρες κι έτσι να πραγματοποιείται πιο εύκολα η αντιστοίχιση των αναγνωσμένων νουκλεοτιδίων και των αντίστοιχων βαθμολογιών τους (Εικόνα 24). Κάθε τεχνολογία

Αλληλούχησης Επόμενης Γενιάς χρησιμοποιεί διαφορετικούς χαρακτήρες ASCII και εύρος τιμών για τη βαθμολογία ποιότητας. Στον Πίνακα 3 παρουσιάζεται η αντιστοίχιση των βαθμών ποιότητας και των χαρακτήρων ASCII για τις πλατφόρμες αλληλούχησης της Illumina.

Στην περίπτωση όπου το πείραμα περιλαμβάνει περισσότερα από ένα δείγματα, η πρωτοταγής ανάλυση περιλαμβάνει και την **αποπολυπλεξία των δειγμάτων (demultiplexing)**. Σε ένα πείραμα με πολυπλεξία (multiplexing), κατά τη διάρκεια της προετοιμασίας των βιολογικών βιβλιοθηκών, στα επιμέρους θραύσματα DNA θα προστεθεί ένας **μοριακός κωδικός (molecular barcode)** ώστε να μπορεί να αναγνωριστεί μετά την αλληλούχηση το δείγμα από όπου προέρχεται το αναγνωσθέν DNA. Η διαδικασία της αποπολυπλεξίας έχει ως αποτέλεσμα την αποθήκευση της κάθε ανάγνωσης DNA σε ξεχωριστό αρχείο ανά δείγμα, ανάλογα τον μοριακό κωδικό του (Εικόνα 25).



Εικόνα 25: Η διαδικασία της πολυπλεξίας στην Αλληλούχηση Επόμενης Γενιάς.

Figure 25: Multiplexing in Next Generation Sequencing.

1.6.2.2. Δευτεροταγής ανάλυση

Η **δευτεροταγής ανάλυση** αποτελεί και το μεγαλύτερο μέρος μιας ροής διοχέτευσης εντολών διεργασιών. Στην περίπτωση της ανάλυσης δεδομένων προερχόμενων από αλληλούχηση DNA, σκοπός της δευτεροταγούς ανάλυσης είναι ο εντοπισμών γενετικών παραλλαγών σε σχέση με κάποιο γονιδίωμα αναφοράς, κι εν προκειμένω, το ανθρώπινο γονιδίωμα αναφοράς (Oliver *et al.*, 2015). Οι παραλλαγές που μπορούν να ανιχνευθούν περιλαμβάνουν σημειακές αλλαγές (point variants), μικρές ενθέσεις και απαλοιφές (small insertions and deletions – indels) οι οποίες συνήθως αφορούν έως 100 βάσεις, καθώς και μεγάλες γονιδιωματικές αναδιατάξεις (large genomic rearrangements) και αλλαγές αριθμού αντιγράφων (copy number variants).

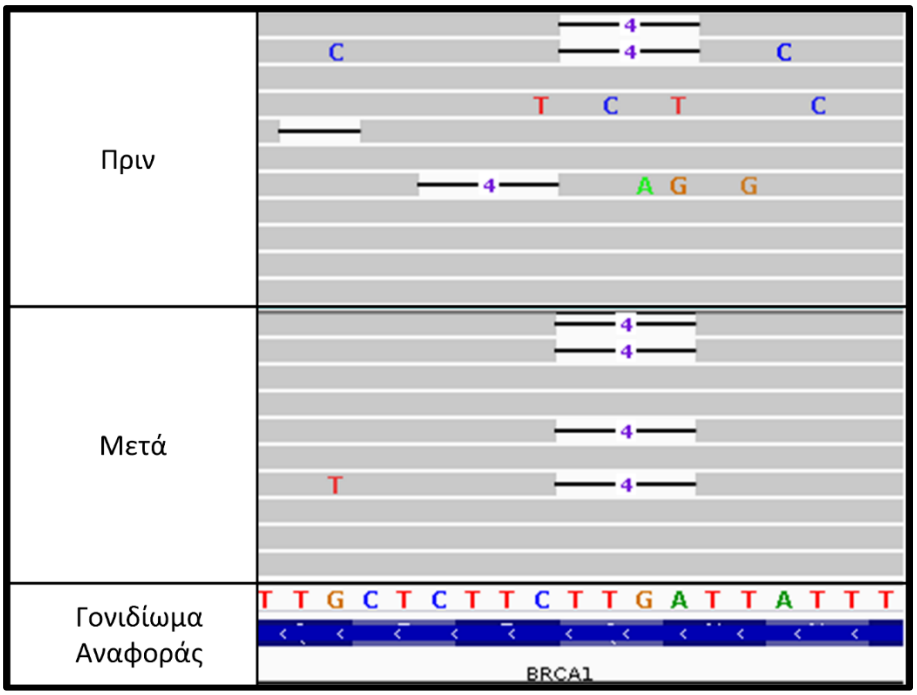
Το πρώτο στάδιο της δευτεροταγούς ανάλυσης περιλαμβάνει την **αντιστοίχιση** (ή ευθυγράμμιση· αγγλ.: sequence alignment) των αναγνώσεων DNA, που παρήχθησαν από την πρωτοταγή ανάλυση, στο ανθρώπινο γονιδίωμα αναφοράς (Oliver *et al.*, 2015). Λόγω του τεράστιου μεγέθους του ανθρωπίνου γονιδιώματος, το οποίο αποτελείται από περίπου τρία δισεκατομμύρια ζεύγη βάσεων, η διαδικασία της αντιστοίχισης είναι αρκετά περίπλοκη και χρονοβόρα. Για τον λόγο αυτό, η συνήθης πρακτική είναι η **ευρετηρίαση (indexing) του γονιδιώματος αναφοράς** πριν την αντιστοίχιση των αναγνώσεων. Η ευρετηρίαση του γονιδιώματος αναφοράς είναι μια διαδικασία που πραγματοποιείται μόνο μία φορά και το αρχείο εξόδου είναι το ευρετήριο του γονιδιώματος αναφοράς, το οποίο θα μπορούσε να παρομοιαστεί με έναν τηλεφωνικό κατάλογο που περιέχει όλες τις υποακολουθίες του γονιδιώματος αναφοράς σε αλφαβητική σειρά αντιστοιχισμένες με την θέση τους στο γονιδίωμα. Στον Πίνακα 4, αναφέρονται κάποια προγραμματιστικά εργαλεία που χρησιμοποιούνται ευρέως για την αντιστοίχιση αναγνώσεων DNA σε γονιδίωμα αναφοράς.

Πίνακας 4: Προγραμματιστικά εργαλεία για την αντιστοίχιση αναγνώσεων DNA σε κάποιο γονιδίωμα αναφοράς. Τροποποίηση από (H. Li & Homer, 2010).

Table 4: Programming tools for DNA read mapping to a reference genome. Modification from (H. Li & Homer, 2010).

Προγραμματιστικό εργαλείο	**Αλγόριθμος**	**Αντιστοίχιση με κενά**	**Αλληλούχηση ζεύγους άκρων**	**Ποιότητα**
Bfast	Hashing	Ναι	Ναι	Όχι
Bowtie	FM-index	Όχι	Ναι	Ναι
BWA	FM-index	Ναι	Ναι	Όχι
MAQ	Hashing	Ναι	Ναι	Ναι
Mosaik	Hashing	Ναι	Ναι	Όχι
Novoaligng	Hashing	Ναι	Ναι	Ναι

Την αντιστοίχιση των αναγνώσεων DNA συνήθως διαδέχονται κάποια βήματα βελτίωσης της αντιστοίχισης (DePristo *et al.*, 2011). Η πρώτη διεργασία που μπορεί να εκτελεστεί είναι η επισήμανση ή ακόμα και το φιλτράρισμα των διπλότυπων αναγνώσεων, μιας και συνήθως αποτελούν τεχνουργήματα που έχουν εισαχθεί από την αλυσιδωτή αντίδραση πολυμεράσης που πραγματοποιείται κατά την προετοιμασία των βιβλιοθηκών. Επιπλέον, συνήθως πραγματοποιείται μία **εκ νέου αντιστοίχιση (realignment)** σε περιοχές που παρουσιάζουν έντονη ποικιλομορφία σε σχέση με το γονιδίωμα αναφοράς, καθώς αυτές οι περιοχές είναι πιθανές τοποθεσίες μικρών ενθέσεων/απαλοιφών (Εικόνα 26). Τέλος, πραγματοποιείται η **εκ νέου βαθμονόμηση της ποιότητας των βάσεων (base recalibration)**, βάσει των δεδομένων ευθυγράμμισης που έχουν προκύψει από την έως τώρα ανάλυση. Οι αναγνώσεις DNA και η αντιστοίχισή τους στο γονιδίωμα αναφοράς, καθώς και διάφορες ποιοτικές πληροφορίες αποθηκεύονται στο **αρχείο SAM** (Sequence Alignment/Mapping).



Εικόνα 26: Παράδειγμα εκ νέου τοπικής αντιστοίχισης σε περιοχή του γονιδιώματος που παρουσιάζει έντονη ποικιλομορφία σε σχέση με το γονιδίωμα αναφοράς. Οι απαλοιφές σε τρεις διαφορετικές θέσεις, οι οποίες αντιπροσωπεύονται από τις μαύρες οριζόντιες γραμμές, ευθυγραμμίζονται μετά την εκ νέου αντιστοίχιση. Οπτικοποίηση στο λογισμικό Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013).

Figure 26: Example of local realignment in a genomic region that is highly diverse relative to the reference genome. Deletions in three different positions, represented by the black horizontal lines, are aligned after realignment. Visualization in Integrative Genomics Viewer (IGV) (Thorvaldsdottir *et al.*, 2013).

Το τελικό στάδιο της δευτεροταγούς ανάλυσης περιλαμβάνει την **κλήση των παραλλαγών (variant calling)**, η οποία αφορά τη διαδοχική σύγκριση των αναγνώσεων DNA με την τοποθεσία του ανθρώπινου γονιδιώματος στην οποία έχουν αντιστοιχηθεί (Oliver *et al.*, 2015). Στη συνέχεια, μετά την εφαρμογή τεχνικών στατιστικής μοντελοποίησης, ταυτοποιούνται οι περιοχές που φέρουν παραλλαγή. Οι παραλλαγές, μαζί με διάφορες πληροφορίες ποιότητας, οι οποίες χρησιμεύουν στο φιλτράρισμα των ψευδώς θετικών ευρημάτων, αποθηκεύονται στο **αρχείο VCF** (Variant Calling Format). Ανάλογα το είδος των παραλλαγών που αναζητούνται υπάρχουν και διαφορετικοί αλγόριθμοι. Έτσι, υπάρχουν διαφορετικά προγραμματιστικά εργαλεία για την ανίχνευση σημειακών παραλλαγών και μικρών ενθέσεων/απαλοιφών και διαφορετικά εργαλεία για την ανίχνευση μεγάλων γονιδιωματικών αναδιατάξεων. Επιπλέον, υπάρχει διάκριση και μεταξύ των εργαλείων που ανιχνεύουν παραλλαγές σε DNA γαμετικής σειράς και εργαλείων που ανιχνεύουν παραλλαγές σε DNA σωματικής σειράς, όπως είναι το γονιδίωμα του όγκου, καθώς η τελευταία κατηγορία φέρει εξαιρετικές προκλήσεις.

Η τεχνολογία Αλληλούχησης Επόμενης Γενιάς είναι λιγότερο ακριβής από τις καθιερωμένες μεθόδους αλληλούχησης, ωστόσο με την πάροδο του χρόνου γίνεται όλο και πιο αξιόπιστη. Για το λόγο αυτό, αναπόφευκτα ανιχνεύονται λανθασμένες κλήσεις παραλλαγών ή αλλιώς **ψευδώς θετικά**

ευρήματα. Για την αποφυγή τέτοιων ευρημάτων, οι αλγόριθμοι κλήσεως παραλλαγών αποθηκεύουν την εκάστοτε παραλλαγή στο τελικό αρχείο VCF, συνοδεία μιας πληθώρας ποιοτικών χαρακτηριστικών που βοηθούν στην αξιολόγησή της. Έτσι, με την ολοκλήρωση της κλήσης των παραλλαγών, το αρχείο VCF φιλτράρεται ή ιεραρχείται βάσει του επιπέδου εμπιστοσύνης της παραλλαγής. Τα κριτήρια για το **φιλτράρισμα ή την ιεράρχηση των παραλλαγών** ποικίλουν ανάλογα την εφαρμογή. Κάποια τυπικά κριτήρια που εξετάζονται είναι το βάθος ανάγνωσης στην περιοχή της παραλλαγής, η συχνότητα του αλληλομόρφου παραλλαγής (Variant Allele Frequency – VAF) στις αναγνώσεις DNA της περιοχής της παραλλαγής, η ποιότητα κλήσης της παραλλαγής (Variant Quality) και η ποιότητα της αντιστοίχισης των αναγνώσεων DNA στην περιοχή (Mapping Quality). Αξίζει να σημειωθεί πως πολλές ψευδείς παραλλαγές αποτελούν τεχνουργήματα της πλατφόρμας αλληλούχησης του εκάστοτε εργαστηρίου και ενδεχομένως πληρούν όλα τα κριτήρια για να θεωρηθούν πραγματικές (Glenn, 2011). Μία καλή πρακτική είναι η διατήρηση μίας τοπικής βάσης δεδομένων για την αποθήκευση επαναλαμβανόμενων ψευδώς θετικών ευρημάτων. Βεβαίως, ένας από τους πιο καθοριστικούς παράγοντες για την ανίχνευση των πραγματικών παραλλαγών είναι η εμπειρία κι η εξειδίκευση του ατόμου που πραγματοποιεί την ανάλυση.

1.6.2.3. Τριτοταγής ανάλυση

Η ανάλυση δεδομένων Αλληλούχησης Επόμενης Γενιάς ολοκληρώνεται με την **τριτοταγή ανάλυση**, όπου οι παραλλαγές που έχουν ανιχνευθεί κι έχουν επισημανθεί ως πραγματικές **εμπλουτίζονται με διάφορες πληροφορίες** (αλλιώς: χαρακτηρισμός παραλλαγών· αγγλ.: variant annotation) από δημόσιες βάσεις δεδομένων και *in silico* εργαλεία με σκοπό να προσδιοριστεί η βιολογική τους σημασία και να καταστεί δυνατή η λειτουργική τους ιεράρχηση και η μεταγενέστερη ερμηνεία τους (Oliver *et al.*, 2015). Ο εμπλουτισμός των παραλλαγών αφορά πληροφορίες όπως η επίπτωση που η εκάστοτε παραλλαγή έχει στην ακολουθία του γονιδίου, η ονοματολογία της παραλλαγής βάσει των κανόνων που έχουν οριστεί από την Εταιρία Παραλλαγών στο Ανθρώπινο Γονιδίωμα (Human Genome Variation Society – HGVS) (den Dunnen *et al.*, 2016), η συχνότητα του αλληλομόρφου της παραλλαγής (Allele Frequency – AF) σε διάφορους πληθυσμούς, καθώς και προβλέψεις από *in silico* εργαλεία σχετικά με την επίπτωση που ενδεχομένως έχει η παραλλαγή στη δομή και τη λειτουργικότητα της εκάστοτε πρωτεΐνης (Jalali Sefid Dashti & Gamieldien, 2017). Στον Πίνακα 5 απαριθμούνται κάποιες βασικές πηγές δεδομένων που χρησιμοποιούνται για τον εμπλουτισμό των δεδομένων Αλληλούχησης Επόμενης Γενιάς.

Πίνακας 5: Βασικές πηγές δεδομένων που χρησιμοποιούνται για τον εμπλουτισμό των δεδομένων Αλληλούχησης Επόμενης Γενιάς. Τροποποίηση από (Oliver *et al.*, 2015).

Table 5: Data sources used for Next Generation Sequencing data annotation. Original table from (Oliver *et al.*, 2015).

	Πηγή εμπλουτισμού	Περιγραφή	Σύνδεσμος
Βασισμένη στη συχνότητα στους πληθυσμούς	1000 Genomes Project	Αλληλούχηση ολόκληρου γονιδιώματος 2.500 υγιών ανθρώπων	http://www.1000genomes.org
	NHLBI Cohort	Αλληλούχηση μόνο των κωδικών περιοχών του γονιδιώματος 6.500 ασθενών με καρδιακή, πνευμονική και/ή αιματολογική ασθένεια	https://esp.gs.washington.edu/drupal/
	HapMap Project	Σύνολο δεδομένων βασισμένων σε κοινούς πολυμορφισμούς για τον καθορισμό απλοτύπων σε 270 διαφορετικούς εθνοτικά διαφορετικούς ανθρώπους	http://hapmap.ncbi.nlm.nih.gov
	gnomAD	Αλληλούχηση μόνο των κωδικών περιοχών του γονιδιώματος 125.748 και ολόκληρου γονιδιώματος 15.708 υγιών ανθρώπων	https://gnomad.broadinstitute.org/about
	FLOSSIES	Αλληλούχηση DNA γαμετικής σειράς σε 27 γονίδια που προδιαθέτουν σε καρκίνο περίπου 10.000 γυναικών, ηλικίας άνω των 70 ετών που δεν έχουν εμφανίσει ποτέ καρκίνο.	https://whi.color.com/
Βασισμένη στην ακολουθία	SnpEff	Επίπτωση της παραλλαγής στη δομή των κωδικονίων και των γονιδίων	http://snpeff.sourceforge.net/SnpEff.html
	VEP	Επίπτωση της παραλλαγής στο γονίδιο, το μετάγραφο και την αλληλουχία της πρωτεΐνης	http://www.ensembl.org/info/docs/tools/vep/index.html
Βασισμένη σε προβλέψεις	SIFT	Διατήρηση της ακολουθίας	http://sift.jcvi.org/
	POLYPHEN	Φυλογενετικά και δομικά χαρακτηριστικά	http://genetics.bwh.harvard.edu/pph/

	CONDEL	Ενσωμάτωση μετα-προβλέψεων	http://omictools.com/sequencing/genome-resequencing/driver-mutations/condel-s654.html
	MutPred	Μέθοδος πρόβλεψης τυχαίου δάσους	http://mutpred.mutdb.org/
	CADD	Μετα-πρόβλεψη και βαθμολογία εμπλουτισμού	http://cadd.gs.washington.edu
	VAAST	Φυλογενετική και βασισμένη σε ασθένειες διατήρηση της ακολουθίας	http://www.yandell-lab.org/software/vaast.html
	MutationTaster	Ενσωμάτωση μετα-δεδομένων	http://www.mutationtaster.org
	ANNOVAR	Ενσωμάτωση μετα-δεδομένων και μετα-προβλέψεων	http://www.openbioinformatics.org/annovar/
Βασισμένη σε αποδεικτικά στοιχεία	OMIM	Σχέση γονιδίου/φαινοτύπου ασθενειών	http://www.omim.org
	Leiden Open Variation Database	Κλινική σχέση γενοτύπου/φαινοτύπου	http://www.lovd.nl/3.0/home
	Human Gene Mutation Database	Βλάβες γονιδίων σε κληρονομικές νόσους στον άνθρωπο	http://www.hgmd.org
	ClinVar	Κλινική σχέση γενοτύπου/φαινοτύπου	http://www.ncbi.nlm.nih.gov/clinvar/

Η διαδικασία της κλήσης παραλλαγών δίνει ως αποτέλεσμα ένα μεγάλο αριθμό παραλλαγών που μπορεί να κυμαίνεται από μερικές εκατοντάδες έως εκατομμύρια παραλλαγές, ανάλογα την εφαρμογή της Αλληλούχησης Επόμενης Γενιάς. Ένας εύκολος τρόπος να ιεραρχηθούν αυτές οι παραλλαγές είναι η εκτίμησή τους βάσει της συχνότητας αλληλομόρφου τους στον γενικό πληθυσμό, λαμβάνοντας πάντα υπόψη την ασθένεια που μελετάται. Στην περίπτωση του κληρονομικού καρκίνου που αντιμετωπίζεται ως μονογονιδιακή νόσος οι παραλλαγές ενδιαφέροντος είναι πολύ οι σπάνιες, δηλαδή οι παραλλαγές με συχνότητα αλληλομόρφου μικρότερη του 0.5% και 2% για τα σύνδρομα με επικρατή και υπολειπόμενο τρόπο κληρονόμησης αντίστοιχα, με τη λογική ότι οι παραλλαγές που είναι συχνές στον πληθυσμό δεν μπορεί να ευθύνονται για έναν τόσο διακριτό φαινότυπο (Karki *et al.*, 2015). Τα όρια αυτά θα πρέπει να προσαρμόζονται ανάλογα την ασθένεια και τον πληθυσμό που εξετάζονται. Λόγου χάρη, οι ιδρυτικές παθογόνοι παραλλαγές στο γονίδιο *BRCA1* βρίσκονται σε συχνότητα που αγγίζει το 2% στον πληθυσμό των Εβραίων Εσκενάζι (Levy-Lahad *et al.*, 1997), μιας και πρόκειται για έναν σχετικά γενετικά ομογενή πληθυσμό, παρόλο που οι παθογόνοι παραλλαγές σε αυτό το γονίδιο είναι εξαιρετικά σπάνιες στον γενικό πληθυσμό.

Με την έλευση της τεχνολογίας Αλληλούχησης Επόμενης Γενιάς, πολλές μελέτες πληθυσμών κατάφεραν να ολοκληρωθούν και πλέον υπάρχουν σημαντικές πηγές δεδομένων που διαθέτουν

πληροφορίες σχετικά με τις **συχνότητες αλληλομόρφων** των παραλλαγών στους πληθυσμούς. Από τις πρώτες τέτοιου είδους πηγές ήταν η βάση δεδομένων του Προγράμματος Αλληλούχησης 1.000 γονιδιωμάτων (1000 Genomes Project) η οποία περιείχε δεδομένα από την αλληλούχηση ολόκληρου του γονιδιώματος 2.500 υγιών ατόμων (L. Clarke *et al.*, 2012), η Κοόρτη του Εθνικού Ινστιτούτου Καρδιάς, Πνευμόνων και Αίματος (National Heart, Leung and Blood Institute – NHLBI Cohort) που περιλαμβάνει δεδομένα από την αλληλούχηση των κωδικών περιοχών του γονιδιώματος 6.500 ασθενών με καρδιακά, πνευμονικά και/ή αιματολογικά νοσήματα (Tennessen *et al.*, 2012) και το πρόγραμμα HapMap που συνείσφερε με ένα σύνολο δεδομένων βασισμένων σε κοινούς πολυμορφισμούς για τον καθορισμό απλοτύπων σε 270 ανθρώπους διαφορετικής εθνικότητας (International HapMap *et al.*, 2007). Σήμερα, η πιο σημαντική βάση πληθυσμιακών δεδομένων είναι το gnomAD, στο οποίο περιέχονται πληροφορίες από την αλληλούχηση μόνο των κωδικών περιοχών του γονιδιώματος 125.748 (Karczewski *et al.*, 2020) και ολόκληρου του γονιδιώματος 15.708 ατόμων. Εκτός από τον τεράστιο όγκο δεδομένων του gnomAD, μία πολύ σημαντική πληροφορία που προσφέρεται είναι η κατηγοριοποίηση των ατόμων με βάση την εμφάνιση ή όχι κάποιας νόσου έως τη στιγμή που αναλύθηκε το DNA τους. Με αυτόν τον τρόπο, τα γονιδιώματα που φέρουν την ετικέτα "όχι-καρκίνος – non-cancer" μπορούν να χρησιμεύσουν ως δεδομένα ελέγχου για μια μελέτη κληρονομικού καρκίνου, μιας και αφορούν ανθρώπους που δεν έχουν νοσήσει με κάποιον τύπο καρκίνου. Ένα σημαντικό σύνολο δεδομένων που μπορεί να χρησιμεύσει ως ομάδα ελέγχου σε μελέτες του κληρονομικού καρκίνου είναι επίσης οι «Υπέροχες Κυρίες άνω των 70» (Fabulous Ladies Over Seventy – FLOSSIES· πρόσβαση: https://whi.color.com/). Όλες οι γενετικές παραλλαγές που περιέχονται σε αυτή την πηγή δεδομένων προέρχονται από την αλληλούχηση DNA γαμετικής σειράς περίπου 10.000 γυναικών, ηλικίας άνω των 70 ετών που δεν έχουν εμφανίσει ποτέ καρκίνο.

Σημαντικό ρόλο στην ιεράρχηση των παραλλαγών έχουν τα προγραμματιστικά εργαλεία που υλοποιούν αλγορίθμους για την **πρόβλεψη της επίπτωσης της παραλλαγής** στη δομή και τη λειτουργικότητα του γονιδίου και της παραγόμενης πρωτεΐνης. Λογισμικά όπως το VEP (McLaren *et al.*, 2016) και το snpEff (Cingolani *et al.*, 2012) μας δίνουν πληροφορίες για την επίπτωση της παραλλαγής στο γονίδιο, το μετάγραφο και την προκύπτουσα αλλαγή των αμινοξέων στην αλληλουχία της πρωτεΐνης. Η επίπτωση στη συνέχεια κατηγοριοποιείται βάσει σαφώς ορισμένων κανόνων και περιγράφεται βάσει της οντολογίας ακολουθίας (Sequence Ontology – SO) (Eilbeck *et al.*, 2005). Λόγου χάρη, οι μη νοηματικές παραλλαγές χαρακτηρίζονται ως παραλλαγές με εξαιρετικά υψηλή επίπτωση (Πίνακας 6).

Πίνακας 6: Επίπτωση των παραλλαγών βάσει της οντολογίας ακολουθίας SO. Τροποποίηση από (Ensembl).

Table 6: Sequence ontology terms for variant consequence. Modification from (Ensembl).

SO όρος	**Ελληνική μετάφραση SO όρου**	**Περιγραφή**	**Επίπτωση παραλλαγής**
transcript_ablation	κατάργηση μεταγράφου	Κατάργηση χαρακτηριστικών στην οποία η διαγραμμένη περιοχή περιλαμβάνει ένα χαρακτηριστικό μεταγραφής	Υψηλή
splice_acceptor_variant	παραλλαγή δέκτη ματίσματος	Παραλλαγή ματίσματος που αλλάζει την περιοχή των 2 βάσεων στο 3' άκρο ενός ιντρονίου	Υψηλή
splice_donor_variant	παραλλαγή δότη ματίσματος	Παραλλαγή ματίσματος που αλλάζει την περιοχή των 2 βάσεων στο 5' άκρο ενός ιντρονίου	Υψηλή
stop_gained	εισαγωγή πρόωρου κωδικόνιου τερματισμού	Παραλλαγή ακολουθίας η οποία αλλάζει τουλάχιστον μία βάση κωδικονίου, με αποτέλεσμα ένα πρόωρο κωδικόνιο τερματισμού, που οδηγεί σε μικρότερο μετάγραφο	Υψηλή
frameshift_variant	πλαισιοτροποποιητική παραλλαγή	Παραλλαγή ακολουθίας που τροποποιεί το μεταφραστικό πλαίσιο ανάγνωσης, επειδή ο αριθμός των νουκλεοτιδίων που εισάγονται ή διαγράφονται δεν είναι πολλαπλάσιο των τριών	Υψηλή
stop_lost	απώλεια κωδικονίου τερματισμού	Παραλλαγή ακολουθίας όπου αλλάζει τουλάχιστον μία βάση του κωδικονίου τερματισμού (stop), με αποτέλεσμα ένα μεγαλύτερο μετάγραφο	Υψηλή
start_lost	απώλεια κωδικονίου έναρξης	Παραλλαγή σε κωδικόνιο που αλλάζει τουλάχιστον μία βάση του κανονικού κωδικονίου έναρξης	Υψηλή
transcript_amplification	ενίσχυση μεταγράφου	Ενίσχυση χαρακτηριστικών μιας περιοχής που περιέχει ένα μετάγραφο	Υψηλή
inframe_insertion	εισαγωγή εντός πλαισίου ανάγνωσης	Μη συνώνυμη παραλλαγή εντός πλαισίου ανάγνωσης που εισάγει βάσεις στην κωδική ακολουθία	Μέτρια
inframe_deletion	απαλοιφή εντός πλαισίου ανάγνωσης	Μη συνώνυμη παραλλαγή εντός πλαισίου ανάγνωσης που διαγράφει βάσεις από την κωδική ακολουθία	Μέτρια
missense_variant	παρανοηματική παραλλαγή	Παραλλαγή ακολουθίας, που αλλάζει μία ή περισσότερες βάσεις, με αποτέλεσμα μια διαφορετική αλληλουχία αμινοξέων αλλά όπου διατηρείται το μήκος	Μέτρια
protein_altering_variant	παραλλαγή τροποποίησης πρωτεΐνης	Παραλλαγή ακολουθίας που προβλέπεται να αλλάξει την πρωτεΐνη που κωδικοποιείται	Μέτρια
splice_region_variant	παραλλαγή περιοχής ματίσματος	Παραλλαγή ακολουθίας στην οποία έχει συμβεί μια αλλαγή στην περιοχή της θέσης ματίσματος,	Χαμηλή

		είτε εντός 1-3 βάσεων του εξονίου είτε 3-8 βάσεων του ιντρονίου	
incomplete_terminal_ codon_variant	ελλιπής παραλλαγή κωδικονίου τερματισμού	Παραλλαγή ακολουθίας όπου αλλάζει τουλάχιστον μία βάση του τελικού κωδικονίου ενός ελλιπώς σχολιασμένου μεταγράφου	Χαμηλή
start_retained_variant	παραλλαγή διατήρησης έναρξης	Παραλλαγή ακολουθίας όπου αλλάζει τουλάχιστον μία βάση στο κωδικόνιο έναρξης, αλλά η έναρξη διατηρείται	Χαμηλή
stop_retained_variant	παραλλαγή διατήρησης τερματισμού	Παραλλαγή ακολουθίας όπου αλλάζει τουλάχιστον μία βάση στο κωδικόνιο τερματισμού, αλλά ο τερματισμός διατηρείται	Χαμηλή
synonymous_variant	συνώνυμη παραλλαγή	Παραλλαγή αλληλουχίας όπου δεν προκύπτει αλλαγή στο κωδικοποιημένο αμινοξύ	Χαμηλή
coding_sequence_variant	παραλλαγή κωδικής ακολουθίας	Παραλλαγή ακολουθίας που αλλάζει την κωδική ακολουθία	Τροποποίηση
mature_miRNA_variant	παραλλαγή ωρίμου miRNA	Παραλλαγή σε μετάγραφο που βρίσκεται με την αλληλουχία του ώριμου miRNA	Τροποποίηση
5_prime_UTR_variant	παραλλαγή 5 άκρου UTR	Παραλλαγή στην 5' αμετάφραστη περιοχή	Τροποποίηση
3_prime_UTR_variant	παραλλαγή 3 άκρου UTR	Παραλλαγή στην 3'αμετάφραστη περιοχή	Τροποποίηση
non_coding_transcript_ exon_variant	παραλλαγή εξονίου μη κωδικού μεταγράφου	Παραλλαγή ακολουθίας που αλλάζει την αλληλουχία ενός μη-κωδικού εξονίου σε ένα μη-κωδικό μετάγραφο	Τροποποίηση
intron_variant	παραλλαγή ιντρονίου	Παραλλαγή ενός μεταγράφου που βρίσκεται εντός του ιντρονίου	Τροποποίηση
NMD_transcript_variant	παραλλαγή μεταγράφου ΜΔΑ	Παραλλαγή ενός μεταγράφου που είναι στόχος μη νοηματικά διαμεσολαβούμενης αποικοδόμησης (ΜΔΑ)	Τροποποίηση
non_coding_transcript_ variant	παραλλαγή μη κωδικού μεταγράφου	Παραλλαγή ενός μεταγράφου ενός μη κωδικού RNA γονιδίου	Τροποποίηση
upstream_gene_variant	παραλλαγή περιοχής ανωδικά γονιδίου	Παραλλαγή ακολουθίας που βρίσκεται ανοδικά ενός γονιδίου	Τροποποίηση
downstream_gene_variant	παραλλαγή περιοχής καθοδικά γονιδίου	Παραλλαγή ακολουθίας που βρίσκεται καθοδικά ενός γονιδίου	Τροποποίηση
TFBS_ablation	κατάργηση περιοχής πρόσδεσης μεταγραφικού παράγοντα	Κατάργηση χαρακτηριστικών όπου η διαγραμμένη περιοχή περιλαμβάνει μια τοποθεσία πρόσδεσης μεταγραφικού παράγοντα	Τροποποίηση
TFBS_amplification	ενίσχυση περιοχής πρόσδεσης μεταγραφικού παράγοντα	Ενίσχυση χαρακτηριστικών όπου η ενισχυμένη περιοχή περιλαμβάνει μια τοποθεσία πρόσδεσης μεταγραφικού παράγοντα	Τροποποίηση
TF_binding_site_variant	παραλλαγή περιοχής πρόσδεσης μεταγραφικού παράγοντα	Παραλλαγή ακολουθίας που βρίσκεται εντός θέσης πρόσδεσης μεταγραφικού παράγοντα	Τροποποίηση
regulatory_region_ ablation	κατάργηση ρυθμιστικής περιοχής	Κατάργηση χαρακτηριστικών όπου η διαγραμμένη περιοχή περιλαμβάνει μια ρυθμιστική περιοχή	Μέτρια

regulatory_region_ amplification	ενίσχυση ρυθμιστικής περιοχής	Ενίσχυση ενός χαρακτηριστικού σε μία περιοχή που περιλαμβάνει μια ρυθμιστική περιοχή	Τροποποίηση
feature_elongation	επέκταση χαρακτηριστικού	Παραλλαγή ακολουθίας που προκαλεί επέκταση ενός γονιδιωματικού χαρακτηριστικού σε σχέση με την ακολουθία αναφοράς	Τροποποίηση
regulatory_region_variant	παραλλαγή ρυθμιστικής περιοχής	Παραλλαγή ακολουθίας που βρίσκεται εντός ρυθμιστικής περιοχής	Τροποποίηση
feature_truncation	κολόβωση χαρακτηριστικού	Παραλλαγή ακολουθίας που προκαλεί τη κολόβωση ενός γονιδιωματικού χαρακτηριστικού σε σχέση με την ακολουθία αναφοράς	Τροποποίηση
intergenic_variant	διαγονιδιακή παραλλαγή	Παραλλαγή ακολουθίας που βρίσκεται στην περιοχή μεταξύ δύο γονιδίων	Τροποποίηση

Οι σχολιασμοί με βάση την πρόβλεψη πραγματοποιούνται με τη βοήθεια ***in silico* εργαλείων** που χρησιμοποιούν δεδομένα όπως ο βαθμός της εξελικτικής συντήρησης της ακολουθίας σε διάφορους οργανισμούς, πίνακες βαθμολογίας αντικατάστασης αμινοξέων και η πρόβλεψη της επίπτωσης που η παραλλαγή έχει στην τρισδιάστατη δομή της πρωτεΐνης (Gunning *et al.*, 2020). Συνήθως, οι πληροφορίες αυτές αποτελούν τα δεδομένα εισόδου σε ένα σύστημα μηχανικής μάθησης. Τα εργαλεία αυτά χρησιμοποιούν αλγορίθμους όπως τα Κρυφά Μαρκοβιανά μοντέλα ή τα νευρωνικά δίκτυα για την ταξινόμηση των παραλλαγών σε αυτές που προκαλούν βλάβη και σε αυτές που δεν έχουν κάποια επίπτωση στην παραγόμενη πρωτεΐνη, προσδίδοντάς τους συνήθως μια βαθμολογία (Oliver *et al.*, 2015). Παρόλο που τα εργαλεία αυτά συνήθως είναι ακριβή στις προβλέψεις τους, η ταξινόμηση που πραγματοποιούν θα πρέπει να χρησιμοποιείται ως μία ένδειξη σχετικά με την παθογένεια των παραλλαγών, ιδιαίτερα ελλείψει περαιτέρω στοιχείων όπως οι λειτουργικές μελέτες.

Ιδιαίτερα σημαντικό ρόλο έχουν και οι βάσεις δεδομένων που καταγράφουν την **κλινική σημασία των παραλλαγών**, όπως αυτή έχει αξιολογηθεί από διάφορες ομάδες κλινικών κι εργαστηριακών γενετιστών κι ερευνητών ανά τον κόσμο. Η βάση ClinVar του National Center for Biotechnology Information (NCBI) φιλοξενεί δεδομένα για 913.565 μοναδικές παραλλαγές (μέχρι τη στιγμή που γραφόταν αυτή η διατριβή) τις οποίες έχουν επιμεληθεί και κατηγοριοποιήσει χειρωνακτικά και στη συνέχεια υποβάλει, περίπου 2.000 ερευνητικές ομάδες ανά τον κόσμο (Landrum *et al.*, 2020). Η βάση δεδομένων LOVD (Leiden Open Variation Database) είναι μία ξεχωριστή περίπτωση, καθώς παρέχει και το αντίστοιχο λογισμικό ανοιχτού κώδικα το οποίο επιτρέπει στους χρήστες να δημιουργήσουν τη δική τους βάση γενετικών δεδομένων. Η κεντρική εγκατάσταση της βάσης LOVD παρέχει πληροφορίες για την κατηγοριοποίηση 328.322 μοναδικών παραλλαγών (μέχρι τη στιγμή που γραφόταν αυτή η διατριβή) (Fokkema *et al.*, 2011).

Παρόλο που η πλήρως αυτοματοποιημένη κατηγοριοποίηση των παραλλαγών θα ήταν χρονικά αποδοτική, ακόμα δεν είναι εφικτή. Υπάρχουν κάποια κριτήρια που πρέπει να ληφθούν υπόψη και η αξιολόγησή τους δε δύναται να αυτοματοποιηθεί πλήρως, είτε επειδή ακόμα δεν έχει βρεθεί βέλτιστος αλγόριθμος για την αυτοματοποίησή τους, είτε επειδή είναι ξεχωριστά για κάθε περίπτωση που μελετάται. Στην πρώτη κατηγορία ανήκουν τα κριτήρια τα οποία για να αξιολογηθούν χρειάζεται μελέτη

της βιβλιογραφίας, όπως η **δοκιμή των παραλλαγών σε λειτουργικές μελέτες**. Λόγου χάρη, η κατηγοριοποίηση μιας παραλλαγής ως επιβλαβούς μέσω μιας λειτουργικής μελέτης αποτελεί ισχυρή ένδειξη για την παθογένεια της παραλλαγής. Στη δεύτερη κατηγορία δύναται να ανήκουν σχεδόν όλα τα κριτήρια. Για παράδειγμα, ενώ είναι συνήθης πρακτική να μελετώνται μόνο οι παραλλαγές που εδράζονται σε κωδικές περιοχές, καθώς εκτιμάται πως το 85% των παθογόνων παραλλαγών βρίσκεται σε εξόνια γονιδίων, θα πρέπει να λαμβάνονται υπόψη κάποιες γνωστές παραλλαγές που εδράζονται σε ιντρονικές περιοχές, όπως η παθογόνος παραλλαγή στο γονίδιο *MUTYH* NM_001128425.1:c.504+19_504+31del (rs781222233) που είναι συχνή στον ελληνικό πληθυσμό (Fostira *et al.*, 2010).

Μετά από την ιεράρχηση των παραλλαγών, κι έχοντας καταλήξει στις ενδεχομένως κλινικά σημαντικές, θα πρέπει να αξιολογηθούν μία προς μία σε βάθος ως προς την παθογένειά τους. Κατά την αξιολόγηση των παραλλαγών λαμβάνονται υπόψη τόσο τα κριτήρια που έχουν ήδη αναφερθεί, όσο και επιπλέον κριτήρια, όπως ο τρόπος κληρονόμησης της γενετικής νόσου και ο διαχωρισμός (segregation) της παραλλαγής στην οικογένεια που μελετάται. Το Αμερικανικό Κολλέγιο της Ιατρικής Γενετικής (American College of Medical Genetics – ACMG) και η Ένωση για τη Μοριακή Παθολογία (Association for Molecular Pathology – AMP) έχουν θεσπίσει σαφείς κανόνες για την κατηγοριοποίηση παραλλαγών DNA τόσο γαμετικής όσο και σωματικής σειράς (M. M. Li *et al.*, 2017; Richards *et al.*, 2015).

Ανακεφαλαιώνοντας, τα κριτήρια που πρέπει να ληφθούν υπόψη κατά την ιεράρχηση, αλλά και την αξιολόγηση των παραλλαγών θα μπορούσαν να συνοψιστούν ως εξής:

- Συχνότητα της παραλλαγής σε διάφορους πληθυσμούς· όσο πιο κοινή η παραλλαγή τόσο λιγότερο πιθανό να προδιαθέτει σε κάποια Μενδελική ασθένεια.
- Αποτέλεσμα της παραλλαγής στην ακολουθία του γονιδίου· ανάλογα το αποτέλεσμα η παραλλαγή έχει διαφορετική επίπτωση. Για παράδειγμα, οι μη νοηματικές και πλαισιοτροποποιητικές παραλλαγές θεωρούνται παραλλαγές με εξαιρετικά υψηλή επίπτωση.
- Δεδομένα από *in silico* εργαλεία· η συμφωνία πολλαπλών υπολογιστικών εργαλείων ως προς τον χαρακτηρισμό μιας παραλλαγής ως επιβλαβούς ή μη, αποτελεί ισχυρή ένδειξη της παθογένειας ή ουδετερότητας της παραλλαγής αντίστοιχα.
- Δεδομένα από λειτουργικές μελέτες· η δοκιμή μιας παραλλαγής σε κάποια λειτουργική μελέτη όπου έχει αποδειχτεί ότι εκτρέπεται το φυσιολογικό μάτισμα ή επέρχεται διακοπή της φυσιολογικής λειτουργίας της πρωτεΐνης ως αποτέλεσμα της παραλλαγής, είναι ισχυρή ένδειξη της παθογένειάς της.
- Δεδομένα ως προς την κατηγοριοποίηση από δημόσιες βάσεις γενετικών δεδομένων.
- Δεδομένα από ανάλυση διαχωρισμού· ο εντοπισμός μιας παραλλαγής αποκλειστικά στο DNA των ατόμων που έχουν νοσήσει από το ίδιο γενετικό σύνδρομο σε μια οικογένεια είναι ένδειξη για την παθογένειά της.
- Γονίδιο στο οποίο εντοπίζεται η παραλλαγή (σε περίπτωση που μελετάται κάποια νόσος με γνωστό αίτιο).

- Τρόπος κληρονόμησης· λόγου χάρη, όταν η νόσος κληρονομείται με υπολειπόμενο τρόπο κληρονόμησης, μια παθογόνος παραλλαγή σε ετεροζυγωτία μάλλον δεν ευθύνεται για τον φαινότυπο, δεδομένου ότι έχει αποκλειστεί η ύπαρξη συνδυαστικής ετεροζυγωτίας (compound heterozygosity).

Το τελικό βήμα της τριτοταγούς ανάλυσης, πριν την αναφορά του αποτελέσματος του γενετικού ελέγχου, είναι η επιβεβαίωση της παραλλαγής με ανεξάρτητη μέθοδο, λόγου χάρη αλληλούχηση κατά Sanger, με σκοπό την αποφυγή ενός ψευδώς θετικού αποτελέσματος (Mu *et al.*, 2016). Το βήμα αυτό είναι ιδιαίτερα σημαντικό, καθώς το αποτέλεσμα του γενετικού ελέγχου επηρεάζει την μετέπειτα διαχείριση του ασθενούς. Για παράδειγμα, ένας γενετικός έλεγχος όπου έχει εντοπιστεί μία ψευδώς θετική παραλλαγή στο γονίδιο *BRCA1*, μπορεί να έχει ως αποτέλεσμα η ασθενής να προβεί σε προφυλακτικό χειρουργείο χωρίς να το χρειάζεται στην πραγματικότητα.

1.6.2.4. Επιλογή μεταγράφων για τον σχολιασμό των παραλλαγών

Κάθε γονίδιο μπορεί να έχει πολλά μετάγραφα τα οποία κωδικοποιούν διαφορετικές παραλλαγές μίας πρωτεΐνης, ανάλογα το κύτταρο στο οποίο εκφράζονται και την λειτουργία που επιτελούν. Κάθε **μετάγραφο ενός γονιδίου** αποτελείται από διαφορετικό συνδυασμό εξονίων του γονιδίου και μεταφράζεται σε μία μοναδική **ισομορφή της πρωτεΐνης**. Καθώς κάθε μετάγραφο αποτελείται από ένα μοναδικό συνδυασμό εξονίων, η αλληλουχία του είναι επίσης μοναδική (Εικόνα 27). Αυτό σημαίνει πως μια παραλλαγή σε ένα γονίδιο είναι πιθανό να έχει διαφορετική επίπτωση ανάλογα το μετάγραφο που εξετάζεται. Λόγου χάρη, η παραλλαγή NC_000017.10:g.41199682C>T που εδράζεται στο γονίδιο *BRCA1* εισάγει ένα πρώιμο κωδικόνιο τερματισμού στο μετάγραφο NM_007294.4 (NM_007294.4: c.5445G>A, NP_009225.1: p.Trp1815Ter) ενώ οδηγεί απλώς σε αλλαγή ενός ασπαραγινικού οξέος σε ασπαραγίνη στο μετάγραφο NM_007299.4 (NM_007299.4: c.2059G>A, NP_009230.2: p.Asp687Asn). Οι παραλλαγές που εισάγουν πρώιμο κωδικόνιο τερματισμού είναι πολύ πιο πιθανό να είναι παθογόνοι σε σχέση με τις παρανοηματικές παραλλαγές, επομένως ο σωστός σχολιασμός είναι εξαιρετικά σημαντικός σε αυτή την περίπτωση.

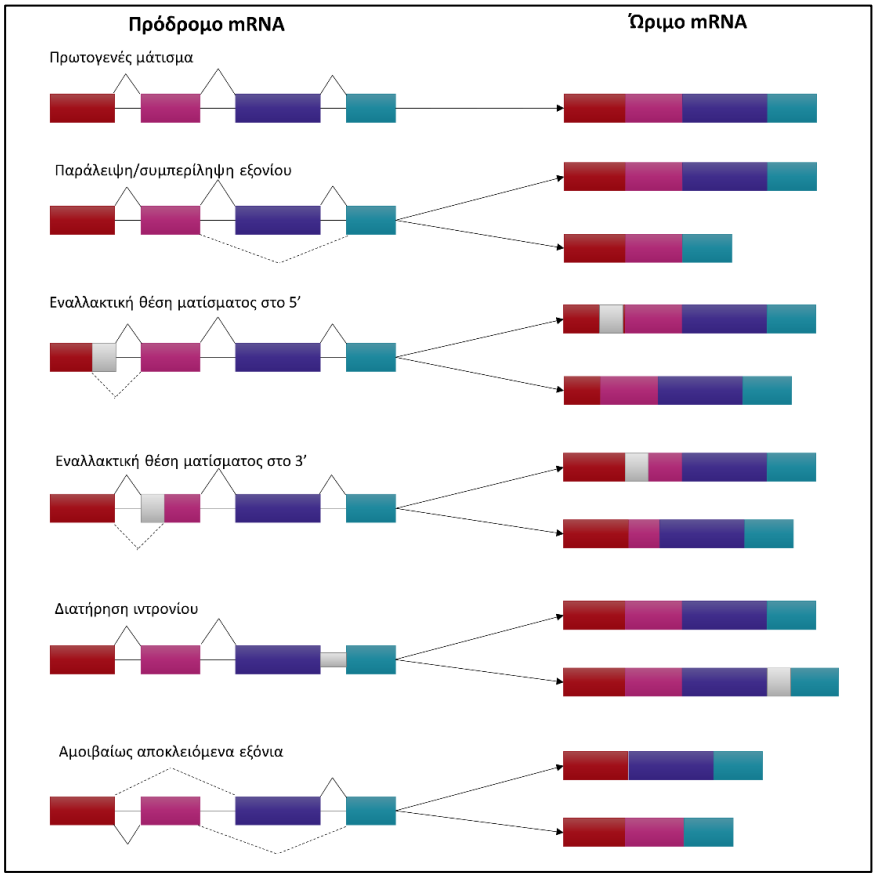
Στο παρελθόν έχουν διεξαχθεί μελέτες που έχουν αναδείξει τη σημασία της προσεκτικής επιλογής μεταγράφων. Πιο συγκεκριμένα, οι McCarthy et al. (McCarthy *et al.*, 2014), συνέκριναν τα εργαλεία ανοιχτού κώδικα VEP και ANNOVAR (K. Wang *et al.*, 2010), καθώς και τα σύνολα μεταγράφων RefSeq (Pruitt *et al.*, 2012) και Ensembl (Flicek *et al.*, 2012). Στη μελέτη τους, πραγματοποιώντας μία άμεση σύγκριση των εργαλείων χρησιμοποιώντας τα ίδια μετάγραφα, το ακριβές ποσοστό συμφωνίας ήταν 86,5%, ενώ στην περίπτωση της σύγκρισης των συνόλων μεταγράφων RefSeq και Ensembl, το ακριβές ποσοστό συμφωνίας ήταν 85%. Μια μεταγενέστερη μελέτη, από τους Frankish et al. (Frankish *et al.*, 2015), όπου συγκρίθηκαν τα σύνολα μεταγράφων GENCODE (Harrow *et al.*, 2012) και RefSeq, έδειξε ότι περίπου το 30% των παραλλαγών απώλειας λειτουργίας είχε εμπλουτιστεί με πληροφορίες με διαφορετικό τρόπο.

Δεν υπάρχει κάποιος σαφώς ορισμένος κανόνας για την επιλογή ενός μεταγράφου, αλλά υπάρχουν κάποιες ενδείξεις που διευκολύνουν τον προσδιορισμό του **κανονικού ή πρωτεύοντος μεταγράφου**

(canonical or primary transcript) για κάθε γονίδιο. Έτσι, συνήθως το κανονικό μετάγραφο είναι αυτό που είναι γνωστό από επιστημονικά δεδομένα ότι εκφράζεται συχνότερα και δίνει το μεγαλύτερο πλαίσιο ανάγνωσης. Γενικά, για τον ορισμό του κανονικού μεταγράφου συνήθως ακολουθούνται τα εξής βήματα ιεραρχικά:

1. Επιλέγεται το μετάγραφο που είναι το πιο διαδεδομένο, δηλαδή το μετάγραφο που εκφράζεται συχνότερα.
2. Ελέγχεται η ακολουθία των μεταγράφων κι επιλέγεται αυτό του οποίου η ακολουθία είναι πιο όμοια με τις ορθόλογες ακολουθίες άλλων ειδών.
3. Επιλέγεται το μετάγραφο που λόγω του μήκους ή της σύνθεσης αμινοξέων του, επιτρέπει τη σαφέστερη περιγραφή των τομέων και των ισομορφών της πρωτεΐνης.
4. Ελλείψει πληροφοριών, επιλέγεται το μετάγραφο με τη μεγαλύτερη ακολουθία.

Η βάση δεδομένων APPRIS φιλοξενεί μία πολύ σημαντική συλλογή δεδομένων που προέρχονται από το χειρωνακτικό και υπολογιστικό σχολιασμό των μεταγράφων όλων των γονιδίων δέκα οργανισμών, μεταξύ των οποίων και ο άνθρωπος, και παρέχει πληροφορίες για το εναλλακτικό μάτισμα, ενώ εντοπίζει τα κανονικά μετάγραφα γι' αυτούς τους οργανισμούς (Rodriguez *et al.*, 2018).



Εικόνα 27: Κοινοί τρόποι φυσιολογικού εναλλακτικού ματίσματος.

Figure 27: Common ways of normal alternative splicing.

1.6.3. Προκλήσεις στη βιοπληροφορική ανάλυση γενετικών δεδομένων με κλινική εφαρμογή

Επί του παρόντος, η συνήθης πρακτική στη μοριακή διάγνωση είναι η αλληλούχηση που στοχεύει σε συγκεκριμένες ομάδες γονιδίων για τον εντοπισμό παθογόνων παραλλαγών σε γονίδια που είναι γνωστό ή για τα οποία υπάρχει υποψία ότι σχετίζονται με Μενδελικές ασθένειες. Η αλληλούχηση μόνο των γονιδίων ενδιαφέροντος είναι οικονομικά αποδοτική σε σύγκριση με την αλληλούχηση ολόκληρου του γονιδιώματος -ή ακόμη και μόνο των κωδικών περιοχών του γονιδιώματος- ενώ το επιτευχθέν βάθος ανάγνωσης είναι μεγαλύτερο, παρέχοντας μεγαλύτερη ακρίβεια στην ανίχνευση παραλλαγών.

Ένα κρίσιμο μέρος της διαδικασίας είναι η βιοπληροφορική ανάλυση και η ερμηνεία των παραγόμενων δεδομένων, διαδικασίες οι οποίες μπορεί να κρύβουν προκλήσεις. Η βιοπληροφορική είναι ένα σχετικώς νέο πεδίο, με τον αριθμό των έμπειρων επιστημόνων βιοπληροφορικής -συγκεκριμένα για το πεδίο της μοριακής διάγνωσης- να είναι περιορισμένος και ως εκ τούτου, οι γενετιστές, που δεν διαθέτουν την απαραίτητη τεχνογνωσία καλούνται να επιλέξουν και να χειριστούν τα απαραίτητα υπολογιστικά εργαλεία και το λογισμικό για την ανάλυση δεδομένων. Μια συνήθης επιλογή είναι το λογισμικό που συνοδεύει τα αντιδραστήρια που χρησιμοποιούνται για την προετοιμασία των βιβλιοθηκών, το οποίο συχνά είναι ένα φιλικό προς το χρήστη και ολοκληρωμένο λογισμικό. Παρόλο που με ένα τέτοιο λογισμικό η ανάλυση μπορεί να πραγματοποιηθεί εύκολα, η διαδικασία απαιτεί και πάλι πολλή προσοχή και ιδιαίτερες δεξιότητες από το άτομο που την εκτελεί (Gomez-Lopez *et al.*, 2019).

1.6.4. Προκλήσεις στην ανάλυση δεδομένων προερχομένων από μαζική παράλληλη αλληλούχηση DNA όγκου

Η μαζική παράλληλη αλληλούχηση πολλαπλών γονιδίων και περιοχών του γονιδιώματος μέσω της μεθόδου αλληλούχησης επόμενη γενιάς χρησιμοποιείται ευρέως ως μέθοδος για τον προσδιορισμό του γονιδιώματος του όγκου. Ως εκ τούτου, μέσω της ταυτοποίησης παραλλαγών που μπορούν να στοχευτούν, προκύπτουν νέες, μοριακές θεραπείες που δύνανται και στοχεύουν στο να βελτιώσουν την έκβαση της νόσου των ασθενών με καρκίνο. Στην πραγματικότητα, για έναν αριθμό ασθενών, η αναγνώριση μιας στοχευόμενης παραλλαγής μπορεί να είναι η τελευταία προσπάθεια για θεραπεία, αφού εξαντληθούν όλες οι άλλες επιλογές.

Στην πράξη, η αλληλούχηση του DNA όγκου μπορεί να είναι εξαιρετικά δύσκολη, λόγω της ετερογένειας του όγκου και της κακής ποιότητας του DNA, η οποία επηρεάζεται σημαντικά από το αρχικό υλικό. Για παράδειγμα, το DNA που προέρχεται από ιστό όγκου που διατηρείται σε παραφίνη μετά από σταθεροποίηση με φορμαλίνη (Formalin-Fixed, Paraffin-Embedded - FFPE) είναι συνήθως πολύ κακής ποιότητας. Αν και αυτή η διαδικασία είναι πολύ αποτελεσματική για την ασφαλή αποθήκευση ιστών για μεγάλα χρονικά διαστήματα (Gaffney *et al.*, 2018), έχει ως αποτέλεσμα τον κατακερματισμό του DNA ή/και τη δημιουργία σταυροσυνδέσεων μεταξύ των μορίων του DNA, η οποία επηρεάζει σημαντικά το αποτέλεσμα της αλληλούχησης «μολύνοντας» της ακολουθίες εξόδου με ψευδώς θετικές παραλλαγές (McDonough *et al.*, 2019; Arreaza *et al.*, 2016).

Ταυτόχρονα, η δυναμική ώθηση του καρκίνου να εξελιχθεί δημιουργεί πολλαπλούς υποπληθυσμούς (κλώνους) καρκινικών κυττάρων, οι οποίοι φέρουν διαφορετικές σωματικές παραλλαγές ή ακόμη και εντελώς διαφορετικά γονιδιώματα (Turajlic *et al.*, 2019; Caldas, 2012; Navin *et al.*, 2011). Επομένως, μπορεί να είναι αρκετά δύσκολο να προσδιοριστούν οι πραγματικές παραλλαγές μεταξύ ενός υποσυνόλου χιλιάδων παραλλαγών, ειδικά όταν ανιχνεύονται σε πολύ χαμηλές συχνότητες. Η ετερογένεια είναι ένα επιπλέον εμπόδιο κατά την αλληλούχηση των γονιδιωμάτων όγκου, καθώς πολλές παραλλαγές μπορούν να εμφανιστούν σε πολύ χαμηλή συχνότητα, με αποτέλεσμα είτε να θεωρηθούν ψευδώς θετικές είτε να παραλειφθούν βάσει του ορίου ανίχνευσης που προσδιορίζεται κατά τον σχεδιασμό της ανάλυσης.

Λόγω αυτών των φαινομένων, η αναγνώριση των πραγματικών σωματικών παραλλαγών, δηλαδή των παραλλαγών που πράγματι υπάρχουν στον όγκο και δεν παράγονται εξαιτίας κάποιου σφάλματος κατά την αλληλούχηση είναι πολύ δύσκολη. Οι ροές διοχέτευσης εντολών διεργασιών για την ταυτοποίηση σωματικών παραλλαγών πρέπει να επιδεικνύουν υψηλή ακρίβεια, ενώ τα παραγόμενα δεδομένα, τα οποία συνήθως είναι πολύ μεγάλα σε όγκο και αφορούν πολλούς ασθενείς, πρέπει τελικά να αξιολογηθούν χειροκίνητα από υψηλά καταρτισμένους επαγγελματίες. Αντίθετα, οι αλγόριθμοι συνήθως δεν συμφωνούν μεταξύ τους σχετικά με την ανίχνευση αληθινών σωματικών παραλλαγών, με το ποσοστό συμφωνίας ανάμεσα σε διαφορετικούς αλγόριθμους ανίχνευσης σωματικών παραλλαγών να φτάνει περίπου στο 0,5% -3% (Q. Wang *et al.*, 2019; Kroigard *et al.*, 2016). Αυτό οδηγεί σε μεγάλο αριθμό κληθέντων παραλλαγών, ενώ καθιστά απαραίτητη την επιβεβαίωση των πραγματικών παραλλαγών με τη χρήση ανεξάρτητης μεθόδου, όπως η αλληλούχηση με τη μέθοδο Sanger (Mu *et al.*, 2016). Η αξιολόγηση ενός τόσο μεγάλου πλήθους ανιχνευμένων παραλλαγών με μια ανεξάρτητη μέθοδο είναι μία χρονοβόρα και πολύπλοκη διαδικασία.

1.6.5. Βάσεις γενετικών δεδομένων

Οι ραγδαίες εξελίξεις στη βιοτεχνολογία τα τελευταία χρόνια είχαν ως αποτέλεσμα την εκθετική αύξηση του όγκου των παραγόμενων γενετικών δεδομένων. Το γεγονός αυτό καθιστά δύσκολη την πρόσβαση από την ερευνητική κοινότητα σε αυτά τα τόσο σημαντικά δεδομένα, ενώ δημιουργεί ερωτήματα σχετικά με τη χρησιμότητά τους. Οι εθνικές βάσεις γενετικών δεδομένων μπορούν να δώσουν λύση στα παραπάνω προβλήματα, λειτουργώντας ως εργαλεία για τη διαχείριση δεδομένων και την πρόσβαση σε αυτά, όχι μόνο παρέχοντας ανεκτίμητα δεδομένα, αλλά και αυξάνοντας τις πληροφορίες που γίνονται διαθέσιμες, οι οποίες διαφορετικά χάνονται, διευκολύνοντας ταυτόχρονα την ενσωμάτωση των δεδομένων σε κεντρικές πηγές γενετικών δεδομένων.

Ήδη από το 2010, οπότε εκδόθηκαν τα πρώτα αποτελέσματα του 1000 Genomes Project (Genomes Project *et al.*, 2015), έγινε φανερό ότι πολλές σπάνιες παραλλαγές παρατηρούνται αποκλειστικά σε συγκεκριμένους πληθυσμούς. Ωστόσο, ορισμένοι πληθυσμοί υποεκπροσωπούνται σε μελέτες μεγάλης κλίμακας, όπως το 1000 Genomes Project, το Exome Sequencing Project (Fu *et al.*, 2013), το Exome Aggregation Consortium (Lek *et al.*, 2016) και το Genome Aggregation Consortium (Koch, 2020). Το γεγονός αυτό είναι αποτέλεσμα ενός συνδυασμού μεταξύ μιας καταγεγραμμένης προτίμησης της ερευνητικής κοινότητας προς τη μελέτη πληθυσμών που είναι ήδη καλά χαρακτηρισμένοι και της

ανεπαρκούς συμπερίληψης σε γενετικές μελέτες άλλων πληθυσμών, οι οποίοι συνήθως αφορούν μικρότερο δείγμα (Bentley *et al.*, 2017).

Οι **εθνικές πηγές γενετικών δεδομένων**, χάρη στην εκτεταμένη καταγραφή της γενετικής ποικιλομορφίας μεταξύ πληθυσμών αλλά και εντός του ίδιου πληθυσμού, δίνουν τη δυνατότητα σε αυτή την ετερογένεια να αναδυθεί. Με αυτόν τον τρόπο, παρέχουν πιο συγκεκριμένο υπολογισμό των συχνοτήτων αλληλομόρφων στους διάφορους πληθυσμούς, επιτρέποντας ταυτόχρονα την ταυτοποίηση σπάνιων παραλλαγών που σχετίζονται με τον κίνδυνο ασθένειας και είναι πιο συχνές σε συγκεκριμένους πληθυσμούς. Επιπλέον, η καταγραφή της γεωγραφικής προέλευσης των ατόμων που φέρουν τις παραλλαγές διευκολύνει τον προσδιορισμό της εντοπιότητας του κάθε αλληλομόρφου, η οποία με τη σειρά της βοηθά στην εξήγηση του γενετικού υποβάθρου και της ποικιλομορφίας σε διαφορετικούς (υπο)πληθυσμούς καθώς και των αλλαγών στις συχνότητες αλληλομόρφων στην πάροδο του χρόνου. Για το λόγο αυτό, τέτοιες προσπάθειες μπορούν να χρησιμεύσουν ως θεμέλιο για την ιατρική ακριβείας σε κάθε χώρα, επιτρέποντας τη θέσπιση πρωτοκόλλων γενετικού ελέγχου για τον εκάστοτε πληθυσμό και έχοντας ως αποτέλεσμα την αποδοτικότητα κόστους της χειρουργικής και φαρμακευτικής θεραπείας.

Εκτός από την υποεκπροσώπηση των πληθυσμών στις υπάρχουσες βάσεις δεδομένων, μία άλλη πρόκληση στη γενετική σήμερα είναι η έλλειψη γενετικών δεδομένων που είναι διαθέσιμα στο κοινό σε συνδυασμό με **φαινοτυπικές και κλινικές πληροφορίες**. Αυτού του είδους τα δεδομένα έχουν μεγάλη σημασία για ένα πλήθος εφαρμογών, όπως ο προσδιορισμός της φαινοτυπικής ετερογένειας, η ανάλυση συσχέτισης γονοτύπου-φαινοτύπου, η παθογένεια των παραλλαγών και η ιατρική ακριβείας. Στις περισσότερες από τις δημόσιες βάσεις δεδομένων γενετικών παραλλαγών, τα γενετικά δεδομένα καταγράφονται χωρίς τη συνοδεία φαινοτυπικών ή/και κλινικών δεδομένων. Για παράδειγμα, η βάση OMIM (Hamosh *et al.*, 2002) περιέχει γενικές πληροφορίες σχετικά με συσχετίσεις μεταξύ γονιδίων και αντίστοιχων φαινοτύπων, χωρίς καταγραφή γενετικών παραλλαγών. Οι βάσεις dbGAP (Tryka *et al.*, 2014) και NHGRI-EBI GWAS Catalogue (Buniello *et al.*, 2019) καταγράφουν τις συσχετίσεις γονοτύπων και φαινοτύπων, όπως έχουν προσδιοριστεί από Μελέτες Συσχέτισης Ολόκληρου του Γονιδιώματος. Ως εκ τούτου, στις βάσεις αυτές καταγράφονται πιο κοινές παραλλαγές στους πληθυσμούς. Τέλος, όσον αφορά την κεντρική εγκατάσταση της βάσης LOVD (Fokkema *et al.*, 2011) και τη βάση δεδομένων ClinVar (Landrum *et al.*, 2020), ενώ ενθαρρύνουν τους συνεισφέροντες χρήστες να υποβάλλουν μαζί με τα γενετικά δεδομένα και τους φαινοτύπους που συσχετίζονται με αυτά, δεν καταγράφουν λεπτομερή κλινικά δεδομένα. Ωστόσο, η σύνδεση γονοτύπων και κλινικών δεδομένων μπορεί να παρέχει σημαντικές πληροφορίες για τα γονίδια και τις γενετικές παραλλαγές που συμβάλλουν στην ανάπτυξη της ανθρώπινης νόσου (Deans *et al.*, 2015).

Τέλος, ένα σημαντικό πρόβλημα που αντιμετωπίζει η ερευνητική κοινότητα είναι ότι εξαιτίας του κατακερματισμού των δραστηριοτήτων προσδιορισμού αλληλουχίας, δημιουργούνται πολλαπλές πηγές δεδομένων μικρής κλίμακας. Το αποτέλεσμα είναι ο διασκορπισμός των γενετικών δεδομένων σε πολλές διαφορετικές μεριές, γεγονός που αναστέλλει την ανοιχτή έρευνα (den Dunnen, 2018). Για το λόγο αυτό, οι τυποποιημένες γενετικές βάσεις δεδομένων έχουν καταστεί ισχυρά εργαλεία που ενισχύουν την πρόοδο της κοινότητας.

1.7. Σκοπός

Η παρούσα διατριβή έχει ως σκοπό την ανάπτυξη και αξιολόγηση εφαρμογών βιοπληροφορικής που έχουν κλινική χρησιμότητα στην ιατρική ακριβείας και πιο συγκεκριμένα στη γενετική του καρκίνου. Η βασική κατεύθυνση είναι η δημιουργία μίας βάσης δεδομένων που καταγράφει τη γενετική ποικιλομορφία και τα αντίστοιχα κλινικά δεδομένα Ελλήνων ασθενών με καρκίνο σε εθνικό επίπεδο. Έτσι, στα επόμενα κεφάλαια θα παρουσιαστεί η βάση δεδομένων **CanVaS** (Cancer Variation Resource – πηγή δεδομένων για τις παραλλαγές στον καρκίνο), μια βάση δεδομένων που καταγράφει τη γενετική ποικιλομορφία και τα αντίστοιχα κλινικά δεδομένα Ελλήνων ασθενών με καρκίνο σε εθνικό επίπεδο. Στη συνέχεια, θα παρουσιαστεί το εργαλείο **VarTrace**, που αποτελεί την υλοποίηση μιας ροής διοχέτευσης εντολών διεργασιών με σκοπό την ακριβή ανίχνευση παραλλαγών DNA σωματικής σειράς, το οποίο προέρχεται από FFPE όγκους. Τέλος, θα παρουσιαστεί μία **ενδελλεχής σύγκριση ανάμεσα σε δύο λογισμικά εμπλουτισμού παραλλαγών**. Η σύγκριση αυτή έχει ως σκοπό την ανάδειξη των παγίδων που κρύβονται στη διαδικασία της ανάλυσης παραλλαγών προερχόμενων από μαζική παράλληλη αλληλούχηση DNA.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1. Ανάπτυξη βάσης γενετικών δεδομένων CanVaS

Η βάση δεδομένων CanVaS αποτελεί μια εθνική βάση δεδομένων που καταγράφει γενετικές παραλλαγές γαμετικής σειράς, συνοδευόμενες από τα σχετικά κλινικά δεδομένα Ελλήνων ασθενών με καρκίνο και μπορεί να προσπελαστεί στο: http://ithaka.rrp.demokritos.gr/CanVaS.

2.1.1. Ομάδα μελέτης

Τα δεδομένα που περιλαμβάνονται στη βάση δεδομένων CanVaS παρήχθησαν από την αλληλούχηση του DNA γαμετικής σειράς 7.363 ασθενών με καρκίνο και υγιών συγγενών τους, που είχαν παραπεμφθεί για γονιδιακή εξέταση λόγω προσωπικού ή οικογενειακού ιστορικού κακοήθειας στο Εργαστήριο Μοριακής Διαγνωστικής (ΕΜΔ) στο Εθνικό Κέντρο Έρευνας Φυσικών Επιστημών (ΕΚΕΦΕ) «Δημόκριτος» μεταξύ των ετών 1999-2019. Πριν από κάθε γονιδιακό έλεγχο, υπογράφηκε από όλους τους εξεταζόμενους γραπτή συγκατάθεση μετά από σχετική ενημέρωση, επιτρέποντας την ανώνυμη χρήση των δεδομένων τους για ερευνητικούς σκοπούς. Η μελέτη εγκρίθηκε από την επιτροπή βιοηθικής του ΕΚΕΦΕ «Δημόκριτος» (Αριθμός αναφοράς: NCSRD-BC report 14/02/2014), σύμφωνα με τη συνθήκη του Ελσίνκι του 1975.

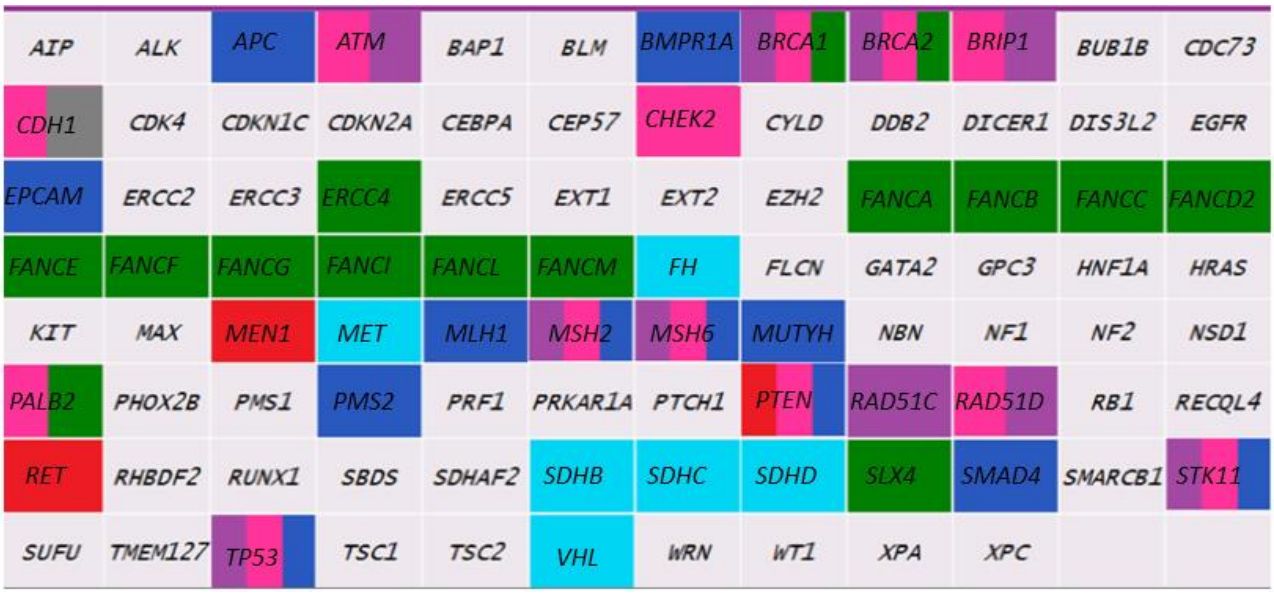
2.1.2. Συλλογή φαινοτυπικών και κλινικών δεδομένων

Για τη συλλογή των στοιχείων του ατομικού και οικογενειακού ιστορικού των εξετασθέντων, διενεργήθηκαν λεπτομερείς συνεντεύξεις από το ΕΜΔ, ενώ κατασκευάστηκε και το οικογενειακό τους δέντρο σε βάθος τριών γενεών. Παράλληλα, από τα άτομα που έχουν νοσήσει ζητήθηκαν οι ιστολογικές εκθέσεις που πιστοποιούν τη διάγνωση του καρκίνου, ώστε να συνεκτιμηθούν και τα ιστοπαθολογικά χαρακτηριστικά του όγκου.

2.1.3. Αλληλούχηση DNA για την ανίχνευση γαμετικών παραλλαγών

Το DNA απομονώθηκε από ολικό περιφερικό αίμα των εξεταζόμενων. Ο γονιδιακός έλεγχος στο DNA τους διεξήχθη με αλληλούχηση κατά Sanger, με Αλληλούχηση Επόμενης Γενιάς με στοχευμένη αλληλούχηση 94 γονιδίων προδιάθεσης σε καρκίνο ή/και με τη μέθοδο πολλαπλής ενίσχυσης ανιχνευτών εξαρτώμενη από την αντίδραση λιγάσης (Multiplex ligation-dependent probe amplification – MLPA) η οποία στοχεύει στην ανίχνευση μεγάλων γονιδιωματικών αναδιατάξεων (διαγραφές/επαναλήψεις), που περιλαμβάνουν ένα ή περισσότερα εξόνια. Πιο συγκεκριμένα, τα άτομα που παραπέμφθηκαν για γονιδιακούς ελέγχους μεταξύ 1999-2012 ελέγχθηκαν αποκλειστικά μέσω αλληλούχησης κατά Sanger, στοχεύοντας γονίδια που προδιαθέτουν στον κλινικό φαινότυπο και στο οικογενειακό ιστορικό του εξεταζόμενου. Μετά το 2012, οι γονιδιακοί έλεγχοι πραγματοποιήθηκαν κυρίως μέσω αλληλούχησης γονιδιακών ομάδων και πιο συγκεκριμένα με τη χρήση του γονιδιακού πάνελ TruSight® Cancer Panel (Illumina®, San Diego, USA) που στοχεύει 94 γονίδια που είναι γνωστά ή

έχουν συσχετιστεί με προδιάθεση στον καρκίνο και 284 κοινές σημειακές παραλλαγές που έχουν συσχετιστεί με καρκίνο μέσω μελετών συσχέτισης ολόκληρου του γονιδιώματος (Εικόνα 28). Η Αλληλούχηση Επόμενης Γενιάς πραγματοποιήθηκε χρησιμοποιώντας αναγνώσεις DNA σε ζεύγη άκρων, μεγέθους 150bp και χημεία v2. Στην περίπτωση που το άτομο είχε σαφή κλινική διάγνωση γενετικού συνδρόμου, πραγματοποιήθηκε στοχευμένη αλληλούχηση του υποψήφιου γονιδίου με τη μέθοδο Sanger. Συνολικά, όλες οι αναλύσεις συμπληρώθηκαν με την τεχνική MLPA για τον εντοπισμό μεγάλων γονιδιακών αναδιατάξεων σε γονίδια σχετικά με τον φαινότυπο του ασθενούς.



Εικόνα 28: Τα 94 γονίδια που περιλαμβάνονται στο γονιδιακό πάνελ TruSight Cancer. Ανάλογα με τον εντοπισμό της κακοήθειας, τα γονίδια έχουν χρωματιστεί και με διαφορετικό χρώμα. Με ροζ παρουσιάζονται τα γονίδια προδιάθεσης στον καρκίνο του μαστού, με μωβ σε καρκίνο των ωοθηκών, με μπλε σε καρκίνο του παχέος εντέρου και πολυποδίαση, με κόκκινο σε καρκίνο του θυρεοειδούς, με γαλάζιο σε καρκίνο των νεφρών και με πράσινο τα γονίδια που σχετίζονται με την αναιμία Fanconi.

Figure 28: The 94 genes included in the TruSight Cancer gene panel. The gene symbols are colored differently, according to the location of the malignancy. Thus, in pink are colored the breast cancer predisposition genes, in purple are the ovarian cancer predisposition genes, in blue are the colorectal cancer and polyposis predisposition genes, in red are the thyroid cancer predisposition genes, in blue are the kidney cancer predisposition genes and in are green the Fanconi anemia related genes.

2.1.4. Βιοπληροφορική ανάλυση δεδομένων από Αλληλούχηση Επόμενης Γενιάς DNA γαμετικής σειράς

Η βιοπληροφορική ανάλυση πραγματοποιήθηκε χρησιμοποιώντας τη ροή διοχέτευσης εντολών διεργασιών MiSeq Reporter Enrichment της Illumina, η οποία χρησιμοποιεί το λογισμικό BWA (H. Li & Durbin, 2009) για την αντιστοίχιση των αναγνώσεων στο ανθρώπινο γονιδίωμα αναφοράς (GRCh37/hg19) και το λογισμικό Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010) για την κλήση παραλλαγών. Στη συνέχεια, στις κληθείσες παραλλαγές εφαρμόστηκαν μία σειρά από φίλτρα βάσει της ποιότητάς τους (ποιότητα γονοτύπου (genotype quality)>40, βάθος ανάγνωσης παραλλαγής>20x). Επιπλέον, απομακρύνθηκαν όλα τα γνωστά τεχνουργήματα που παρουσιάζονται επανειλημμένα κατά τη

διάρκεια των πειραμάτων στο ΕΜΔ. Στο τελικό σύνολο παραλλαγών, συμπεριλήφθηκαν μόνο οι σπάνιες παραλλαγές (συχνότητα αλληλομόρφου στη βάση δεδομένων genome Aggregation Database (gnomAD) σε υγιείς μη Φινλανδούς Ευρωπαίους <0,01).

2.1.5. Επιμέλεια δεδομένων

Η επιμέλεια όλων των δεδομένων, εκτός των γενετικών δεδομένων που παρήχθησαν με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς, πραγματοποιήθηκε χειροκίνητα. Τα δημογραφικά χαρακτηριστικά των ατόμων που περιλαμβάνονται στη βάση δεδομένων προέρχονται από τη συνέντευξη που πραγματοποίησε το ΕΜΔ πριν τον γενετικό έλεγχο. Σε περίπτωση που κάποιο εξεταζόμενο άτομο κατάγεται από μικρό χωριό, στη βάση καταγράφηκε η ευρύτερη περιοχή για λόγους διαφύλαξης της ανωνυμίας.

Η επιμέλεια των φαινοτυπικών δεδομένων και του οικογενειακού ιστορικού πραγματοποιήθηκε εξάγοντας τις αντίστοιχες πληροφορίες από τις ιστοπαθολογικές εκθέσεις και τα λεπτομερή οικογενειακά δέντρα των ασθενών, αντίστοιχα. Ιδιαίτερη σημασία δόθηκε στην εγγραφή των δεδομένων σε σαφώς προσδιορισμένα πεδία και την αποφυγή πεδίων ελεύθερου κειμένου.

Η επιμέλεια των γενετικών δεδομένων πραγματοποιήθηκε χειροκίνητα σε περίπτωση ταυτοποίησης της παραλλαγής με τη μέθοδο Sanger ή MLPA, από το αρχείο των παραλλαγών που διατηρεί το ΕΜΔ. Οι παραλλαγές που ανιχνεύτηκαν με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς εισήχθησαν στη βάση μαζικώς, μέσω του φιλτραρισμένου αρχείου VCF.

Οι γαμετικές σπάνιες παραλλαγές που ανιχνεύτηκαν από την αλληλούχηση DNA κατηγοριοποιήθηκαν βάσει των κανόνων του Αμερικάνικου Κολλεγίου Ιατρικής Γενετικής (American College of Medical Genetics – ACMG) για την ερμηνεία των παραλλαγών (Richards *et al.*, 2015). Η αντιστοίχιση της κατηγοριοποίησης των παραλλαγών βάσει ACMG με την κατηγοριοποίηση στη διεπαφή της βάσης δεδομένων CanVaS έχει ως εξής: οι παθογόνοι παραλλαγές καταγράφονται στη βάση δεδομένων CanVaS ως «Επηρεάζει τη λειτουργία – Affects function», οι πιθανώς παθογόνοι παραλλαγές καταγράφονται ως «Πιθανώς επηρεάζει τη λειτουργία – Probably affects function», οι παραλλαγές αγνώστου σημασίας καταγράφονται ως «Άγνωστη επίπτωση – Effect unknown», οι πιθανώς ουδέτερες παραλλαγές καταγράφονται ως «Πιθανώς δεν επηρεάζει τη λειτουργία – Probably does not affect function» και, τέλος, οι ουδέτερες παραλλαγές καταγράφονται ως «Δεν επηρεάζει τη λειτουργία – Does not affect function».

2.1.6. Υπολογισμός συχνότητας αλληλομόρφων

Για τον υπολογισμό της συχνότητας αλληλομόρφων στον ελληνικό πληθυσμό λήφθηκαν υπόψη τα δεδομένα από την αλληλούχηση DNA ατόμων που δεν έχουν συγγένεια μεταξύ τους και οι οποίοι έχουν εξετασθεί με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς. Έτσι, για κάθε αλληλόμορφο η συχνότητα υπολογίστηκε ως εξής:

$$AF = AC/AN$$

Όπου AF η συχνότητα αλληλομόρφου, AC ο αριθμός των χρωμοσωμάτων που φέρουν την παραλλαγή και AN ο συνολικός αριθμός των χρωμοσωμάτων, των οποίων η ακολουθία στη συγκεκριμένη θέση έχει προσδιοριστεί με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς. Το AN αυξάνεται κατά 2 για κάθε άτομο που έχει εξεταστεί για τη συγκεκριμένη παραλλαγή με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς. Το AC αυξάνεται κατά 2 για κάθε άτομο που φέρει την παραλλαγή σε ομοζυγωτία, κατά 1 για κάθε άτομο που φέρει την παραλλαγή σε ετεροζυγωτία και καθόλου αν κάποιο άτομο δε φέρει την παραλλαγή.

2.1.7. Λογισμικό βάσης δεδομένων

Το λογισμικό βάσης δεδομένων ανοιχτού κώδικα Leiden Open Variation Database (LOVD) v3.0 (Fokkema *et al.*, 2011) χρησιμοποιήθηκε για την ενσωμάτωση των δεδομένων, καθώς επιτρέπει την αυτόματη κοινή χρήση των δεδομένων με την κεντρική εγκατάσταση LOVD. Για να λειτουργήσει το λογισμικό LOVD χρειάζεται έναν εξυπηρετητή διαδικτύου Apache, τη γλώσσα σεναρίου PHP για τη διασύνδεση της διεπαφής χρήστη με τη βάση δεδομένων και το σύστημα βάσεων δεδομένων MySQL. Το λογισμικό LOVD προσφέρει τη δυνατότητα δημιουργίας προσαρμοσμένων στηλών στους ήδη υπάρχοντες πίνακες της βάσης χωρίς την ανάπτυξη κώδικα. Καθώς ο υπολογισμός της συχνότητας αλληλομόρφων είναι μια δυναμική διαδικασία για κάθε παραλλαγή, το συγκεκριμένο πεδίο δημιουργήθηκε χρησιμοποιώντας τη γλώσσα σεναρίου PHP. Οι προσαρμοσμένες στήλες που δημιουργήθηκαν για τη βάση δεδομένων CanVaS παρουσιάζονται στον Πίνακα 7.

Πίνακας 7: Περιγραφή προσαρμοσμένων στηλών στην εγκατάσταση του CanVaS. Οι προσαρμοσμένες στήλες είτε διατέθηκαν από το λογισμικό LOVD είτε καθορίστηκαν κατά την ανάπτυξη του CanVaS.

Table 7: Description of custom columns in CanVaS installation. Custom columns were either provided by LOVD software or specified during CanVaS development.

Στήλη	**Περιγραφή**	**Πρόσβαση**
Individual/Consanguinity†	Υποδεικνύει αν οι γονείς του εξεταζόμενου ατόμου έχουν συγγένεια εξ αίματος	Δημόσια
Individual/Gender†	Το φύλο του συγκεκριμένου ατόμου	Δημόσια
Individual/Origin/Geographic†	Η γεωγραφική καταγωγή του Γονέα #1 του ατόμου	Δημόσια
Individual/Origin/Geographic2‡	Η γεωγραφική καταγωγή του Γονέα #2 του ατόμου	Δημόσια
Individual/Origin/Population†	Ο πληθυσμός στον οποίο ανήκει το άτομο	Δημόσια
Individual/Remarks†	Παρατηρήσεις σχετικές με το άτομο	Δημόσια
Phenotype/Additional†	Συμπληρωματικές πληροφορίες σχετικά με τον φαινότυπο	Δημόσια
Phenotype/Age/Onset†	Η ηλικία κατά την οποία εμφανίστηκαν τα πρώτα συμπτώματα, αν είναι γνωστή	Δημόσια
Phenotype/Behaviour‡	Η συμπεριφορά του όγκου (διαχύτου τύπου/οριοθετημένο)	Ελεγχόμενη

Phenotype/BrCa/FH‡	Οικογενειακό ιστορικό καρκίνου μαστού, ωοθηκών ή/και παγκρέατος. Όχι = κανένας καρκίνος στην οικογένεια, Αδύναμο = ένας καρκίνος στην οικογένεια, Σοβαρό = περισσότεροι από έναν καρκίνοι στην οικογένεια. Μόνο για φαινότυπους καρκίνου μαστού.	Ελεγχόμενη
Phenotype/CRC/CRC_FH‡	Οικογενειακό ιστορικό καρκίνου παχέος εντέρου. Μόνο για φαινότυπους καρκίνου παχέος εντέρου.	Ελεγχόμενη
Phenotype/CRC/Lynch_FH‡	Οικογενειακό ιστορικό καρκίνων που συνδέονται με το σύνδρομο Lynch. Μόνο για φαινότυπους καρκίνου παχέος εντέρου.	Ελεγχόμενη
Phenotype/CRC/Polyposis_FH‡	Οικογενειακό ιστορικό πολυποδιάσεων. Μόνο για φαινότυπους καρκίνου παχέος εντέρου.	Ελεγχόμενη
Phenotype/ER‡	Κατάσταση υποδοχέων οιστρογόνων.	Ελεγχόμενη
Phenotype/PRC/Gleason‡	Δείκτης Gleason. Μόνο για φαινότυπους καρκίνου του προστάτη.	Ελεγχόμενη
Phenotype/Grade‡	Βαθμός όγκου	Ελεγχόμενη
Phenotype/HER2‡	Κατάσταση πρωτεΐνης HER2.	Ελεγχόμενη
Phenotype/LN‡	Θετικότητα λεμφαδένων	Ελεγχόμενη
Phenotype/Morphology‡	Ιστολογία όγκου (πχ πορογενές, λοβιακό κλπ.)	Ελεγχόμενη
Phenotype/MSI‡	Μικροδορυφορική αστάθεια.	Ελεγχόμενη
Phenotype/OvCa/FH‡	Οικογενειακό ιστορικό καρκίνου μαστού, ωοθηκών ή/και παγκρέατος. Μόνο για φαινότυπους καρκίνου ωοθηκών.	Ελεγχόμενη
Phenotype/PR‡	Κατάσταση υποδοχέων προγεστερόνης.	Ελεγχόμενη
Phenotype/Stage‡	Στάδιο όγκου.	Ελεγχόμενη
Phenotype/Tissue‡	Ιστός στον οποίο αναπτύχθηκε ο όγκος.	Ελεγχόμενη
Phenotype/Subtype‡	Υπότυπος όγκου.	Ελεγχόμενη
Screening/Technique†	Τεχνική που χρησιμοποιήθηκε για την ανίχνευση της παραλλαγής.	Δημόσια
VariantOnGenome/AKA‡	Ονοματολογία παραλλαγής με την οποία είναι γνωστή και η οποία δεν ακολουθεί τους κανόνες HGVS.	Δημόσια
VariantOnGenome/Allele_Count‡	Πλήθος κάθε εναλλακτικού αλληλομόρφου σε κάθε θέση στο γονιδίωμα σε όλα τα άτομα χωρίς συγγενική σχέση.	Δημόσια
VariantOnGenome/Allele_Number‡	Συνολικός αριθμός αλληλομόρφων σε κάθε θέση στο γονιδίωμα σε όλα τα άτομα χωρίς συγγενική σχέση.	Δημόσια
VariantOnGenome/dbSNP†	Ο κωδικός αναφοράς της παραλλαγής στη βάση dbSNP.	Δημόσια
VariantOnGenome/DNA†	Περιγραφή της παραλλαγής σε επίπεδο DNA, βασισμένη στην ακολουθία αναφοράς του γενωμικού DNA (ακολουθώντας τους κανόνες HGVS).	Δημόσια

VariantOnGenome/GeographicOrigin‡	Εντοπιότητα παραλλαγής.	Δημόσια
VariantOnGenome/IsGreekFounder‡	Υποδεικνύει αν η παραλλαγή είναι ιδρυτική για τον ελληνικό πληθυσμό.	Δημόσια
VariantOnGenome/Reference†	Αναφορά στη δημοσίευση στην οποία περιγράφεται η παραλλαγή.	Δημόσια
VariantOnGenome/Remarks†	Παρατηρήσεις σχετικές με την παραλλαγή.	Δημόσια
VariantOnGenome/Segregation†	Υποδεικνύει εάν η παραλλαγή διαχωρίζεται με τον φαινότυπο	Δημόσια
VariantOnTranscript/DNA†	Περιγραφή της παραλλαγής σε επίπεδο DNA, βασισμένη στην ακολουθία αναφοράς του κωδικού DNA (ακολουθώντας τους κανόνες HGVS).	Δημόσια
VariantOnTranscript/Exon†	Ο αριθμός του εξονίου/ιντρονίου όπου εδράζεται η παραλλαγή.	Δημόσια
VariantOnTranscript/Protein†	Περιγραφή της παραλλαγής σε επίπεδο πρωτεΐνης (ακολουθώντας τους κανόνες HGVS).	Δημόσια
VariantOnTranscript/RNA†	Περιγραφή της παραλλαγής σε επίπεδο RNA (ακολουθώντας τους κανόνες HGVS).	Δημόσια

† Προσαρμοσμένες στήλες που διατίθενται από το LOVD.

‡ Προσαρμοσμένες στήλες CanVaS.

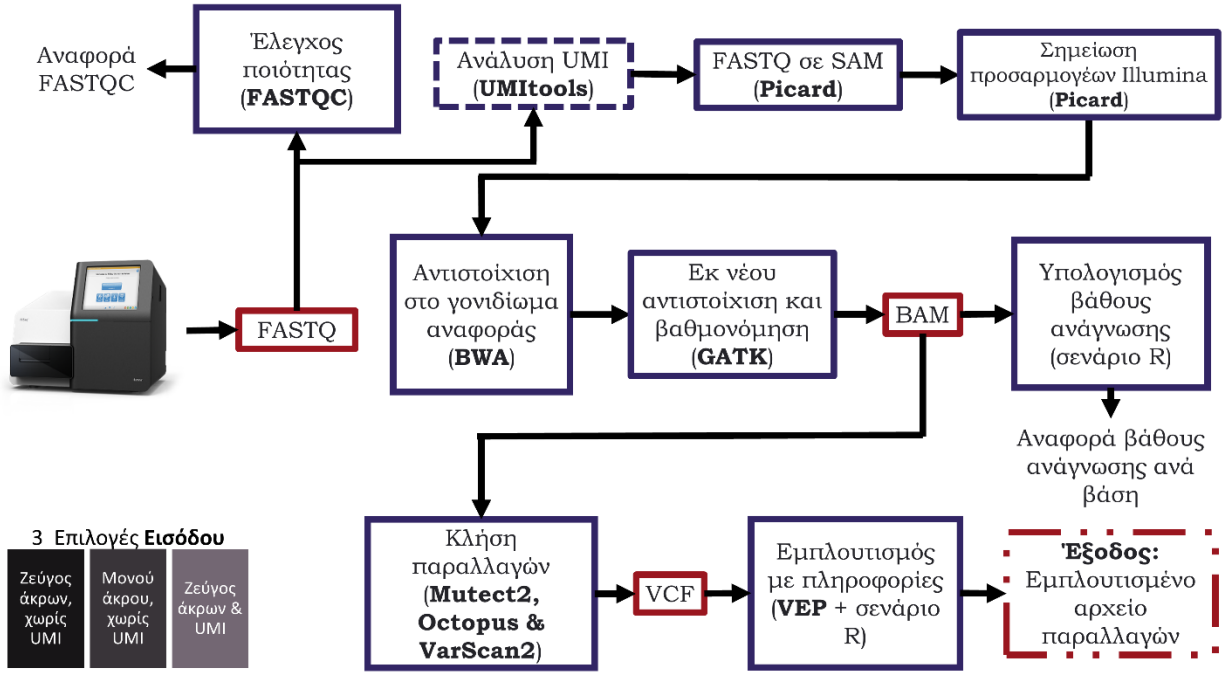
2.2. Ανάπτυξη ροής διοχέτευσης εντολών διεργασιών VarTrace

Το VarTrace είναι μια αυτοματοποιημένη, πλήρως διαμορφώσιμη και παράλληλη ροή διοχέτευσης εντολών διεργασιών που αναπτύχθηκε με σκοπό την ακριβή ανίχνευση πραγματικών σωματικών παραλλαγών σε DNA όγκων που διατηρούνται σε παραφίνη μετά από σταθεροποίηση με φορμαλίνη (Formalin-Fixed, Paraffin-Embedded - FFPE). Τα δεδομένα που παίρνει ως είσοδο το VarTrace πρέπει να έχουν παραχθεί σε πλατφόρμα αλληλούχησης Illumina. Το VarTrace είναι γραμμένο σε γλώσσες R και Perl. Η ροή διοχέτευσης εντολών διεργασιών VarTrace είναι διαθέσιμη στο URL: https://gitlab.com/dkalfakakou/TumorPipeline.

2.2.1. Χρησιμοποιούμενα εργαλεία

Η ροή διοχέτευσης εντολών διεργασιών VarTrace χρησιμοποιεί και ενσωματώνει σε διάφορα βήματα τα ακόλουθα εργαλεία: FastQC (Andrews, 2010), Picard (2019), UMI-tools (Smith *et al.*, 2017), BWA (H. Li & Durbin, 2009), GATK (McKenna *et al.*, 2010), Mutect2 (Benjamin *et al.*, 2019), Octopus (Cooke *et al.*, 2021), Varscan2 (Koboldt *et al.*, 2012) και Variant Effect Predictor (VEP) (McLaren *et al.*, 2016). Το FastQC χρησιμοποιείται για τον έλεγχο της ποιότητας του πειράματος, παρέχοντας μία τυποποιημένη και λεπτομερή αναφορά για την ποιότητα των αναγνώσεων DNA. Το Picard είναι ένα σύνολο εργαλείων γραμμής εντολών ανεπτυγμένο από το Ινστιτούτο Broad των Massachusetts Institute of Technology (MIT) και Harvard, που χρησιμοποιείται για τη διαχείριση αρχείων που έχουν παραχθεί από πειράματα Αλληλούχησης Επόμενης Γενιάς. Κατά τη διάρκεια της ανάλυσης χρησιμοποιούνται πολλά από τα εργαλεία Picard. Το UMI-tools ανιχνεύει τους Μοναδικούς Μοριακούς Κωδικούς (Unique Molecular

Identifiers – UMI), εάν το πείραμα είναι σχεδιασμένο κατά αυτόν τον τρόπο, και αφαιρεί τις διπλές αναγνώσεις DNA που δημιουργήθηκαν κατά το στάδιο της αλυσιδωτή αντίδρασης πολυμεράσης (Polymerase Chain Reaction – PCR) της προετοιμασίας των βιβλιοθηκών, με σκοπό την εξάλειψη των τεχνουργημάτων. Το εργαλείο BWA υλοποιεί τον αλγόριθμο Burrows-Wheeler για την αντιστοίχιση των αναγνώσεων σε κάποιο γονιδίωμα αναφοράς. Το σύνολο εργαλείων GATK έχει επίσης αναπτυχθεί από το Ινστιτούτο Broad των MIT και Harvard και παρέχει μία σειρά από εργαλεία τα οποία εστιάζουν στην ανίχνευση παραλλαγών, συμπεριλαμβανομένου του Mutect2 που είναι το εργαλείο που χρησιμοποιείται για την κλήση σημειακών παραλλαγών και μικρών ενθέσεων και απαλοιφών από δεδομένα προερχόμενα από αλληλούχηση DNA σωματικής σειράς. Εκτός από το Mutect2, στο VarTrace χρησιμοποιούνται και τα εργαλεία κλήσης παραλλαγών Octopus και VarScan2. Τέλος, το VarTrace χρησιμοποιεί το VEP για τον εμπλουτισμό των παραλλαγών με πληροφορίες. Στην Εικόνα 29 παρουσιάζεται μία σχηματική απεικόνιση των βημάτων που ακολουθούνται στην ανάλυση με το VarTrace.



Εικόνα 29: Σχηματική απεικόνιση της ροής διοχέτευσης εντολών διεργασιών VarTrace.

Figure 29: Schematic representation of VarTrace pipeline.

2.2.2. Περιγραφή VarTrace

2.2.2.1. Δεδομένα εισόδου

Τα δεδομένα που δέχεται η ροή διοχέτευσης εντολών διεργασιών ως είσοδο είναι αρχεία FASTQ τα οποία έχουν παραχθεί σε πλατφόρμα αλληλούχησης DNA επόμενης γενιάς Illumina, με οποιοδήποτε συμβατό κιτ δημιουργίας βιβλιοθηκών. Τα αρχεία μπορεί να προέρχονται είτε από αλληλούχηση ζεύγους άκρων είτε από αλληλούχηση μονού άκρου, ενώ ο χρήστης πρέπει να το ορίσει κατά την

παραμετροποίηση του VarTrace. Επιπλέον, το VarTrace παίρνει ως είσοδο το αρχείο FASTA με την ακολουθία του ανθρώπινου γονιδιώματος κι ένα αρχείο BED (Browser Extensible Data) με τις συντεταγμένες των περιοχών που στοχεύονται από το γονιδιακό πάνελ.

2.2.2.2. Έλεγχος ποιότητας

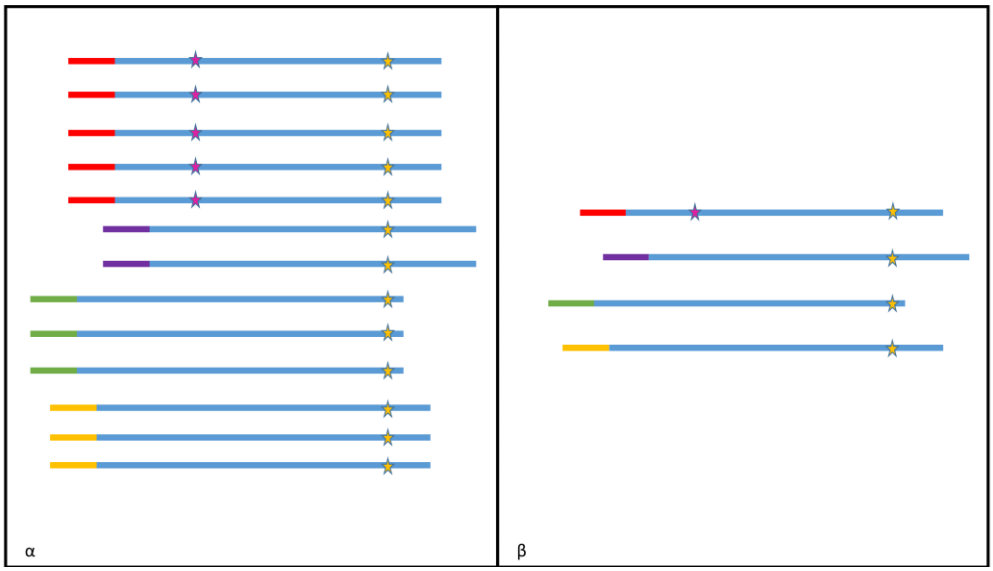
Το πρώτο βήμα στη ροή διοχέτευσης εντολών διεργασιών είναι ο έλεγχος ποιότητας, που εκτελείται με την εφαρμογή FastQC (Andrews, 2010). Το εργαλείο FastQC παρέχει μια λεπτομερή αναφορά η οποία συνοψίζει την αξιολόγηση των δεδομένων παρέχοντας στατιστικά στοιχεία και γραφήματα, προτού ξεκινήσουν τα εντατικά υπολογιστικά μέρη της ανάλυσης.

2.2.2.3. Προεπεξεργασία

Στη συνέχεια της ανάλυσης, πραγματοποιείται μία προεπεξεργασία των δεδομένων, πριν την αντιστοίχισή τους στο γονιδίωμα αναφοράς. Κατά το πρώτο βήμα της προεπεξεργασίας, πραγματοποιείται προαιρετικά η ανάλυση των UMI των αναγνώσεων, αν το πείραμα περιλαμβάνει UMI, με τη χρήση του εργαλείου UMI-tools (Smith *et al.*, 2017). Οι μοναδικοί μοριακοί κωδικοί χρησιμοποιούνται ώστε να επισημάνουν τις αναγνώσεις DNA πριν τον πολλαπλασιασμό τους με τη μέθοδο PCR, έτσι ώστε να διευκολυνθεί κατά τη βιοπληροφορική ανάλυση ο εντοπισμός τεχνουργημάτων που εισήχθησαν από την τεχνική PCR (Εικόνα 30). Το UMI-tools αναγνωρίζει τα UMI, τα αφαιρεί από την ακολουθία και τα τοποθετεί στο όνομα της ανάγνωσης στο FASTQ αρχείο. Σε επόμενο βήμα, και πιο συγκεκριμένα μετά την αντιστοίχιση των αναγνώσεων στο γονιδίωμα αναφοράς, το UMI-tools ομαδοποιεί τις αναγνώσεις που έχουν το ίδιο UMI.

Ακολούθως, τα αρχεία FASTQ μετατρέπονται σε μη αντιστοιχισμένα BAM αρχεία (unaligned BAM – uBAM), χρησιμοποιώντας το εργαλείο FastqToSam της εργαλειοθήκης Picard (BroadInstitute, 2019). Ο λόγος που πραγματοποιείται αυτό το βήμα, είναι ότι τα αρχεία SAM, και κατ' επέκταση η δυαδική μορφή τους, BAM, μπορούν να αποθηκεύσουν περισσότερες πληροφορίες σχετικά με τους μοριακούς κωδικούς και την ποιότητα των αναγνώσεων. Αυτές οι πληροφορίες είναι χρήσιμες στα επόμενα βήματα.

Στο τελευταίο βήμα της προεπεξεργασίας, πραγματοποιείται σημείωση των προσαρμογέων της Illumina με σκοπό την αύξηση της ακρίβειας κατά την αντιστοίχιση των αναγνώσεων στο γονιδίωμα αναφοράς. Το βήμα αυτό πραγματοποιείται με τη χρήση του εργαλείου MarkIlluminaAdapters της εργαλειοθήκης Picard.



Εικόνα 30: Η λειτουργία των UMI. Στο α. οι αναγνώσεις που προέρχονται από το ίδιο θραύσμα DNA (ίδιο χρώμα στο 5’ άκρο, δηλαδή ίδιο UMI) δεν έχουν ομαδοποιηθεί και φαίνεται πως η παραλλαγή που αντιπροσωπεύεται με το ροζ αστέρι εντοπίζεται στο 39% των αναγνώσεων. Στο β. οι αναγνώσεις που προέρχονται από το ίδιο θραύσμα DNA έχουν ομαδοποιηθεί και φαίνεται πως η παραλλαγή που αντιπροσωπεύεται με το ροζ αστέρι εντοπίζεται μόνο σε μία ανάγνωση DNA, το οποίο αποτελεί ένδειξη ότι ενδεχομένως είναι τεχνούργημα.

Figure 30: The function of UMIs. In a: reads from the same DNA fragment (same color at the 5' end, i.e. same UMI) have not been grouped and it appears that the variant represented by the pink star is found in 39% of the readings. In b: The reads from the same DNA fragment have been grouped and it appears that the variant represented by the pink star is found in only one DNA reading, which is an indication that it might be an artifact.

2.2.2.4. Αντιστοίχιση στο γονιδίωμα αναφοράς και επεξεργασία αρχείων BAM

Τα αρχεία που έχουν παραχθεί από τα βήματα της προεπεξεργασίας είναι πλέον έτοιμα για αντιστοίχιση στο γονιδίωμα αναφοράς. Η διαδικασία αυτή πραγματοποιείται από το εργαλείο BWA και συγκεκριμένα από το BWA-MEM (H. Li & Durbin, 2009), το οποίο είναι ιδανικό για μικρές αναγνώσεις μεγέθους μεγαλύτερου από 70 ζεύγη βάσεων, όπως είναι οι αναγνώσεις που παράγονται από τις πλατφόρμες Αλληλούχησης Επόμενης Γενιάς της Illumina.

Μετά την αντιστοίχιση στο γονιδίωμα αναφοράς, πραγματοποιείται μία βελτιστοποίηση της αντιστοίχισης χρησιμοποιώντας το εργαλείο IndelRealigner της εργαλειοθήκης GATK. Το εργαλείο αυτό, πραγματοποιεί τοπική εκ νέου τοπική αντιστοίχιση γύρω από τις περιοχές που παρουσιάζουν έντονη ποικιλομορφία σε σχέση με το γονιδίωμα αναφοράς. Επομένως, επιτρέπει τη διόρθωση σφαλμάτων που έγιναν κατά την κύρια αντιστοίχιση των αναγνώσεων DNA, με αποτέλεσμα την πιο έγκυρη αντιστοίχιση στις περιοχές με μικρές ενθέσεις και απαλοιφές.

Ακολούθως, πραγματοποιείται εκ νέου βαθμονόμηση των βάσεων, σύμφωνα με τα δεδομένα που έχουν παραχθεί από την ευθυγράμμιση. Το βήμα αυτό πραγματοποιείται με το εργαλείο BaseRecalibrator του GATK. Το εργαλείο αυτό εφαρμόζει μεθόδους μηχανικής μάθησης για να εντοπίζει

συστηματικά σφάλματα στη βαθμονόμηση των βάσεων από την πλατφόρμα Αλληλούχησης Επόμενης Γενιάς και να την προσαρμόσει αναλόγως.

2.2.2.5. Υπολογισμός περιοχών με μικρό βάθος ανάγνωσης

Το επόμενο βήμα έχει ως έξοδο ένα ενδιάμεσο αρχείο ελέγχου ποιότητας. Για τον υπολογισμό των περιοχών με μικρό βάθος ανάγνωσης αναπτύχθηκε ένα προσαρμοσμένο σενάριο R, το οποίο λαμβάνει ως είσοδο το αρχείο BAM και δύο κατώφλια, ένα για τον έλεγχο του ελάχιστου βάθους ανάγνωσης (Coverage threshold – CT) κι ένα για τον έλεγχο της ελάχιστης ποιότητας αντιστοίχισης (Quality Threshold – QT). Η έξοδος του σεναρίου είναι ένα αρχείο BED το οποίο περιέχει όλες τις περιοχές στις οποίες έχει αντιστοιχηθεί μικρότερο από CT πλήθος αναγνώσεων οι οποίες έχουν ποιότητα αντιστοίχισης μεγαλύτερη από QT.

2.2.2.6. Κλήση παραλλαγών

Καθώς η κλήση των παραλλαγών από δεδομένα που προέρχονται από αλληλούχηση DNA όγκων είναι εξαιρετικά απαιτητική και υπάρχει μεγάλη ασυμφωνία μεταξύ των αλγορίθμων κλήσης παραλλαγών (Q. Wang *et al.*, 2019; Kroigard *et al.*, 2016), στο VarTrace χρησιμοποιούνται τρεις αλγόριθμοι κλήσης παραλλαγών και συγκεκριμένα οι Mutect2 (Benjamin *et al.*, 2019), Octopus (Cooke *et al.*, 2021) και VarScan2 (Koboldt *et al.*, 2012). Για κάθε αρχείο κλήσης παραλλαγών εφαρμόζεται μία σειρά φίλτρων βάσει της ποιότητας των παραλλαγών, για την εξάλειψη των ψευδώς θετικών ευρημάτων. Τα αρχεία που εξάγονται από το κάθε εργαλείο συνδυάζονται σε ένα αρχείο με τη χρήση του CombineVariants της εργαλειοθήκης GATK.

2.2.2.7. Εμπλουτισμός παραλλαγών με πληροφορίες

Το τελευταίο βήμα της ανάλυσης με τη ροή διοχέτευσης διεργασιών VarTrace περιλαμβάνει τον εμπλουτισμό των παραλλαγών με πληροφορίες. Για τη διαδικασία του εμπλουτισμού χρησιμοποιείται το εργαλείο VEP της Ensembl (McLaren *et al.*, 2016). Στη συνέχεια, το αρχείο εξόδου του VEP υπόκειται σε περαιτέρω επεξεργασία με τη χρήση προσαρμοσμένων σεναρίων R και Perl, ώστε να εξαχθούν όλες οι πληροφορίες που χρειάζονται για την αξιολόγηση των παραλλαγών και η τελική μορφή του να είναι αρχείου κειμένου οριοθετημένου με στηλοθέτες.

2.2.3. Αξιολόγηση της ροής διοχέτευσης εντολών διεργασιών VarTrace

2.2.3.1. Ομάδα ελέγχου

Τα δείγματα που χρησιμοποιήθηκαν για την αξιολόγηση της ροής διοχέτευσης εντολών διεργασιών VarTrace προέρχονται από FFPE όγκους 75 ασθενών με επιθηλιακό καρκίνωμα των ωοθηκών που προηγουμένως είχαν ελεγχθεί για γαμετικές παθογόνους παραλλαγές σε γονίδια προδιάθεσης στον καρκίνο των ωοθηκών και είχαν αρνητικό αποτέλεσμα.

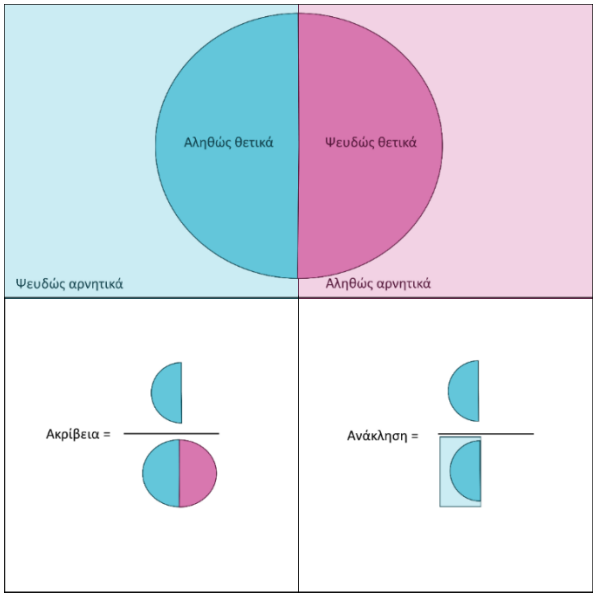
2.2.3.2. Αλληλούχηση DNA

Το DNA των όγκων απομονώθηκε από FFPE δείγματα καρκίνου των ωοθηκών χρησιμοποιώντας το κιτ ιστού QIAamp® DNA FFPE, σύμφωνα με τις οδηγίες του κατασκευαστή (Qiagen, Dusseldorf, Γερμανία). Όλοι οι όγκοι των ωοθηκών εκτιμήθηκαν από έναν παθολογοανατόμο για την περιεκτικότητα τους σε καρκινικά κύτταρα (ΠΚΚ) και νέκρωση, ενώ τα κατώφλια ποιότητας του ιστού ορίστηκαν σε >50% και <20% για ΠΚΚ και νέκρωση, αντίστοιχα.

2.2.3.3. Σύγκριση αποτελεσμάτων

Για την αξιολόγηση της ροής διοχέτευσης εντολών διεργασιών VarTrace, πραγματοποιήθηκε σύγκριση με το εμπορικό πακέτο λογισμικού Biomedical Genomics Workbench (BGWB). Τα αρχεία FASTQ που παρήχθησαν από την Αλληλούχηση Επόμενης Γενιάς, αναλύθηκαν και με τα δύο εργαλεία. Όσον αφορά την ανάλυση των δεδομένων με το BGWB, χρησιμοποιήθηκαν δύο ξεχωριστές ροές διοχέτευσης εντολών διεργασιών· η ροή διοχέτευσης εντολών διεργασιών με τις προκαθορισμένες παραμέτρους του προγράμματος (BGWBdef) και η ροή διοχέτευσης εντολών διεργασιών που λαμβάνει υπόψη τα UMI (BGWBumi), η οποία ήταν και η προτεινόμενη από την εταιρία ροή. Το VarTrace δοκιμάστηκε χωρίς την προαιρετική παράμετρο για την ανάλυση των UMI.

Στη συνέχεια, οι παραλλαγές που ανιχνεύτηκαν τουλάχιστον από μία εκ των τριών ροών και οι οποίες πληρούσαν κάποια κριτήρια, επιβεβαιώθηκαν με τη μέθοδο Sanger. Οι παραλλαγές που ελέγχθηκαν έπρεπε να είναι σπάνιες, παθογόνοι/πιθανώς παθογόνοι παραλλαγές και να έχουν συχνότητα αλληλομόρφου παραλλαγής (Variant Allele Frequency – VAF) μεγαλύτερη του 10%, ενώ το βάθος ανάγνωσης στη γενωμική θέση της παραλλαγής έπρεπε να είναι μεγαλύτερο του 50x.



Εικόνα 30: Υπολογισμός μετρήσεων ακρίβειας και ανάκλησης.

Figure 31: Precision and recall metrics calculation.

Στη συνέχεια, έχοντας στη διάθεσή μας το σύνολο των πραγματικών παραλλαγών μετά την επιβεβαίωσή τους με τη μέθοδο Sanger, υπολογίστηκαν δύο μετρήσεις για την αξιολόγηση των ροών και τελικά, τη σύγκρισή τους· η **ακρίβεια (precision)** και η **ανάκληση (recall).** Η ακρίβεια είναι ο λόγος των παραλλαγών που χαρακτηρίστηκαν ως πραγματικές από έναν αλγόριθμο και ήταν πραγματικές, διά του συνολικού αριθμού των παραλλαγών που χαρακτηρίστηκαν ως πραγματικές, ενώ η ανάκληση είναι ο λόγος των παραλλαγών που χαρακτηρίστηκαν ως πραγματικές από έναν αλγόριθμο και ήταν πραγματικές διά του συνολικού αριθμού των πραγματικών παραλλαγών (Εικόνα 31). Ουσιαστικά, η ακρίβεια παρέχει μία εκτίμηση του ποσοστού των παραλλαγών που είναι πραγματικές και ανιχνεύτηκαν και η ανάκληση μία εκτίμηση του ποσοστού των παραλλαγών που χαρακτηρίστηκαν εσφαλμένα ως πραγματικές. Όσο πιο κοντά στη μονάδα είναι και οι δύο μετρήσεις, τόσο πιο ακριβής είναι ο αλγόριθμος.

2.3. Αξιολόγηση εμπορικού λογισμικού για τον εμπλουτισμό παραλλαγών με πληροφορίες

Η συγκεκριμένη μελέτη αφορά τη σύγκριση του χαρακτηρισμού παραλλαγών ως προς την επίπτωσή τους, όπως πραγματοποιήθηκε από δύο λογισμικά, το Illumina® VariantStudio v3.0 (VS) και το VEP, ώστε να επισημανθούν οι διαφορές που ενδέχεται να έχουν αντίκτυπο στην κλινική ερμηνεία των παραλλαγών.

2.3.1. Δεδομένα

Τα δεδομένα παρήχθησαν από την αλληλούχηση DNA γαμετικής σειράς 857 ασθενών με καρκίνο που παραπέμφθηκαν για γονιδιακό έλεγχο στο ΕΜΔ μεταξύ του Απρίλιου του 2017 και του Νοεμβρίου του 2018. Από αυτούς, 615 ήταν ασθενείς με καρκίνο του μαστού ή/και των ωοθηκών, 57 ήταν ασθενείς με καρκίνο του παχέος εντέρου, ενώ 185 είχαν άλλους τύπους καρκινικών ή προ-καρκινικών διαγνώσεων, ικανοποιώντας τα κριτήρια επιλογής για γονιδιακό έλεγχο.

2.3.2. Αλληλούχηση DNA

Το DNA απομονώθηκε από ολικό περιφερικό αίμα των εξεταζόμενων. Οι βιβλιοθήκες που παρασκευάστηκαν από το DNA γαμετικής σειράς των ασθενών ακολουθώντας τις τυπικές διαδικασίες, αλληλουχήθηκαν με τη χρήση του αναλυτή επόμενης γενιάς Illumina® MiSeq®, χρησιμοποιώντας το πάνελ TruSight® Cancer Panel (Illumina®, San Diego, USA). Η αλληλούχηση πραγματοποιήθηκε χρησιμοποιώντας αναγνώσεις DNA σε ζεύγη, μεγέθους 150bp και χημεία v2.

2.3.3. Βιοπληροφορική ανάλυση δεδομένων από Αλληλούχηση Επόμενης Γενιάς DNA γαμετικής σειράς

Η βιοπληροφορική ανάλυση πραγματοποιήθηκε χρησιμοποιώντας τη ροή διοχέτευσης εντολών διεργασιών MiSeq Reporter Enrichment, η οποία χρησιμοποιεί το λογισμικό BWA (H. Li & Durbin, 2009) για την αντιστοίχιση των αναγνώσεων στο ανθρώπινο γονιδίωμα αναφοράς (GRCh37/hg19) και το

λογισμικό GATK (McKenna *et al.*, 2010) για την κλήση παραλλαγών. Στη συνέχεια, οι κληθείσες παραλλαγές φιλτραρίστηκαν βάσει της ποιότητάς τους (ποιότητα γονοτύπου (genotype quality)>40, βάθος ανάγνωσης παραλλαγής>20x). Επιπλέον, απομακρύνθηκαν και όλα τα γνωστά τεχνουργήματα που παρουσιάζονται επανειλημμένα κατά τη διάρκεια των πειραμάτων στο ΕΜΔ.

2.3.4. Εμπλουτισμός παραλλαγών με πληροφορίες

Οι παραλλαγές σχολιάστηκαν με τη χρήση των προγραμμάτων VS v3.0 και VEP v89. Για το VS, χρησιμοποιήθηκαν οι προεπιλεγμένες ρυθμίσεις για το χαρακτηρισμό των παραλλαγών, δηλαδή με βάση το μετάγραφο με το μεγαλύτερο μήκος στην περιοχή, λόγω περιορισμών του λογισμικού. Από το λογισμικό εξήχθη ένα αρχείο κειμένου οριοθετημένο με στηλοθέτες (tab delimited) για περαιτέρω ανάλυση.

2.3.5. Επιλογή κανονικών μεταγράφων

Για τον εμπλουτισμό των παραλλαγών με το VEP επιλέχθηκε ένα σύνολο κανονικών μεταγράφων των γονιδίων που περιλαμβάνονται στο πάνελ TruSight Cancer, χρησιμοποιώντας τη βάση δεδομένων APPRIS (Rodriguez *et al.*, 2018). Τα μετάγραφα επιλέχθηκαν από το σύνολο δεδομένων RefSeq (Pruitt *et al.*, 2007), και συγκεκριμένα από το RefSeq105, το οποίο είναι συμβατό με το ανθρώπινο γονιδίωμα GRCh37/hg19, προκειμένου να είναι δυνατή η άμεση σύγκριση με το VS. Σε περίπτωση πολλαπλών κανονικών μεταγράφων, χρησιμοποιήθηκε το μεγαλύτερο. Το σύνολο των επιλεγμένων μεταγράφων συνοψίζεται στον Πίνακα 8.

Πίνακας 8: Επιλεγμένα μετάγραφα για κάθε υπό μελέτη γονίδιο για τον εμπλουτισμό των παραλλαγών με πληροφορίες με τη χρήση του λογισμικού VEP.

Table 8: Selected transcripts for each gene under study for the annotation of variants using VEP.

Γονίδιο	Μετάγραφο	Γονίδιο	Μετάγραφο
AIP	NM_003977.2	*HRAS*	NM_005343.2
ALK	NM_004304.4	*KIT*	NM_000222.2
APC	NM_000038.5	*MAX*	NM_002382.4
ATM	NM_000051.3	*MEN1*	NM_130799.2
BAP1	NM_004656.3	*MET*	NM_000245.2
BLM	NM_000057.2	*MLH1*	NM_000249.3
BMPR1A	NM_004329.2	*MSH2*	NM_000251.2
BRCA1	NM_007294.3	*MSH6*	NM_000179.2
BRCA2	NM_000059.3	*MUTYH*	NM_001048172.1
BRIP1	NM_032043.2	*NBN*	NM_002485.4
BUB1B	NM_001211.5	*NF1*	NM_000267.3
CDC73	NM_024529.4	*NF2*	NM_000268.3
CDH1	NM_004360.3	*NSD1*	NM_022455.4
CDK4	NM_000075.3	*PALB2*	NM_024675.3
CDKN1C	NM_000076.2	*PHOX2B*	NM_003924.3
CDKN2A	NM_058195.3	*PMS1*	NM_000534.4
CEBPA	NM_004364.3	*PMS2*	NM_000535.5

CEP57	NM_014679.4	*PRF1*	NM_005041.4
CHEK2	NM_007194.3	*PRKAR1A*	NM_002734.4
CYLD	NM_001042412.1	*PTCH1*	NM_000264.3
DDB2	NM_000107.2	*PTEN*	NM_000314.4
DICER1	NM_177438.2	*RAD51C*	NM_058216.2
DIS3L2	NM_152383.4	*RAD51D*	NM_002878.3
EGFR	NM_005228.3	*RB1*	NM_000321.2
EPCAM	NM_002354.2	*RECQL4*	NM_004260.3
ERCC2	NM_000400.3	RET	NM_020975.4
ERCC3	NM_000122.1	RHBDF2	NM_001005498.3
ERCC4	NM_005236.2	RUNX1	NM_001754.4
ERCC5	NM_000123.3	SBDS	NM_016038.2
EXT1	NM_000127.2	SDHAF2	NM_017841.2
EXT2	NM_207122.1	SDHB	NM_003000.2
EZH2	NM_004456.4	SDHC	NM_003001.3
FANCA	NM_000135.2	SDHD	NM_003002.3
FANCB	NM_152633.2	SLX4	NM_032444.2
FANCC	NM_000136.2	SMAD4	NM_005359.5
FANCD2	NM_033084.3	SMARCB1	NM_003073.3
FANCE	NM_021922.2	STK11	NM_000455.4
FANCF	NM_022725.3	SUFU	NM_016169.3
FANCG	NM_004629.1	TMEM127	NM_017849.3
FANCI	NM_001113378.1	TP53	NM_000546.5
FANCL	NM_018062.3	TSC1	NM_000368.4
FANCM	NM_020937.2	TSC2	NM_000548.3
FH	NM_000143.3	VHL	NM_000551.3
FLCN	NM_144997.5	WRN	NM_000553.4
GATA2	NM_032638.4	WT1	NM_024426.4
GPC3	NM_004484.3	XPA	NM_000380.3
HNF1A	NM_000545.5	XPC	NM_004628.4

2.3.6. Σύγκριση εναλλακτικών χαρακτηρισμών

Για τη σύγκριση μεταξύ των εναλλακτικών χαρακτηρισμών των παραλλαγών ως προς την επίπτωσή τους από τα δύο λογισμικά δημιουργήθηκε ένα αρχείο κειμένου οριοθετημένο με στηλοθέτες, χρησιμοποιώντας ειδικά διαμορφωμένο σενάριο ανεπτυγμένο σε γλώσσα R. Τα πεδία που καταγράφονται στο αρχείο περιλαμβάνουν τη θέση της παραλλαγής στο γονιδίωμα, την παραλλαγή, τα μετάγραφα που χρησιμοποιήθηκαν από το VS και το VEP, την ονοματολογία κατά HGVS σε σχέση με το εκάστοτε μετάγραφο και την επίπτωση της παραλλαγής, όπως αναφέρεται και από τα δύο εργαλεία. Οι σημειακές παραλλαγές που αφορούν γενετικούς τόπους που έχουν ανιχνευθεί μέσω μελετών συσχέτισης ολόκληρου του γονιδιώματος αποκλείστηκαν από την ανάλυση. Στη συνέχεια, όλες οι παραλλαγές που βρέθηκε να έχουν διαφορετική επίπτωση από τα δύο λογισμικά απομονώθηκαν.

2.3.7. Χαρακτηρισμός παραλλαγών βάσει της προβλεπόμενης επίπτωσής τους

Τόσο το λογισμικό VS όσο και το VEP χρησιμοποιούν τους όρους Οντολογίας Ακολουθιών για να περιγράψουν την επίπτωση των παραλλαγών, γεγονός που καθιστά τη σύγκριση των δύο συνόλων δεδομένων ευκολότερη. Για τους σκοπούς της συγκεκριμένης μελέτης, οι αναφερόμενες επιπτώσεις των

παραλλαγών χωρίστηκαν σε τέσσερις κατηγορίες: απώλεια λειτουργίας (Loss of function – LoF), περιοχής ματίσματος (splice region), κωδικής περιοχής (coding) και μη κωδικής περιοχής (non-coding). Στην κατηγορία των παραλλαγών απώλειας λειτουργίας, συμπεριλήφθηκαν οι πλαισιοτροποποιητικές ενθέσεις και απαλοιφές και οι παραλλαγές που έχουν ως αποτέλεσμα κέρδος πρόωρου κωδικονίου τερματισμού και απώλεια κωδικονίου έναρξης ή τερματισμού, καθώς συνήθως έχουν σοβαρή επίπτωση στη πρωτεΐνη. Οι παραλλαγές θέσεων ματίσματος, που βρίσκονται σε θέση δότη ή δέκτη ματίσματος, συχνά προκαλούν παρεκκλίνον μάτισμα mRNA. Ωστόσο, η παθογένειά τους δεν μπορεί να επιβεβαιωθεί χωρίς να έχουν δοκιμαστεί σε λειτουργική μελέτη, επομένως μελετήθηκαν ανεξάρτητα. Οι παραλλαγές που βρίσκονται σε κωδική περιοχή περιλαμβάνουν παρανοηματικές και συνώνυμες παραλλαγές, καθώς και μικρές ενθέσεις και απαλοιφές εντός πλαισίου ανάγνωσης. Η επίπτωση των παραλλαγών αυτής της κατηγορίας μπορεί να ποικίλει, με τις περισσότερες παραλλαγές αγνώστου κλινικής σημασίας να εμπίπτουν σε αυτήν την κατηγορία. Οι παραλλαγές που βρίσκονται σε μη κωδικές περιοχές, περιλαμβάνουν παραλλαγές που εντοπίζονται σε ιντρόνια, στις 3’ και 5’ αμετάφραστες περιοχές (Untranslated region - UTR), και διαγονιδιακές παραλλαγές.

2.3.8. Αξιολόγηση παραλλαγών σε σχέση με την κλινική τους σημασία

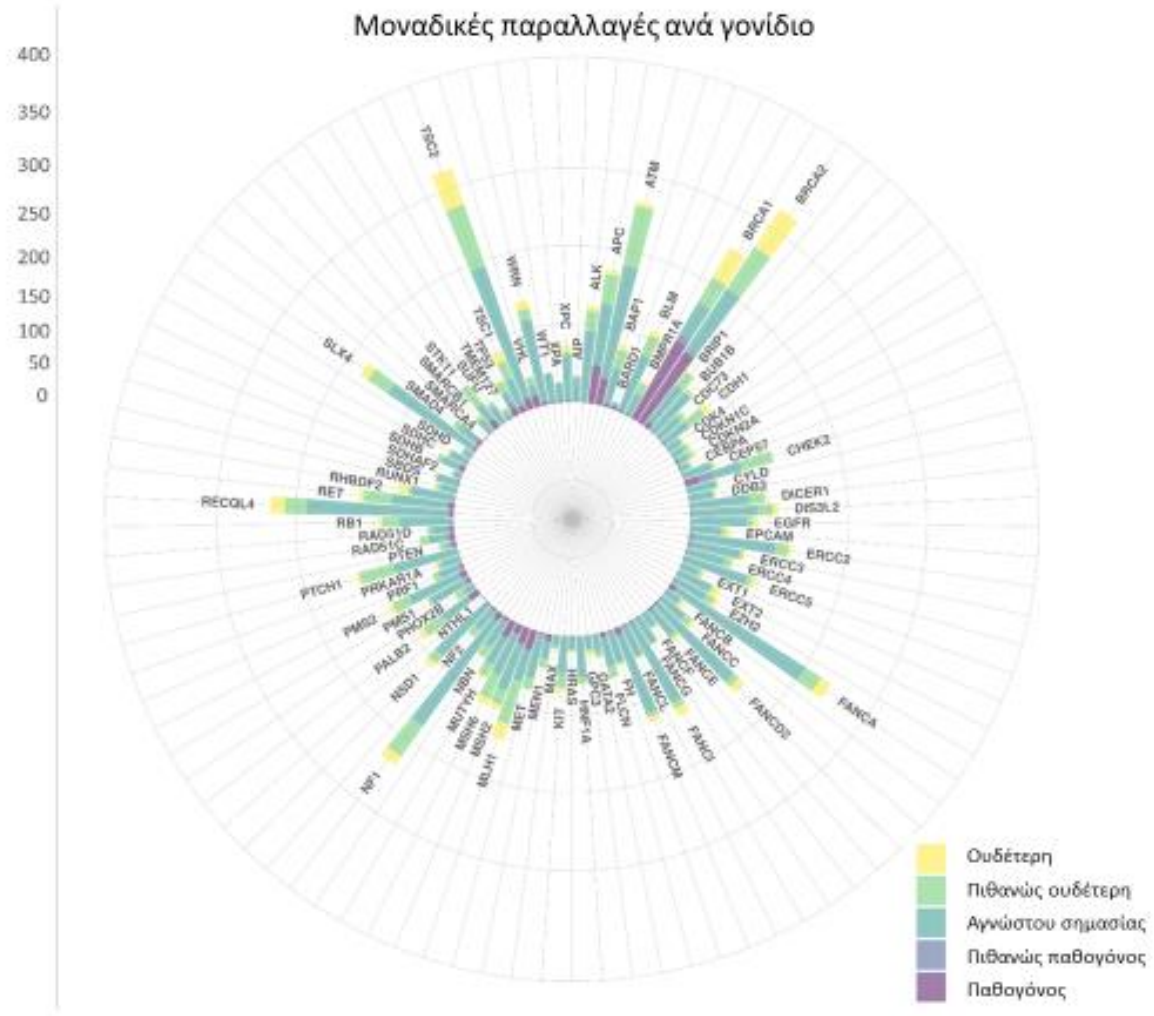
Οι παραλλαγές που ήταν λιγότερο πιθανό να είναι παθογόνοι, δηλαδή οι παραλλαγές σε θέση ματίσματος και οι παραλλαγές σε κωδικές και μη κωδικές περιοχές, αξιολογήθηκαν σε σχέση με τη παθογένειά τους σύμφωνα με τα δεδομένα από τη βάση ClinVar (Landrum *et al.*, 2020). Στην περίπτωση πολλαπλών αντικρουόμενων ερμηνειών μιας παραλλαγής, χρησιμοποιήθηκε η ερμηνεία που προτάθηκε από την πλειοψηφία των συνεισφερόντων.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1. Βάση γενετικών δεδομένων CanVaS

3.1.1. Στατιστικά στοιχεία βάσης δεδομένων

Τα δεδομένα που καταγράφονται στη βάση CanVaS προέρχονται από το γονιδιακό έλεγχο του DNA γαμετικής σειράς 1-97 γονιδίων που προδιαθέτουν σε κάποιον τύπο καρκίνου (Πίνακας 9). Τα δεδομένα αυτά προέρχονται από την ανάλυση του DNA 7.363 ατόμων, εκ των οποίων 6.271 (85,17%) είναι γυναίκες και 1.092 (14,83%) είναι άνδρες.



Εικόνα 31: Κατανομή των καταχωρήσεων των μοναδικών παραλλαγών και της κατηγοριοποίησής τους σε κάθε γονίδιο που καταγράφεται στο CanVaS. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021).

Figure 32: Distribution of unique variant entries and their classification in each gene recorded in CanVaS. Modification from (Kalfakakou, Fostira*, et al.*, 2021).

Συνολικά, εντοπίστηκαν και καταγράφτηκαν 22.089 σπάνιες παραλλαγές σε γονίδια τα οποία είναι γνωστά ή είναι ύποπτα για προδιάθεση σε κάποιον καρκίνο ή καρκινικό σύνδρομο. Αυτές οι σπάνιες παραλλαγές αντιστοιχούν σε 7.968 μοναδικές παραλλαγές. Από αυτές, οι 2.204 αποτελούν παθογόνους/πιθανώς παθογόνους παραλλαγές (Pathogenic Variants/Likely Pathogenic Variants – PV/LPV· 9,98%), οι 10.696 αποτελούν παραλλαγές άγνωστης σημασίας (Variants of Unknown Significance – VUS· 48,42%) και οι 9.189 παραλλαγές αποτελούν ουδέτερες/πιθανώς ουδέτερες παραλλαγές (Benign Variants/Likely Bening Variants – LBV· 41,60%) (Εικόνα 32). Οι υπόλοιπες παραλλαγές βρίσκονται σε διαγονιδιακές περιοχές οι οποίες έχουν ανιχνευθεί σε μελέτες συσχέτισης ολόκληρου το γονιδιώματος και δεν μελετώνται περαιτέρω εδώ. Από όλες τις PV/LPV, 1.817 (82,5%) ανιχνεύονται σε γονίδια υψηλής διεισδυτικότητας (S. Y. Yang *et al.*, 2016; Menko *et al.*, 2014; Salpea & Stratakis, 2014; Puntervoll *et al.*, 2013; Houweling *et al.*, 2011; Janoueix-Lerosey *et al.*, 2008; Mosse *et al.*, 2008), δηλαδή σε γονίδια παθογόνοι παραλλαγές στα οποία επιφέρουν μεγάλο ρίσκο εμφάνισης καρκίνου, 266 (12,0%) σε γονίδια μέτριας διεισδυτικότητας (Daly *et al.*, 2020; Provenzale *et al.*, 2020; Stewart *et al.*, 2019), ενώ 121 (5,5%) βρίσκονται σε γονίδια με χαμηλή ή αβέβαιη διεισδυτικότητα (Daly *et al.*, 2020; Y. Li *et al.*, 2020; Provenzale *et al.*, 2020). Ιδιαίτερα σημαντικό είναι πως οι 1.698 (77%) PV/LPV που καταγράφονται στη βάση δεδομένων CanVaS αφορούν γονίδια που περιλαμβάνονται στον κατάλογο ACMG με τα 59 γονίδια για τα οποία συνιστάται η αναφορά τυχαίων ευρημάτων (Miller *et al.*, 2021) (Πίνακας 9).

Πίνακας 9: Κατανομή παραλλαγών στα γονίδια που περιλαμβάνονται στη βάση δεδομένων CanVaS. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021).

Table 9: Distribution of variants in the genes included in CanVaS. Modification from (Kalfakakou, Fostira*, et al.*, 2021)

Γονίδιο	Ασθένεια/Σύνδρομο	PV	LPV	VUS	LBV	BV
Γονίδια υψηλής διεισδυτικότητας						
ALK	Παιδιατρικό Νευρωβλάστωμα	0	0	193	56	50
APC†	Οικογενής αδενωματώδης πολυποδίαση	86	23	140	135	70
BLM‡	Σύνδρομο Bloom	9	0	109	115	53
BMPR1A†	Νεανική πολυποδίαση	2	0	42	44	13
BRCA1†	Κληρονομικός καρκίνος μαστού/ωοθηκών	761	70	64	64	175
BRCA2†	Κληρονομικός καρκίνος μαστού/ωοθηκών	283	6	121	138	389
CDH1	Κληρονομικός γαστρικός καρκίνος διαχύτου τύπου	1	2	77	156	52
CDK4	Μελάνωμα	0	0	44	7	4
CDKN2A	Μελάνωμα και καρκίνος του παγκρέατος	3	0	20	21	8
EPCAM	Καρκίνος του παχέος εντέρου	0	1	52	21	53
FANCA‡	Αναιμία Fanconi	9	0	391	162	91
FANCB‡	Αναιμία Fanconi	0	0	31	9	42
FANCC‡	Αναιμία Fanconi	0	3	102	22	33
FANCD2‡	Αναιμία Fanconi	3	1	351	15	108

FANCE[‡]	Αναιμία Fanconi	4	1	93	19	43
FANCF[‡]	Αναιμία Fanconi	0	0	50	6	27
FANCG[‡]	Αναιμία Fanconi	1	0	52	1	23
FANCI[‡]	Αναιμία Fanconi	3	0	248	37	114
FANCL[‡]	Αναιμία Fanconi	6	14	77	9	16
FANCM[‡]	Αναιμία Fanconi	16	4	235	9	31
FH	Κληρονομική Λειομυομάτωση και καρκίνος των νεφρών	0	4	60	13	1
FLCN	Σύνδρομο Birt-Hogg-Dubé	18	0	114	55	3
MAX[†]	Παραγαγγλίωμα/Φαιοχρωμοκύττωμα	0	0	17	6	1
MEN1[†]	Πολλαπλή Ενδοκρινής Νεοπλασία Τύπου Ι	14	2	40	17	35
MET	Καρκίνος των νεφρών	2	0	159	56	1
MLH1[†]	Σύνδρομο Lynch	41	9	87	345	405
MSH2[†]	Σύνδρομο Lynch	43	0	133	94	86
MSH6[†]	Σύνδρομο Lynch	16	0	106	141	113
MUTYH[†‡]	MUTYH σχετιζόμενη αδενωματώδης πολυποδίαση	99	5	110	43	17
NF1	Νευροϊνωμάτωση 1	10	0	353	127	46
NF2[†]	Νευροϊνωμάτωση 2	5	0	39	43	0
NTHL1[‡]	NTHL1 σχετιζόμενο καρκινικό σύνδρομο	4	0	0	0	0
PALB2[†]	Καρκίνος μαστού και παγκρέατος	54	0	67	118	47
PTCH1	Σύνδρομο Gorlin	0	1	126	183	52
PTEN[†]	Σύνδρομο Cowden	7	0	11	78	7
PRKAR1A	Σύμπλεγμα Carney	0	0	84	10	3
RB1[†]	Ρετινοβλάστωμα	1	0	285	86	40
RECQL4[‡]	Σύνδρομο Rothmund-Thomson	6	2	529	76	109
RET[†]	Πολλαπλή Ενδοκρινής Νεοπλασία Τύπου ΙΙ	19	2	160	71	10
SDHAF2[†]	Παραγαγγλίωμα	0	0	12	16	0
SDHB[†]	Παραγαγγλίωμα	17	3	64	83	1
SDHC[†]	Παραγαγγλίωμα	0	0	90	4	2
SDHD[†]	Παραγαγγλίωμα	1	0	48	0	0
SMAD4[†]	Νεανική πολυποδίαση	7	1	50	26	35
SMARCA4	Μικροκυτταρικό καρκίνωμα ωοθηκών, υπερασβεσταιμικού τύπου	4	0	0	0	0
SMARCB1	Σύνδρομο προδιάθεσης ραβδοειδών όγκων	0	0	62	16	9
STK11[†]	Σύνδρομο Peutz-Jeghers	17	0	40	54	0
SUFU	Σύνδρομο Gorlin	0	0	30	109	0
TP53[†]	Σύνδρομο Li-Fraumeni Syndrome	16	25	22	43	26
TSC1[†]	Σύμπλεγμα Οζώδους Σκλήρυνσης	8	0	80	65	64
TSC2[†]	Σύμπλεγμα Οζώδους Σκλήρυνσης	14	0	377	226	285

VHL[†]	Σύνδρομο Von Hippel Lindau	27	0	16	33	0
WT1[†]	Όγκος Wilm's 1	1	0	114	0	0
Γονίδια μέτριας διεισδυτικότητας						
AIP	Αδενώματα υπόφυσης	0	0	40	66	0
ATM	Καρκίνος μαστού	57	5	281	367	66
BRIP1	Καρκίνος ωοθηκών	14	3	117	70	5
CHEK2	Καρκίνος μαστού	36	97	162	135	0
DICER1	DICER1 σχετιζόμενο σύνδρομο	0	1	111	78	6
PMS2[†]	Σύνδρομο Lynch	11	1	246	151	13
RAD51C	Καρκίνος ωοθηκών	27	8	49	69	1
RAD51D	Καρκίνος ωοθηκών	6	0	32	10	0
Γονίδια χαμηλής/αβέβαιης διεισδυτικότητας						
BAP1	Σύνδρομο προδιάθεσης σε όγκους	3	2	84	98	56
BARD1	Καρκίνος μαστού	6	0	16	0	0
BUB1B[‡]	Καρκίνος παχέος εντέρου	0	0	119	9	71
CDC73	Αδενώματα παραθυρεοειδούς & καρκίνος	0	0	66	12	2
CDKN1C	Σύνδρομο Beckwith-Wiedemann	0	0	27	17	4
CEBPA	Οξεία Μυελογενής Λευχαιμία	5	0	21	7	8
CEP57	Άγνωστο	0	0	87	10	25
CYLD	Οικογενής κυλινδρομάτωση	0	0	94	2	2
DDB2[‡]	Μελαγχρωματική ξηροδερμία, DDB-αρνητικός υπότυπος	0	0	47	15	23
DIS3L2[‡]	Σύνδρομο Perlman	0	1	278	54	49
EGFR	Καρκίνος πνεύμονα	0	0	243	37	17
ERCC2[‡]	Μελαγχρωματική ξηροδερμία	1	9	175	32	4
ERCC3[‡]	Μελαγχρωματική ξηροδερμία	0	9	157	12	36
ERCC4[‡]	Μελαγχρωματική ξηροδερμία	0	2	87	51	44
ERCC5[‡]	Μελαγχρωματική ξηροδερμία	0	4	221	17	15
EXT1	Άγνωστο	0	0	38	16	3
EXT2	Άγνωστο	0	2	67	10	38
EZH2	Σύνδρομο Weaver	0	3	75	21	20
GATA2	Σύνδρομο Emberger	0	0	28	7	6
GPC3	Άγνωστο	0	0	43	7	31
HNF1A	Άγνωστο	0	0	124	38	18
HRAS	Άγνωστο	0	8	37	2	5
KIT	Γαστρεντερικοί στρωματικοί όγκοι	0	0	157	39	68

NBN	Καρκίνος μαστού	12	3	78	42	4
NSD1	Σύνδρομο Sotos	1	1	122	50	45
PHOX2B	Συγγενές Σύνδρομο Κεντρικού Υποαερισμού	0	0	16	33	4
PMS1	Καρκίνος παχέος εντέρου	8	0	105	5	8
PRF1	Άγνωστο	5	1	80	82	8
RHBDF2	Τύλωση με καρκίνο του οισοφάγου	0	0	133	17	73
RUNX1	Οξεία Μυελογενής Λευχαιμία	0	0	23	16	40
SBDS‡	Σύνδρομο Shwachman-Diamond	7	10	119	9	2
SLX4‡	Άγνωστο	6	0	256	76	220
TMEM127†	Παραγαγγλίωμα/Φαιοχρωμοκύττωμα	1	0	9	10	7
WRN‡	Σύνδρομο Werner	4	3	251	57	107
XPA‡	Μελαγχρωματική ξηροδερμία	2	1	31	3	0
XPC‡	Μελαγχρωματική ξηροδερμία	1	0	167	5	95

PV: Παθογόνος παραλλαγή· LPV: Πιθανώς παθογόνος παραλλαγή· VUS: Παραλλαγή αγνώστου σημασίας· LBV: Πιθανώς ουδέτερη παραλλαγή· BV: Ουδέτερη παραλλαγή.

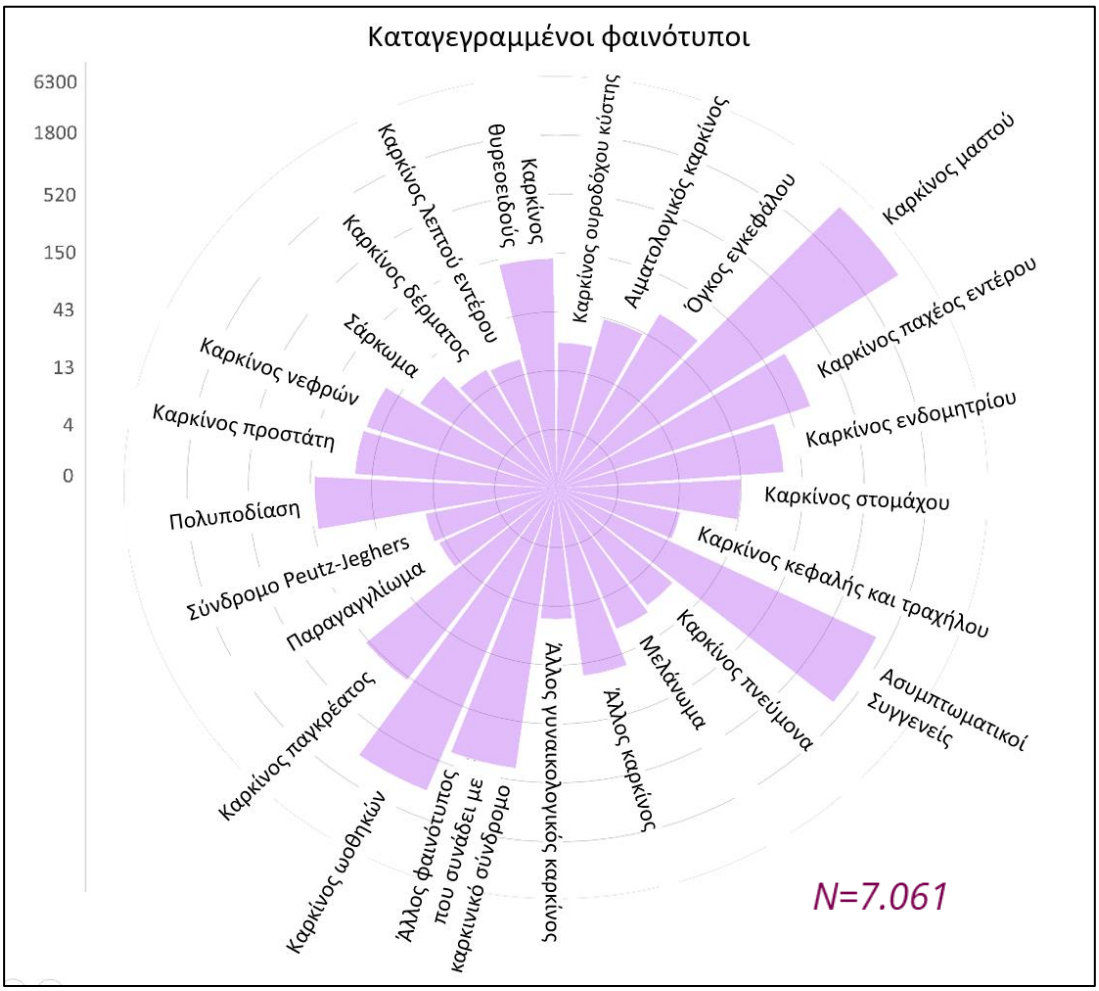
† Γονίδια που περιλαμβάνονται στον κατάλογο ACMG με τα 59 γονίδια για τα οποία συνιστάται η αναφορά τυχαίων ευρημάτων.

‡ Αυτοσωμικός υπολειπόμενος τρόπος κληρονόμησης.

Συνολικά, στη βάση δεδομένων CanVaS καταγράφονται 7.061 διαφορετικοί φαινότυποι ασθενειών σε 6.001 (81,5%) άτομα, ενώ 1.362 άτομα (18,5%) είναι ασυμπτωματικοί συγγενείς εξ αίματος των προαναφερθέντων ασθενών. Ο πιο διαδεδομένος φαινότυπος είναι ο καρκίνος του μαστού, που αντιστοιχεί σε 3.970 άτομα (53,92% · μέση ηλικία ± SD: 46,82 ± 12.05), εκ των οποίων οι 103 (1,04% όλων των ατόμων) είναι ασθενείς με ανδρικό καρκίνο μαστού (μέση ηλικία ± SD: 61,95 ± 13.4). Επιπλέον, 1.017 (13,81%) ασθενείς έχουν διαγνωστεί με καρκίνο των ωοθηκών (μέση ηλικία ± SD: 53,08 ± 13,04), ενώ οι 510 (6,93%) έχουν ιστορικό πολυπόδων, πολυποδίασης ή/και καρκίνου του παχέος εντέρου (μέση ηλικία ± SD του καρκίνου παχέος εντέρου: 50,6 ± 13,85). Επιπλέον, 158 (2,15%), 131 (1,78%) και 102 (1,39%) άτομα έχουν διαγνωστεί με καρκίνο του παγκρέατος (μέση ηλικία ± SD: 60,14 ± 12,19), του θυρεοειδούς (μέση ηλικία ± SD: 46,42 ± 15,01) και του ενδομητρίου (μέση ηλικία ± SD: 52,08 ± 12,75), αντίστοιχα. Οι υπόλοιπες διαγνώσεις που καταγράφονται στη βάση δεδομένων CanVaS περιλαμβάνουν πιο σπάνιους τύπους καρκίνου, όπως καρκίνο του εγκεφάλου ή γαστρικό καρκίνο διαχύτου τύπου (συνολικά 857 ή 11,64% των ασθενών έχουν διαγνωστεί με άλλους καρκινικούς φαινότυπους) και μη καρκινικές διαγνώσεις που όμως συνάδουν με καρκινικά σύνδρομα, όπως η οζώδης σκλήρυνση και η νευροϊνωμάτωση (317 ή 4,3% των ασθενών με μη καρκινικούς φαινοτύπους) (Εικόνα 33). Συνολικά 709 (9,63%) άτομα στο CanVaS παρουσιάζουν περισσότερους από έναν φαινότυπους.

Μεταξύ των ατόμων που είχαν τουλάχιστον μία διάγνωση καρκίνου, 91 (1,52%) ήταν παιδιά ή έφηβοι (0-18 ετών) κατά τη στιγμή της διάγνωσης, 360 (6%) ήταν νεαροί ενήλικες (19-30 ετών), 2.158 (35,97%) διαγνώστηκαν μεταξύ των ηλικιών 31-45, 1.846 (30,77%) διαγνώστηκαν μεταξύ των ηλικιών 46-60, ενώ

1.018 (16,97%) άτομα διαγνώστηκαν μετά την ηλικία των 60. Η ηλικία στην πρώτη διάγνωση δεν ήταν διαθέσιμη για 527 (8,78%) άτομα.



Εικόνα 32: Καταχωρήσεις φαινοτύπων στο CanVaS. Οι τιμές απεικονίζονται σε κλίμακα $\log_2$. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021)

Figure 33: Phenotype entries in CanVaS. Values are displayed on a $\log_2$ scale. Modification from (Kalfakakou, Fostira*, et al.*, 2021)

3.1.2. Παρεχόμενες πληροφορίες

Κάθε άτομο που είναι εξοικειωμένο με το λογισμικό ανοιχτού κώδικα LOVD μπορεί εύκολα να πλοηγηθεί στη βάση δεδομένων CanVaS, καθώς η μορφή της βάσης δεδομένων είναι τυποποιημένη, ενώ το LOVD χρησιμοποιείται συχνά για το στήσιμο βάσεων γενετικών δεδομένων. Ωστόσο, εκτός από τις τυποποιημένες πληροφορίες που περιλαμβάνονται στις βάσεις δεδομένων που έχουν αναπτυχθεί με τη χρήση του λογισμικού LOVD, οι πίνακες της βάσης δεδομένων CanVaS έχουν εμπλουτιστεί με επιπλέον, ειδικά διαμορφωμένες στήλες (Πίνακας 10), έτσι ώστε να μπορούν να παρέχονται συμπληρωματικές πληροφορίες, που εξυπηρετούν τον σκοπό μιας εθνικής πηγής γενετικών δεδομένων και της καταγραφής των φαινοτύπων με όσο γίνεται περισσότερη λεπτομέρεια.

Πίνακας 10: Περιγραφή ειδικά διαμορφωμένων πεδίων που περιλαμβάνονται στη βάση δεδομένων CanVaS. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021).

Table 10: Description of custom fields included in CanVaS. Modification from (Kalfakakou, Fostira*, et al.*, 2021).

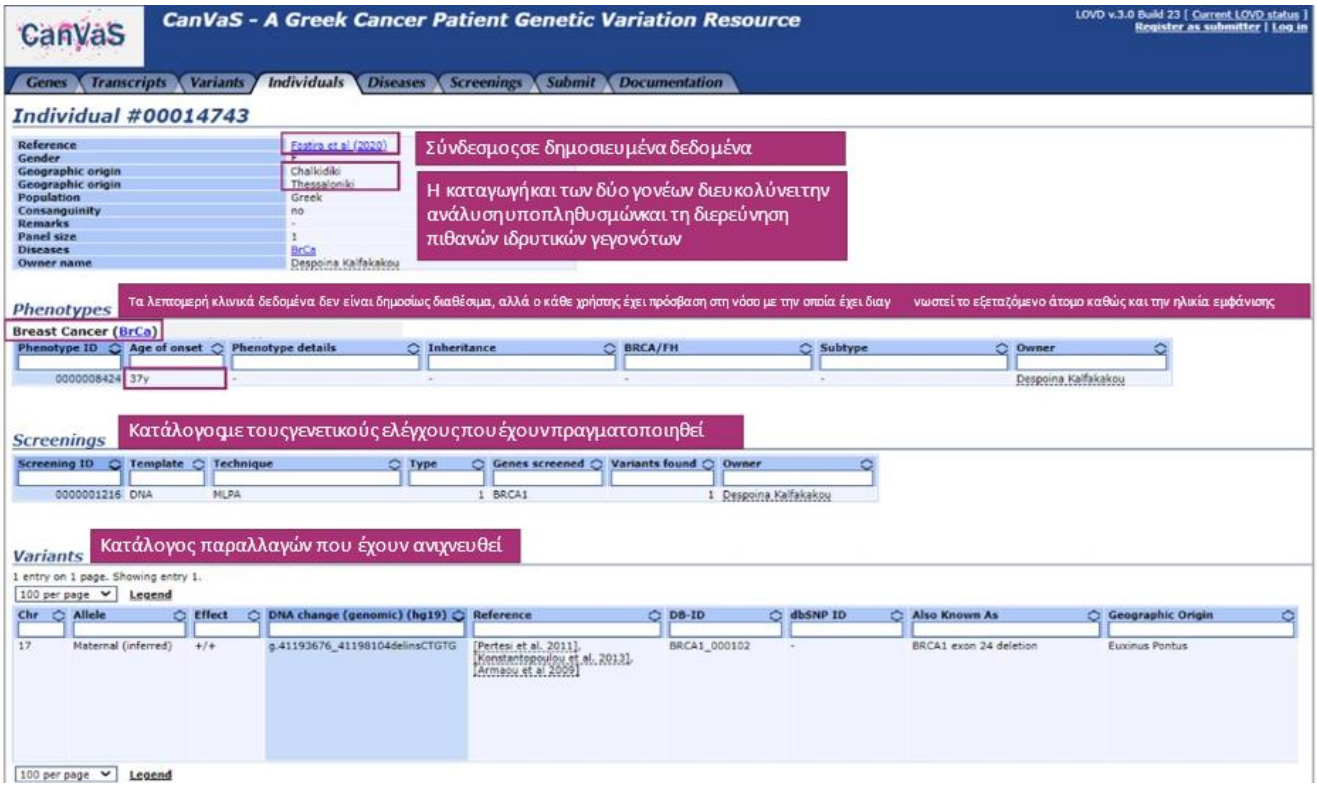
Πεδίο	**Περιγραφή**	**Πρόσβαση**
Individual/Consanguinity†	Υποδεικνύει αν οι γονείς του εξεταζόμενου ατόμου έχουν συγγένεια εξ αίματος	Δημόσια
Individual/Gender†	Το φύλο του συγκεκριμένου ατόμου	Δημόσια
Individual/Origin/Geographic†	Η γεωγραφική καταγωγή του Γονέα #1 του ατόμου	Δημόσια
Individual/Origin/Geographic2‡	Η γεωγραφική καταγωγή του Γονέα #2 του ατόμου	Δημόσια
Individual/Origin/Population†	Ο πληθυσμός στον οποίο ανήκει το άτομο	Δημόσια
Individual/Remarks†	Παρατηρήσεις σχετικές με το άτομο	Δημόσια
Phenotype/Additional†	Συμπληρωματικές πληροφορίες σχετικά με τον φαινότυπο	Δημόσια
Phenotype/Age/Onset†	Η ηλικία κατά την οποία εμφανίστηκαν τα πρώτα συμπτώματα, αν είναι γνωστή	Δημόσια
Phenotype/Behaviour‡	Η συμπεριφορά του όγκου (διαχύτου τύπου/οριοθετημένο)	Ελεγχόμενη
Phenotype/BrCa/FH‡	Οικογενειακό ιστορικό καρκίνου μαστού, ωοθηκών ή/και παγκρέατος. Όχι = κανένας καρκίνος στην οικογένεια, Αδύναμο = ένας καρκίνος στην οικογένεια, Σοβαρό = περισσότεροι από έναν καρκίνοι στην οικογένεια. Μόνο για φαινότυπους καρκίνου μαστού.	Ελεγχόμενη
Phenotype/CRC/CRC_FH‡	Οικογενειακό ιστορικό καρκίνου παχέος εντέρου. Μόνο για φαινότυπους καρκίνου παχέος εντέρου.	Ελεγχόμενη
Phenotype/CRC/Lynch_FH‡	Οικογενειακό ιστορικό καρκίνων που συνδέονται με το σύνδρομο Lynch. Μόνο για φαινότυπους καρκίνου παχέος εντέρου.	Ελεγχόμενη
Phenotype/CRC/Polyposis_FH‡	Οικογενειακό ιστορικό πολυποδιάσεων. Μόνο για φαινότυπους καρκίνου παχέος εντέρου.	Ελεγχόμενη
Phenotype/ER‡	Κατάσταση υποδοχέων οιστρογόνων.	Ελεγχόμενη
Phenotype/PRC/Gleason‡	Δείκτης Gleason. Μόνο για φαινότυπους καρκίνου του προστάτη.	Ελεγχόμενη
Phenotype/Grade‡	Βαθμός όγκου	Ελεγχόμενη
Phenotype/HER2‡	Κατάσταση πρωτεΐνης HER2.	Ελεγχόμενη
Phenotype/LN‡	Θετικότητα λεμφαδένων	Ελεγχόμενη

Phenotype/Morphology‡	Ιστολογία όγκου (πχ πορογενές, λοβιακό κλπ)	Ελεγχόμενη
Phenotype/MSI‡	Μικροδορυφορική αστάθεια.	Ελεγχόμενη
Phenotype/OvCa/FH‡	Οικογενειακό ιστορικό καρκίνου μαστού, ωοθηκών ή/και παγκρέατος. Μόνο για φαινότυπους καρκίνου ωοθηκών.	Ελεγχόμενη
Phenotype/PR‡	Κατάσταση υποδοχέων προγεστερόνης.	Ελεγχόμενη
Phenotype/Stage‡	Στάδιο όγκου.	Ελεγχόμενη
Phenotype/Tissue‡	Ιστός στον οποίο αναπτύχθηκε ο όγκος.	Ελεγχόμενη
Phenotype/Subtype‡	Υπότυπος όγκου.	Ελεγχόμενη
Screening/Technique†	Τεχνική που χρησιμοποιήθηκε για την ανίχνευση της παραλλαγής.	Δημόσια
VariantOnGenome/AKA‡	Ονοματολογία παραλλαγής με την οποία είναι γνωστή και η οποία δεν ακολουθεί τους κανόνες HGVS.	Δημόσια
VariantOnGenome/Allele_Count‡	Πλήθος κάθε εναλλακτικού αλληλομόρφου σε κάθε θέση στο γονιδίωμα σε όλα τα άτομα χωρίς συγγενική σχέση.	Δημόσια
VariantOnGenome/Allele_Number‡	Συνολικός αριθμός αλληλομόρφων σε κάθε θέση στο γονιδίωμα σε όλα τα άτομα χωρίς συγγενική σχέση.	Δημόσια
VariantOnGenome/dbSNP†	Ο κωδικός αναφοράς της παραλλαγής στη βάση dbSNP.	Δημόσια
VariantOnGenome/DNA†	Περιγραφή της παραλλαγής σε επίπεδο DNA, βασισμένη στην ακολουθία αναφοράς του γενωμικού DNA (ακολουθώντας τους κανόνες HGVS).	Δημόσια
VariantOnGenome/GeographicOrigin‡	Εντοπιότητα παραλλαγής.	Δημόσια
VariantOnGenome/IsGreekFounder‡	Υποδεικνύει αν η παραλλαγή είναι ιδρυτική για τον ελληνικό πληθυσμό.	Δημόσια
VariantOnGenome/Reference†	Αναφορά στη δημοσίευση στην οποία περιγράφεται η παραλλαγή.	Δημόσια
VariantOnGenome/Remarks†	Παρατηρήσεις σχετικές με την παραλλαγή.	Δημόσια
VariantOnGenome/Segregation†	Υποδεικνύει εάν η παραλλαγή διαχωρίζεται με τον φαινότυπο	Δημόσια
VariantOnTranscript/DNA†	Περιγραφή της παραλλαγής σε επίπεδο DNA, βασισμένη στην ακολουθία αναφοράς του κωδικού DNA (ακολουθώντας τους κανόνες HGVS).	Δημόσια
VariantOnTranscript/Exon†	Ο αριθμός του εξονίου/ιντρονίου όπου εδράζεται η παραλλαγή.	Δημόσια
VariantOnTranscript/Protein†	Περιγραφή της παραλλαγής σε επίπεδο πρωτεΐνης (ακολουθώντας τους κανόνες HGVS).	Δημόσια
VariantOnTranscript/RNA†	Περιγραφή της παραλλαγής σε επίπεδο RNA (ακολουθώντας τους κανόνες HGVS).	Δημόσια

† Ειδικά διαμορφωμένα πεδία που παρέχονται από την LOVD.

‡Ειδικά διαμορφωμένα πεδία που υπάρχουν μόνο στη βάση CanVaS.

Κάθε καταχώριση που σχετίζεται με ένα άτομο, συνοδεύεται από δημογραφικά στοιχεία, λόγου χάρη το φύλο και την ηλικία. Αξίζει να σημειωθεί ότι επιπλέον παρέχονται πληροφορίες σχετικά με τη συγκεκριμένη γεωγραφική περιοχή από την οποία κατάγεται το εξεταζόμενο άτομο, επιτρέποντας την καταγραφή της εντοπιότητας της παραλλαγής, καθώς και των απομονωμένων πληθυσμών. Εκτός από τα δημογραφικά στοιχεία, κάθε καταχώριση που σχετίζεται με ένα άτομο συμπληρώνεται και με έναν σύνδεσμο προς το δημοσιευμένο άρθρο στο οποίο έχουν συμπεριληφθεί τα αντίστοιχα δεδομένα, όπου είναι διαθέσιμο. Επιπλέον, παρέχεται μία σύνοψη των φαινοτύπων τους οποίους εμφανίζει το άτομο, του γονιδιακού ελέγχου και των παραλλαγών που φέρει το άτομο (Εικόνα 34).

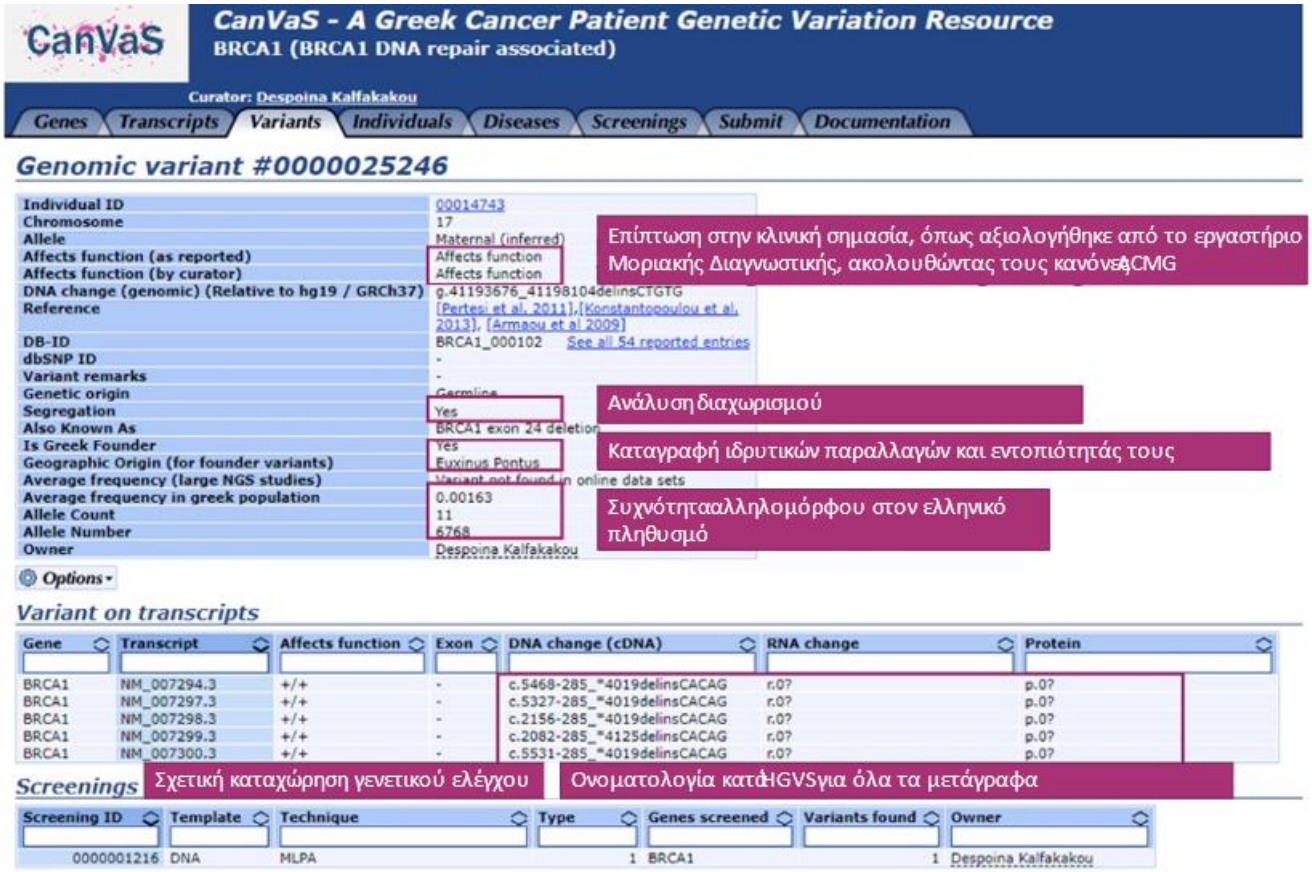


Εικόνα 33: Στιγμιότυπο της διεπαφής του CanVaS το οποίο απεικονίζει μία σελίδα καταχώρησης ενός ατόμου. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021).

Figure 34: CanVaS interface snapshot of an “Individual” entry. Modification from (Kalfakakou, Fostira*, et al.*, 2021).

Κάθε καταχώριση για μια παραλλαγή που έχει ανιχνευθεί σε κάποιο εξεταζόμενο άτομο, περιλαμβάνει πληροφορίες σχετικά με τη θέση της στο γονιδίωμα, την περιγραφή της σε όλα τα μετάγραφα σύμφωνα με τους κανόνες ονοματολογίας παραλλαγών HGVS (den Dunnen *et al.*, 2016), την ανάλυση διαχωρισμού στην οικογένεια και την προέλευση της παραλλαγής (πατρική ή μητρική). Επιπλέον, κάθε μοναδική παραλλαγή αξιολογήθηκε ξεχωριστά και κατηγοριοποιήθηκε ανάλογα σε σχέση με την κλινική της σημασία, σύμφωνα με τις οδηγίες ACMG. Όλες οι προαναφερθείσες πληροφορίες συνοδεύονται από προσαρμοσμένα δεδομένα, τα οποία εξυπηρετούν το σκοπό μιας βάση

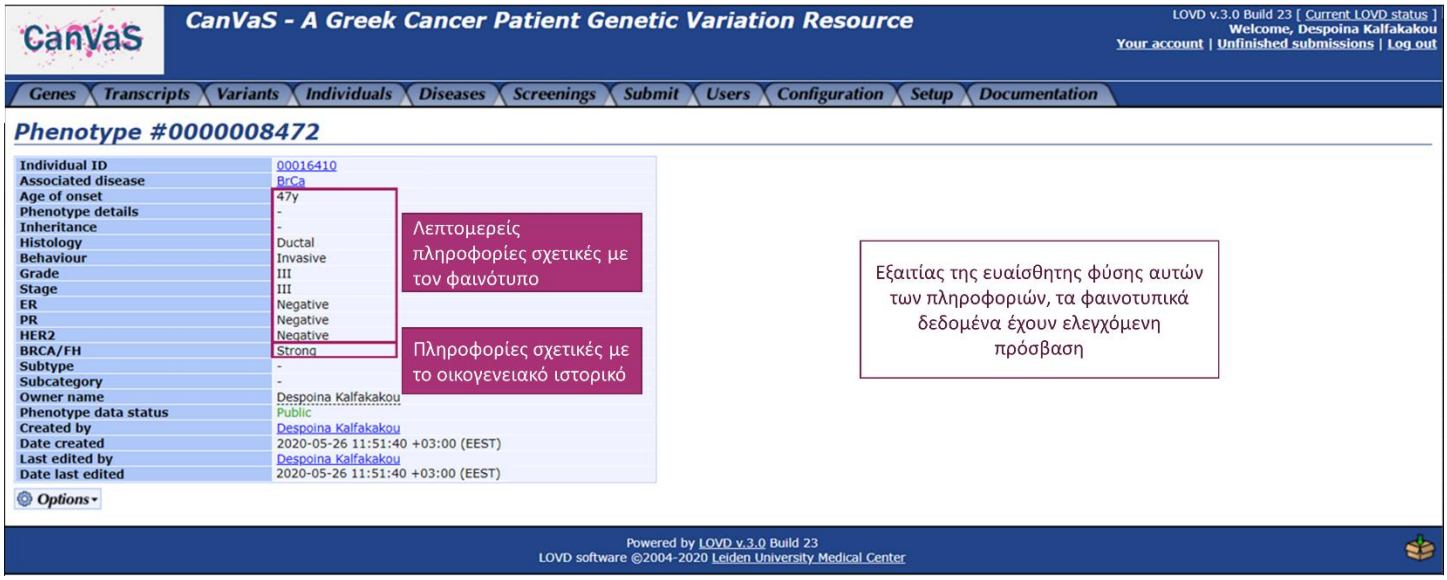
πληθυσμιακών δεδομένων. Έτσι, κάθε καταχώριση για μια παραλλαγή, συνοδεύεται επίσης από την εντοπιότητά της, καθώς και πληροφορίες σχετικά με το αν είναι ιδρυτική ή όχι, όπου είναι διαθέσιμες. Επιπλέον, καταγράφεται η συχνότητα εμφάνισής της στον ελληνικό πληθυσμό. Εκτός από την ονοματολογία κατά HGVS, καταγράφονται επιπλέον «δημοφιλή» ονόματα της παραλλαγής για ευκολότερη αναγνώριση, ένα χαρακτηριστικό το οποίο φαίνεται ιδιαίτερα χρήσιμο για την καταγραφή μεγάλων γονιδιωματικών αναδιατάξεων, οι οποίες συχνά περιγράφονται μόνο σε γονιδιωματικό επίπεδο. Ωστόσο, οι πληροφορίες σχετικά με την απουσία μιας παραλλαγής είναι εξίσου ισχυρές όσο και οι πληροφορίες σχετικά με την παρουσία της σε γενετικές μελέτες. Επομένως, η καταχώριση του γονιδιακού ελέγχου ενός ατόμου καταγράφει όχι μόνο τη μέθοδο με την οποία το DNA του εξεταζόμενου αναλύθηκε, αλλά και το αν εντοπίστηκε κάποια παραλλαγή ή όχι (Εικόνα 35).



Εικόνα 34: Στιγμιότυπο της διεπαφής του CanVaS το οποίο απεικονίζει μία σελίδα καταχώρησης μιας παραλλαγής. Η παραλλαγή κατηγοριοποιείται ως “Επηρεάζει τη λειτουργία (Affects function)”, που αντιστοιχεί σε κατηγοριοποίηση ως παθογόνος σύμφωνα με τις οδηγίες ACMG. Τροποποίηση από (Kalfakakou, Fostira, *et al.*, 2021).

Figure 35: CanVaS interface snapshot of a “Variant” entry. The variant is classified as "Affects function", which corresponds to a pathogenic classification to the ACMG guidelines. Modification from (Kalfakakou, Fostira, *et al.*, 2021).

Ιδιαίτερη προσοχή δόθηκε στην καταγραφή των φαινοτυπικών δεδομένων, καθώς μπορούν να χρησιμεύσουν ως ένα ισχυρό εργαλείο για κάθε ανάλυση. Για αυτόν τον λόγο, δόθηκε έμφαση στην περιγραφή των δεδομένων με δομημένο και λεπτομερή τρόπο, ελαχιστοποιώντας τις καταχωρήσεις ελεύθερου κειμένου. Επομένως, κάθε καταχώριση που σχετίζεται με κάποιο φαινότυπο συνδυάζεται με σημαντικά κλινικά δεδομένα, όπως η μορφολογία του όγκου του ασθενούς, ο βαθμός, το στάδιο και η κατάσταση των ορμονικών υποδοχέων του όγκου και η κατάσταση των λεμφαδένων. Παράλληλα, παρέχεται το ιατρικό ιστορικό οποιουδήποτε σχετικού φαινοτύπου της οικογένειας του ατόμου (Εικόνα 36). Εξαιτίας της ευαίσθητης φύσης αυτών των πληροφοριών, τα φαινοτυπικά δεδομένα έχουν ελεγχόμενη πρόσβαση και είναι διαθέσιμα μόνο κατόπιν αιτήματος μέσω της διεπαφής της βάσης δεδομένων.

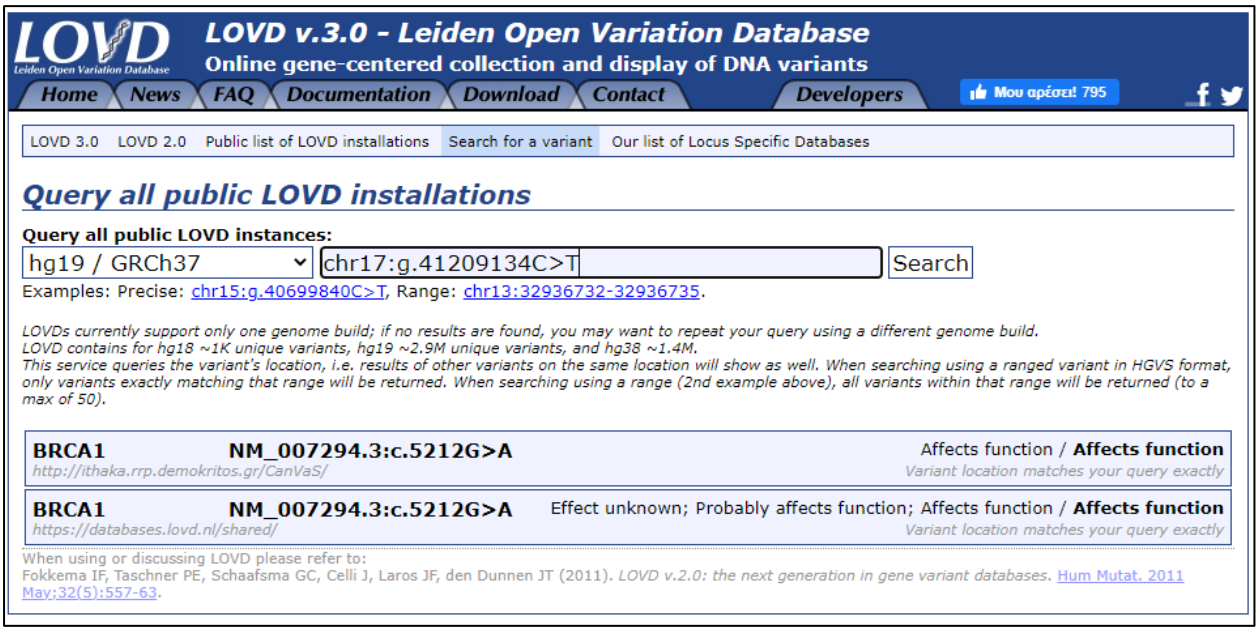


Εικόνα 35: Στιγμιότυπο της διεπαφής του CanVaS το οποίο απεικονίζει μία σελίδα καταχώρησης ενός φαινότυπου. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021)

Figure 36: CanVaS interface snapshot of a “Phenotype” entry. Modification from (Kalfakakou, Fostira*, et al.*, 2021).

3.1.3. Ενσωμάτωση στην κεντρική εγκατάσταση της LOVD

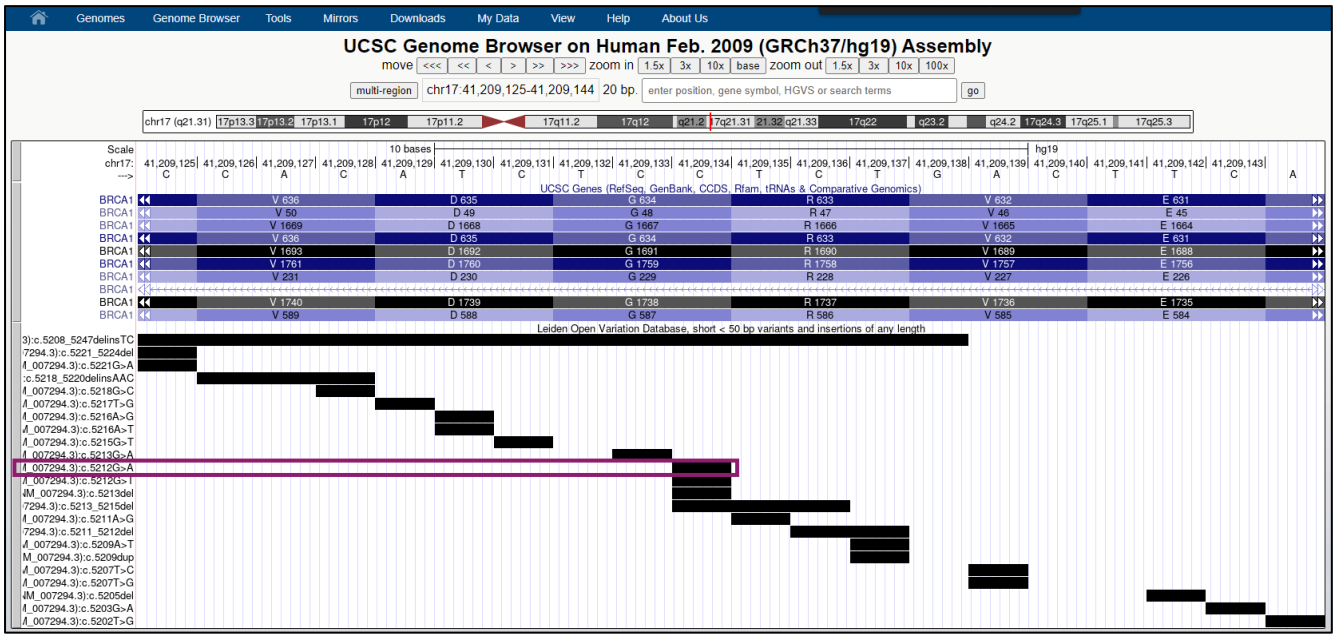
Όλες οι εγκαταστάσεις του λογισμικού ανοικτού κώδικα LOVD για τις οποίες είναι ενεργοποιημένη η επιλογή να είναι δημόσιες, συνδέονται εξ ορισμού με την κεντρική εγκατάσταση της LOVD. Έτσι, και οι παραλλαγές που καταγράφονται στη βάση δεδομένων CanVaS, περιλαμβάνονται αυτόματα στη μηχανή αναζήτησης της κεντρικής εγκατάστασης LOVD. Επομένως, κάθε φορά που ένας χρήστης υποβάλει ένα ερώτημα στην κεντρική εγκατάσταση της LOVD για μια συγκεκριμένη παραλλαγή που καταγράφεται στη CanVaS, του επιστρέφεται η καταχώρηση της παραλλαγής στη βάση CanVaS. Για παράδειγμα, όταν ένας χρήστης αναζητήσει την παραλλαγή NM_007294.3:c.5212G>A (p.Gly1738Arg) του γονιδίου *BRCA1,* η οποία είναι μία ιδρυτική παθογόνος παραλλαγή για τον ελληνικό πληθυσμό, του επιστρέφεται η καταχώρηση της παραλλαγής στη CanVaS και στην κεντρική εγκατάσταση της LOVD (Εικόνα 37).



Εικόνα 36: Παράδειγμα αναζήτησης της παραλλαγής NM_007294.3:c.5212G>A (p.Gly1738Arg) του γονιδίου *BRCA1* στην κεντρική εγκατάσταση της βάσης LOVD. Η αναζήτηση επιστρέφει την καταχώρηση της παραλλαγής στο CanVaS και στην κεντρική LOVD.

Figure 37: Example query for the *BRCA1* gene variant NM_007294.3: c.5212G>A (p.Gly1738Arg) in the central installation of the LOVD database. The search returns the variant entries both in CanVaS and the central LOVD.

Παρομοίως, μιας και η LOVD είναι συνδεδεμένη με το πρόγραμμα περιήγησης γονιδιωμάτων του Πανεπιστημίου της Καλιφόρνια – Σάντα Κρουζ (UCSC Genome browser) (Kent *et al.*, 2002) κάθε χρήστης που χρησιμοποιεί το συγκεκριμένο πρόγραμμα περιήγησης γονιδιωμάτων, μπορεί να μεταφερθεί άμεσα στη βάση δεδομένων CanVaS, σε περίπτωση που θέλει να μελετήσει κάποια παραλλαγή που είναι καταχωρημένη σ’ αυτή (Εικόνα 38).



Εικόνα 37: Παράδειγμα περιήγησης στη θέση της παραλλαγής NM_007294.3:c.5212G>A (p.Gly1738Arg) του γονιδίου *BRCA1* με τη χρήση του γονιδιωματικού περιηγητή του UCSC. Η παραλλαγή εμφανίζεται, καθώς είναι καταχωρημένη στο CanVaS και, συνεπώς, στην κεντρική εγκατάσταση της LOVD.

Figure 38: Example of navigation, using the UCSC genomic browser, in the position of the *BRCA1* variant NM_007294.3: c.5212G>A (p.Gly1738Arg). The variant appears as it is registered in CanVaS and therefore in the central LOVD installation.

3.2. Ροή διοχέτευσης εντολών διεργασιών VarTrace

3.2.1. Αρχεία εξόδου

Το εργαλείο VarTrace αναπτύχθηκε με σκοπό την ακριβή ανίχνευση σωματικών παραλλαγών στο DNA καρκινικών κυττάρων. Τα κύρια αρχεία εξόδου του VarTrace είναι το αρχείο BAM που παράγεται μετά την αντιστοίχιση στο γονιδίωμα αναφοράς, την εκ νέου αντιστοίχιση των αναγνώσεων και την εκ νέου βαθμονόμηση των βάσεων, και το αρχείο VCF που περιέχει τις παραλλαγές που έχουν ανιχνευτεί από όλα τα εργαλεία κλήσης παραλλαγών, μετά το φιλτράρισμα των παραλλαγών που δεν πληρούν τα ποιοτικά κριτήρια. Ο χρήστης έχει τη δυνατότητα να επιλέξει αν θέλει να αποθηκευτούν και τα ενδιάμεσα αρχεία VCF των επιμέρους προγραμμάτων.

Εκτός από τα αρχεία BAM και VCF, το VarTrace έχει ως έξοδο ένα εμπλουτισμένο με πληροφορίες αρχείο παραλλαγών και δύο αρχεία αναφοράς για τον έλεγχο ποιότητας του πειράματος Αλληλούχησης Επόμενης Γενιάς. Το εμπλουτισμένο με πληροφορίες αρχείο παραλλαγών έχει δημιουργηθεί με το πρόγραμμα εμπλουτισμού VEP της Ensembl κι έχει επεξεργαστεί περαιτέρω για την εξαγωγή πληροφοριών που βρίσκονται στο VCF, αλλά δεν παρέχονται από το VEP, όπως το βάθος ανάγνωσης στη συγκεκριμένη θέση και η συχνότητα αλληλομόρφου παραλλαγής (Πίνακας 11). Επιπλέον, παρέχεται η αναφορά ποιότητας των αναγνώσεων DNA, όπως εξέρχεται από το εργαλείο FastQC. Τέλος, το VarTrace έχει ως έξοδο ένα αρχείο με όλες τις περιοχές που έχουν μικρό βάθος ανάγνωσης ή/και χαμηλή ποιότητα

αντιστοίχισης αναγνώσεων. Το αρχείο αυτό είναι μορφής BED και είναι συμβατό με τους περιηγητές γονιδιωμάτων (Εικόνα 39).

Πίνακας 11: Πεδία που περιλαμβάνει το εμπλουτισμένο με πληροφορίες αρχείο παραλλαγών και η επεξήγησή τους.

Table 11: Fields included in the annotated variant file and their explanation.

Πεδίο	**Επεξήγηση**	**Πεδίο**	**Επεξήγηση**
CHROM	Χρωμόσωμα	CDS_position	Θέση στην κωδική αλληλουχία
POS	Θέση	Protein_position	Θέση στην πρωτεΐνη
ID	Γνωστός μοναδικός κωδικός	Amino_acids	Αμινοξέα
REF	Αλληλόμορφο αναφοράς	Codons	Κωδικόνια
ALT	Εναλλακτικό αλληλόμορφο	Existing_variation	Κωδικοί παραλλαγής σε διάφορες βάσεις δεδομένων
QUAL	Ποιότητα	STRAND	Κλώνος DNA
FILTER	Φίλτρα	ENSP	Κωδικός πρωτεΐνης
GT†	Γονότυπος	SIFT	Ταξινόμηση βάσει του εργαλείου SIFT
AD†	Βάθος ανάγνωσης αλληλομόρφων	PolyPhen	Ταξινόμηση βάσει του εργαλείου PolyPhen
AF†	Συχνότητα αλληλομόρφου	Global_AF	Συχνότητα αλληλομόρφου σε όλους τους πληθυσμούς από το πρόγραμμα 1000 γονιδιωμάτων
DP†	Βάθος ανάγνωσης	AFR_AF	Συχνότητα αλληλομόρφου στους Αφρικανούς από το πρόγραμμα 1000 γονιδιωμάτων
TLOD†	Πιθανότητα να είναι σωματική παραλλαγή	AMR_AF	Συχνότητα αλληλομόρφου στους Βορειοαμερικάνους από το πρόγραμμα 1000 γονιδιωμάτων
ALT_F1R2†	Αριθμός των αναγνώσεων στον προσανατολισμό F1R2 που υποστηρίζει το εναλλακτικό αλληλόμορφο	EAS_AF	Συχνότητα αλληλομόρφου στους Ανατολικοασιάτες από το πρόγραμμα 1000 γονιδιωμάτων
ALT_F2R1†	Αριθμός των αναγνώσεων στον προσανατολισμό F2R1 που υποστηρίζει το εναλλακτικό αλληλόμορφο	EUR_AF	Συχνότητα αλληλομόρφου στους Ευρωπαίους από το πρόγραμμα 1000 γονιδιωμάτων
FOXOG†	Αριθμός των εναλλακτικών αναγνώσεων που δείχνουν σφάλμα οξείδωσης κατά την προετοιμασία των βιβλιοθηκών	SAS_AF	Συχνότητα αλληλομόρφου στους Νοτιοασιάτες από το πρόγραμμα 1000 γονιδιωμάτων
QSS†	Άθροισμα των βαθμολογιών ποιότητας βάσεων για κάθε αλληλόμορφο	AA_AF	Συχνότητα αλληλομόρφου στους Αφροαμερικανούς από το πρόγραμμα 1000 γονιδιωμάτων

REF_F1R2†	Αριθμός των αναγνώσεων στον προσανατολισμό F1R2 που υποστηρίζει το αλληλόμορφο αναφοράς	EA_AF	Συχνότητα αλληλομόρφου στους Αμερικανούς ευρωπαϊκής καταγωγής από το πρόγραμμα 1000 γονιδιωμάτων
REF_F2R1†	Αριθμός των αναγνώσεων στον προσανατολισμό F2R1 που υποστηρίζει το αλληλόμορφο αναφοράς	ExAC_AF	Συχνότητα αλληλομόρφου σε όλους τους πληθυσμούς από το ExAC
Alleles	Αλληλόμορφα	ExAC_Adj_AF	Κανονικοποιημένη συχνότητα αλληλομόρφου σε όλους τους πληθυσμούς από το ExAC
Consequence	Επίπτωση	ExAC_AFR_AF	Συχνότητα αλληλομόρφου στους Αφρικανούς από το ExAC
IMPACT	Επίδραση	ExAC_AMR_AF	Συχνότητα αλληλομόρφου στους Βορειοαμερικάνους από το ExAC
SYMBOL	Σύμβολο γονιδίου	ExAC_EAS_AF	Συχνότητα αλληλομόρφου στους Ανατολικοασιάτες από το ExAC
Gene	Γονίδιο	ExAC_FIN_AF	Συχνότητα αλληλομόρφου στους Φινλανδούς από το ExAC
Feature_type	Τύπος μορίου	ExAC_NFE_AF	Συχνότητα αλληλομόρφου στους Ευρωπαίους, των Φινλανδών εξαιρουμένων, από το ExAC
Feature	Μόριο	ExAC_OTH_AF	Συχνότητα αλληλομόρφου σε άλλους πληθυσμούςαπό το ExAC
BIOTYPE	Βιοτύπος	ExAC_SAS_AF	Συχνότητα αλληλομόρφου στους Νοτιοασιάτες από το ExAC
EXON	Εξόνιο	MAX_AF	Μεγαλύτερη συχνότητα αλληλομόρφου
INTRON	Ιντρόνιο	MAX_AF_POPS	Πληθυσμός με τη μεγαλύτερη συχνότητα
HGVSc	Ονοματολογία HGVSc	CLIN_SIG	Κλινική σημασία από το ClinVar
HGVSp	Ονοματολογία HGVSp	SOMATIC	Έχει αναφερθεί ξανά ως σωματική παραλλαγή (ναι/όχι)
cDNA_position	Θέση στο συμπληρωματικό DNA	PUBMED	Δημοσίευση στο Pubmed

† Πληροφορίες για την ποιότητα της παραλλαγής που δεν παρέχονται από το VEP.

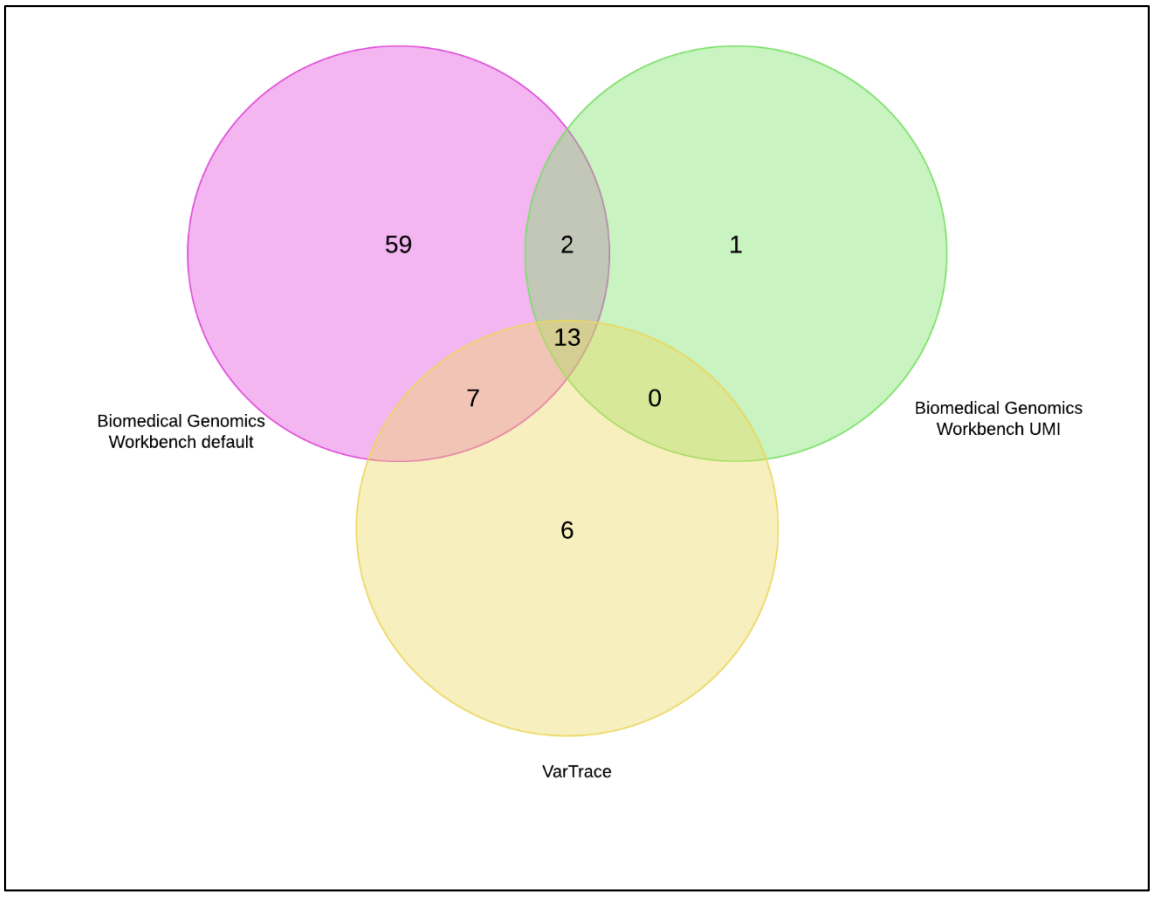
Εικόνα 38: Στιγμιότυπο οθόνης που δείχνει την οπτικοποίηση στον περιηγητή γονιδιωμάτων IGV του αρχείου BED που περιέχει τις περιοχές με μικρό βάθος ανάγνωσης. Στο συγκεκριμένο στιγμιότυπο φαίνεται το έβδομο εξόνιο του γονιδίου *BRCA1*, οι αναγνώσεις που έχουν αντιστοιχηθεί στο εξόνιο και κάτω-κάτω μια περιοχή που έχει μικρό βάθος ανάγνωσης (εδώ: βάθος ανάγνωσης < 100x).

Figure 39: Screenshot of the visualization in the IGV genome browser of the BED file containing the low-coverage genomic regions. This snapshot shows the seventh exon of *BRCA1*, the reads that have been aligned to the exon and finally a region that has low read depth (here: read depth <100x).

3.2.2. Αξιολόγηση

3.2.2.1. Ανίχνευση παθογόνων/πιθανώς παθογόνων παραλλαγών

Από τα 75 δείγματα DNA όγκου που αναλύθηκαν για την αξιολόγηση του VarTrace, ανιχνεύτηκαν 88 παθογόνοι/πιθανώς παθογόνοι παραλλαγές που πληρούσαν τα κριτήρια ποιότητας, σε 40 δείγματα. Το ποσοστό ασυμφωνίας μεταξύ των τριών ροών που συγκρίθηκαν (VarTrace και οι δύο ξεχωριστές ροές διοχέτευσης εντολών διεργασιών του λογισμικού BGWB) ήταν αρκετά μεγάλο, καθώς μόνο οι δεκατρείς παραλλαγές (14,8%) εντοπίστηκαν και από τι τρεις. Η ροή BGWBdef είχε το μεγαλύτερο ποσοστό ανίχνευσης παραλλαγών, καθώς ανίχνευσε 81 παραλλαγές, εκ των οποίων οι 59 (67% του συνόλου) ήταν μοναδικές. Η ροή BGWBumi ανίχνευσε μόλις 16 παραλλαγές εκ των οποίων η μία (1,3% του συνόλου) ήταν μοναδική. Τέλος, η ροή VarTrace ανίχνευσε 26 παραλλαγές εκ των οποίων οι 6 (6.8% του συνόλου) ήταν μοναδικές (Εικόνα 40). Τα 35 δείγματα στα οποία δεν ανιχνεύτηκε καμία παραλλαγή από καμία από τις τρεις ροές θεωρήθηκαν αληθώς αρνητικά κατά το στάδιο της αξιολόγησης.



Εικόνα 40: Διάγραμμα Venn που δείχνει τον αριθμό των παραλλαγών που ανιχνεύτηκαν από τις τρεις ροές διοχέτευσης εντολών διεργασιών.

Figure 40: Venn diagram showing the number of variants detected by the three pipelines.

3.2.2.2. Επαλήθευση παραλλαγών και αξιολόγηση ροών διοχέτευσης εντολών διεργασιών

Η επαλήθευση των παθογόνων/πιθανώς παθογόνων παραλλαγών που ανιχνεύτηκαν από τις τρεις ροές διεργασιών έδειξε πως από τις 88 παραλλαγές μόλις οι 13 ήταν πραγματικές. Αμφότερες οι ροές BGWBdef και VarTrace κατάφεραν να εντοπίσουν και τις 13 παραλλαγές που επαληθεύτηκαν με τη μέθοδο αλληλούχησης κατά Sanger, γεγονός που καθιστά την ανάκλησή τους ίση με 1, ενώ η ροή BGWBumi δεν ανίχνευσε δύο παραλλαγές κι έτσι η ανάκλησή της ανέρχεται στο 0,85. Ωστόσο, φαίνεται πως η ροή BGWBumi έδωσε το μικρότερο αριθμό ψευδώς θετικών παραλλαγών -μόλις πέντε- με την ακρίβεια της ροής να ανέρχεται στο 0,69. Στον αντίποδα, η ροή BGWBdef επέστρεψε 68 παραλλαγές οι οποίες δεν επαληθεύτηκαν με την ανεξάρτητη μέθοδο με αποτέλεσμα η ακρίβεια της ροής να είναι πολύ χαμηλή, και συγκεκριμένα 0,16. Το ποσοστό ανίχνευσης ψευδώς θετικών παραλλαγών από τη ροή διοχέτευσης εντολών διεργασιών VarTrace είναι επίσης μεγαλύτερο από της ροής BGWBumi, αλλά πολύ μικρότερο από της ροής BGWBdef, καθώς επέστρεψε 13 ψευδώς αληθείς παραλλαγές, που αντιστοιχούν σε ακρίβεια 0,5 (Πίνακας 12).

Πίνακας 12: Αποτελέσματα αξιολόγησης ροών διοχέτευσης εντολών διεργασιών.

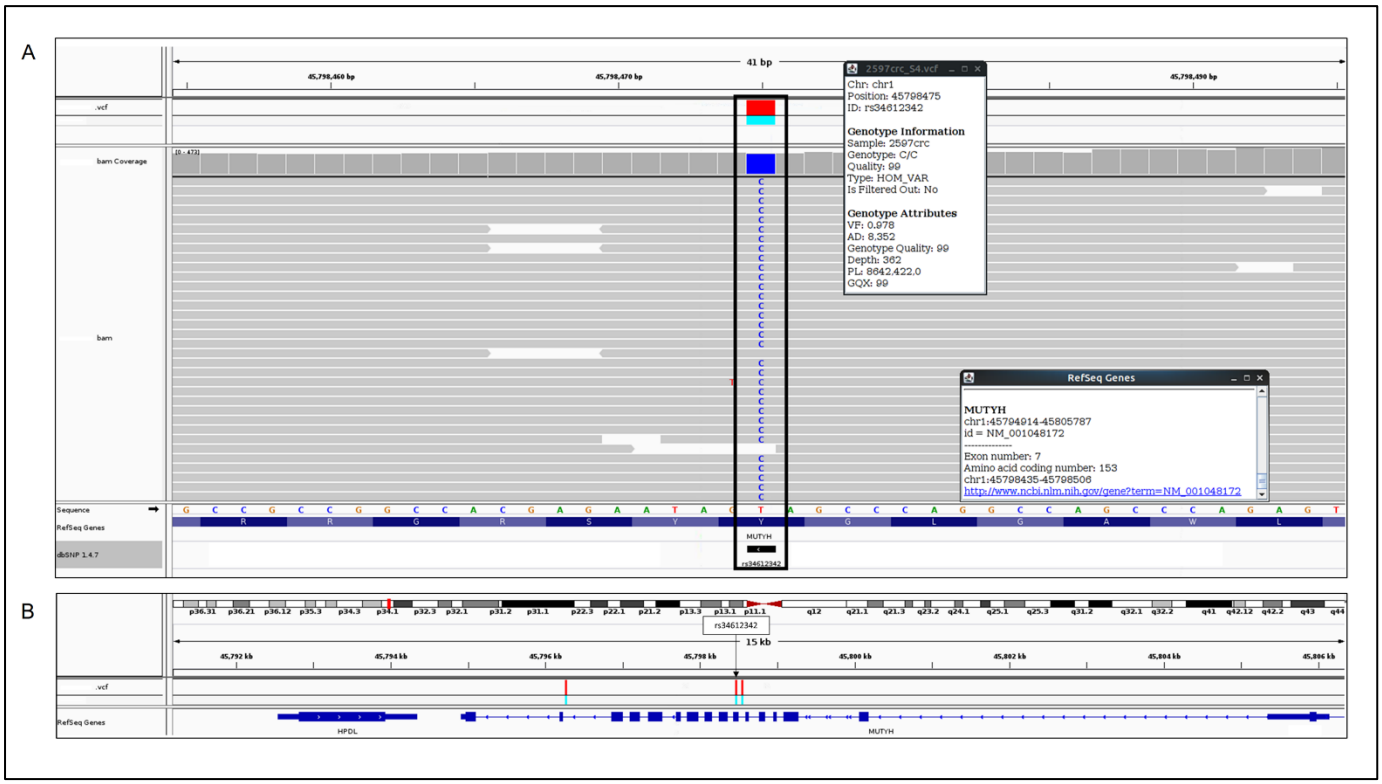
Table 12: Evaluation of the three pipelines.

	Αρνητικά		Θετικά			
	Αληθώς αρνητικά	**Ψευδώς θετικά**	**Αληθώς θετικά**	**Ψευδώς αρνητικά**	**Ακρίβεια**	**Ανάκληση**
BGWB default	42	68	13	0	*0,16*	*1*
BGWB UMI	65	5	11	2	*0,69*	*0,85*
VarTrace	57	13	13	0	*0,5*	*1*

3.3. Αξιολόγηση εμπορικού λογισμικού για τον εμπλουτισμό παραλλαγών με πληροφορίες

3.3.1. Λειτουργία των δύο λογισμικών κατά τη διαδικασία του χαρακτηρισμού των παραλλαγών

Τα δύο σύνολα μεταγράφων χαρακτηρίστηκαν χρησιμοποιώντας τα δύο λογισμικά, VS και VEP, με σκοπό την αξιολόγηση των διαφορών τους κατά το χαρακτηρισμό των παραλλαγών. Κατά τη διαδικασία του χαρακτηρισμού των παραλλαγών που εκτελείται από το VS, επιλέγεται εξ ορισμού το μεγαλύτερο μετάγραφο στην περιοχή, ανεξάρτητα από την επίπτωση της παραλλαγής στο συγκεκριμένο μετάγραφο. Στην περίπτωση όπου μία παραλλαγή βρίσκεται σε μετάγραφα που συμπίπτουν στην ίδια γονιδιωματική περιοχή ή τα οποία είναι γειτονικά (απόσταση <5.000 ζευγών βάσεων), η παραλλαγή χαρακτηρίζεται βάσει του μεγαλύτερου μεταγράφου του γονιδίου που βρίσκεται στο 5’ της συγκεκριμένης γονιδιωματικής περιοχής. Είναι αξιοσημείωτο ότι το γονίδιο αυτό ενδέχεται να μην ανήκει στα γονίδια που αποτελούν στόχους του γονιδιακού πάνελ που χρησιμοποιείται κατά την Αλληλούχηση Επόμενης Γενιάς, ενώ συχνά το μετάγραφο δεν είναι καν κωδικό. Ένα τυπικό παράδειγμα που επισημαίνεται στη μελέτη μας, περιλαμβάνει τον παραπλανητικό χαρακτηρισμό των παραλλαγών που εντοπίζονται στα εξόνια 5-16 του γονιδίου *MUTYH*, οι οποίες χαρακτηρίστηκαν βάσει του γονιδίου *HPDL* (NM_032756.2), το οποίο δεν ανήκει στην γονιδιακή ομάδα που στοχεύεται από το TruSight® Cancer Panel (Εικόνα 41).



Εικόνα 39: Στιγμιότυπο οθόνης από το πρόγραμμα περιήγησης γονιδιώματος IGV που δείχνει την ανιχνευμένη σε ομοζυγωτία παρανοηματική παραλλαγή NM_001048172.1: c.455A>G (p.Tyr152Cys) (rs34612342) του γονιδίου *MUTYH*, η οποία κατηγοριοποιείται ως παθογόνος από το ClinVar. Σύμφωνα με το VariantStudio®, χαρακτηρίστηκε με βάση το γονίδιο *HPDL* (μετάγραφο NM_032756.2), ενώ σύμφωνα με το επιλεγμένο σύνολο μεταγραφών, σχολιάστηκε με βάση το *MUTYH* (μετάγραφο NM_001048172). Η παραλλαγή *MUTYH* c.455A>G βρίσκεται περίπου 4.000 ζεύγη βάσεων καθοδικά του *HPDL*. Τροποποίηση από (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Figure 41: Screenshot from the IGV genome browser showing the *MUTYH* missense variant NM_001048172.1: c.455A>G (p.Tyr152Cys) (rs34612342) detected in homozygosity, which is classified as pathogenic in ClinVar. According to VariantStudio®, it was characterized based on the *HPDL* gene (transcript NM_032756.2), while according to the selected set of transcripts, it was characterized based on *MUTYH* (transcript NM_001048172). The *MUTYH* c.455A>G variant is found approximately 4,000 base pairs downstream of HPDL. Modification from (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Παραδόξως, ενώ υπάρχει η επιλογή να αλλάξουν οι προεπιλεγμένες παράμετροι και να οριστούν τα προεπιλεγμένα μετάγραφα για κάθε γονίδιο, δε φαίνεται να είναι λειτουργική, καθώς η διαμόρφωση του συνόλου των μεταγράφων από το χρήστη δεν αποθηκεύεται από το VS. Παρόλα ταύτα, υπάρχει μια εναλλακτική επιλογή για τον ορισμό των προτιμώμενων μεταγραφών κατά τη διάρκεια μιας ανάλυσης. Αυτή η επιλογή επιτρέπει στο χρήστη να επιλέξει το προτιμώμενο μετάγραφο με μη αυτόματο τρόπο από μία λίστα, επιτρέποντας την άμεση εξέταση των επιπτώσεων μιας παραλλαγής σε διαφορετικά μετάγραφα. Ωστόσο, αυτό το βήμα πρέπει να εκτελεστεί ξεχωριστά για κάθε παραλλαγή που εμπίπτει σε πολλά μετάγραφα, ενώ πρέπει να επαναλαμβάνεται για κάθε ανάλυση που πραγματοποιείται, καθιστώντας έτσι τον σχολιασμό περίπλοκο, χρονοβόρο και επιρρεπή σε σφάλματα.

Για τη σύγκριση του χαρακτηρισμού των παραλλαγών που πραγματοποιείται με το σύνολο των μεταγράφων που επιλέγονται από το VS (VS transcript set - VSts) με το χαρακτηρισμό των παραλλαγών βάσει του επιλεγμένο συνόλου μεταγραφών (Selected transcript set – Sts), χρησιμοποιήθηκε το λογισμικό εμπλουτισμού με πληροφορίες VEP της Ensembl. Με το VEP, ο εμπλουτισμός των παραλλαγών με πληροφορίες γίνεται με ευέλικτο τρόπο μέσω ενός σεναρίου γραμμής εντολών σε γλώσσα Perl, επιτρέποντας στον χρήστη να ορίζει παραμέτρους και φίλτρα, σύμφωνα με τις ανάγκες της εκάστοτε ανάλυσης. Ο σχολιασμός μέσω ενός εργαλείου γραμμής εντολών απαιτεί εξειδίκευση και εμπειρία, αλλά ο καθορισμός των προτιμώμενων παραμέτρων είναι μια εφάπαξ διαδικασία, για την οποία απαιτούνται ελάχιστες ή καθόλου προσαρμογές κατά τη διάρκεια επαναληπτικών αναλύσεων.

3.3.2. Ανίχνευση και χαρακτηρισμός παραλλαγών

Συνολικά, ανιχνεύτηκαν 5.912 παραλλαγές σε περιοχές που στοχεύονται από το γονιδιακό πάνελ TruSight® Cancer, εκ των οποίων οι 558 αφορούσαν περιοχές που έχουν ανιχνευθεί σε μελέτες συσχέτισης ολόκληρου του γονιδιώματος, και εξαιρέθηκαν από την ανάλυση. Επομένως, 5.354 παραλλαγές υποβλήθηκαν τελικά σε επεξεργασία για εμπλουτισμό με πληροφορίες από τα δύο λογισμικά.

Με βάση το VSts, 116 παραλλαγές χαρακτηρίστηκαν ως απώλειας λειτουργίας και 31 ως παραλλαγές σε θέση ματίσματος. Επιπλέον, 2.574 και 2.633 παραλλαγές ήταν σε κωδικές και μη κωδικές περιοχές αντίστοιχα, και δεν είχαν ως αποτέλεσμα την απώλεια της λειτουργίας της πρωτεΐνης. Αφετέρου, ο εμπλουτισμός με παραλλαγές βάσει του Sts είχε ως αποτέλεσμα, 131 παραλλαγές να χαρακτηριστούν ως απώλειας λειτουργίας και 33 ως παραλλαγές σε θέση ματίσματος. Επιπλέον, 2.828 και 2.362 παραλλαγές εντοπίστηκαν σε κωδικές και μη κωδικές περιοχές αντίστοιχα, χωρίς απώλεια της λειτουργίας της πρωτεΐνης. Ο Πίνακας 13 συνοψίζει το αποτέλεσμα του εμπλουτισμού των παραλλαγών με πληροφορίες χρησιμοποιώντας καθένα από τα δύο σύνολα μεταγράφων.

Πίνακας 13: Σύνοψη του αριθμού των σχολιασμένων παραλλαγών και αντιστοιχία μεταξύ των δύο συνόλων μεταγράφων που χρησιμοποιήθηκαν, κατανεμημένα ανά τύπο παραλλαγής. Τροποποίηση από (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Table 13: Summary of the number of annotated variants and concordance between the two transcript sets used, broken down by variant type. Modification from (Kalfakakou, Konstantopoulou*, et al.*, 2021)

	VSts+Sts*	VSts†	Sts‡	Ποσοστό συμφωνίας (%) §
Συνολικός αριθμός χαρακτηρισμών	**5859**	**5354**	**5354**	**82.82**
Παραλλαγές απώλειας λειτουργίας	131	116	131	88.55
πλαισιοτροποποιητική_παραλλαγή	72	59	72	81.94
απώλεια_κωδικονίου_έναρξης	3	3	3	100
εισαγωγή_πρόωρου_κωδικόνιου_τερματισμού	53	51	53	96.22
εισαγωγή_πρόωρου_κωδικόνιου_τερματισμού, πλαισιοτροποποιητική_παραλλαγή	2	2	2	100
εισαγωγή_πρόωρου_κωδικόνιου_τερματισμού, παραλλαγή_περιοχής_ματίσματος	1	1	1	100
Παραλλαγές σε θέση ματίσματος	33	31	33	93.94
παραλλαγή_δέκτη_ματίσματος	14	12	14	85.71
παραλλαγή_δότη_ματίσματος	16	16	16	100
παραλλαγή_δότη_ματίσματος, παραλλαγή_κωδικής_ακολουθίας	1	1	1	100
παραλλαγή_δότη_ματίσματος,παραλλαγή_κωδικής_ακολουθίας,παραλλαγή_ιντρονίου	1	1	1	100
παραλλαγή_δότη_ματίσματος, παραλλαγή_ιντρονίου	1	1	1	100
Άλλες παραλλαγές σε κωδικές περιοχές	2831	2574	2828	90.92
απαλοιφή_εντός_πλαισίου_ανάγνωσης	31	28	31	90.32
απαλοιφή_εντός_πλαισίου_ανάγνωσης, παραλλαγή_περιοχής_ματίσματος	1	1	1	100
εισαγωγή_εντός_πλαισίου_ανάγνωσης	6	5	6	83.33
παρανοηματική_παραλλαγή	1587	1433	1585	90.29
παρανοηματική_παραλλαγή, παραλλαγή_περιοχής_ματίσματος	47	39	47	82.98
παραλλαγή_περιοχής_ματίσματος, συνώνυμη_παραλλαγή	26	24	26	92.31
συνώνυμη_παραλλαγή	1133	1044	1132	92.14
Παραλλαγές σε μη κωδικές περιοχές	2864	2633	2362	82.47
παραλλαγή_3_άκρου_UTR	95	95	79	83.16
παραλλαγή_5_άκρου_UTR	74	58	74	78.37
παραλλαγή_περιοχής_3_άκρου	344	344	0	0
παραλλαγή_ιντρονίου	1975	1783	1975	90.28
παραλλαγή_ιντρονίου, παραλλαγή _μη_κωδικού_μεταγράφου	11	11	0	0
παραλλαγή_εξονίου_μη_κωδικού_μεταγράφου,παραλλαγή_μη_κωδικού_μεταγράφου	7	7	0	0
παραλλαγή_περιοχής_ματίσματος, παραλλαγή_3_άκρου_UTR	1	0	1	0
παραλλαγή_περιοχής_ματίσματος, παραλλαγή_ιντρονίου	229	208	229	90.82
παραλλαγή_περιοχής_ματίσματος, παραλλαγή_εξονίου_μη_κωδικού_μεταγράφου	1	1	0	0
παραλλαγή_περιοχής_5_άκρου	127	127	4	31.49

* Αριθμός χαρακτηρισμών που εξήχθησαν από τη χρήση και των δύο συνόλων μεταγράφων.

† Αριθμός χαρακτηρισμών που εξήχθησαν από τη χρήση του συνόλου μεταγράφων του VariantStudio.

‡ Αριθμός χαρακτηρισμών που εξήχθησαν από τη χρήση του συνόλου των επιλεγμένων μεταγράφων.

§ Το ποσοστό συμφωνίας υπολογίστηκε διαιρώντας το σύνολο των συμφωνηθέντων χαρακτηρισμών διά τον συνολικό αριθμό των χαρακτηρισμών που εξήχθησαν για κάθε κατηγορία χρησιμοποιώντας και τα δύο σύνολα μεταγράφων.

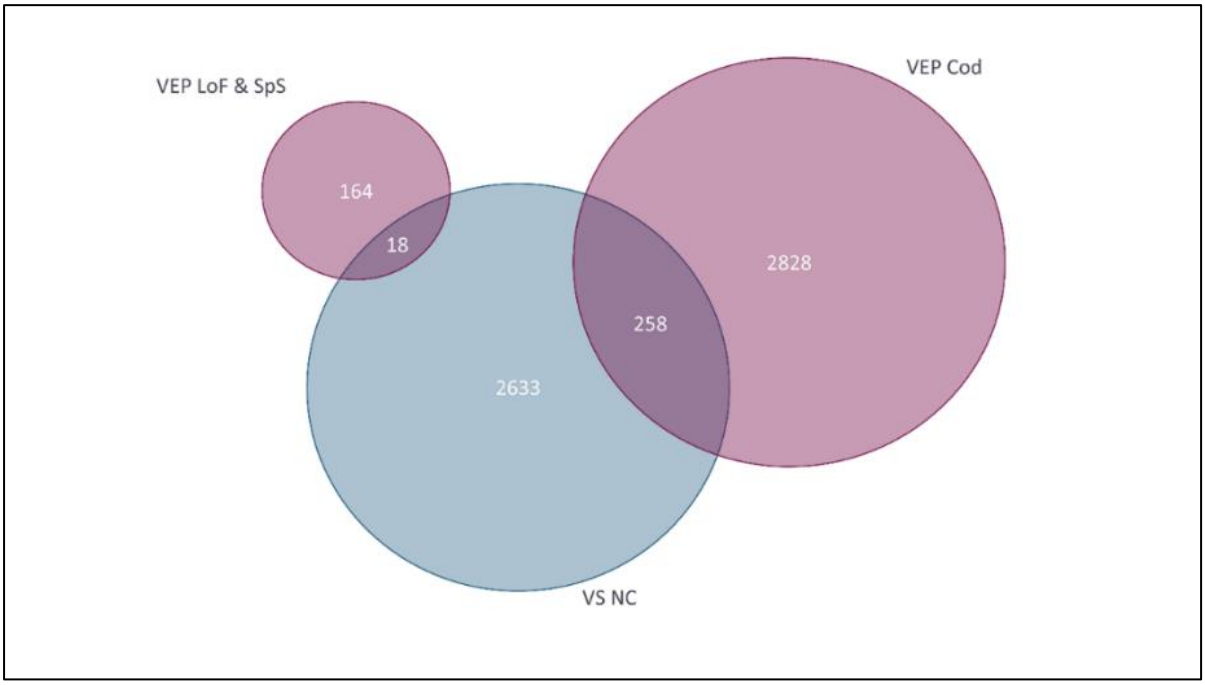
3.3.3. Ασυμφωνία μεταξύ των δύο συνόλων μεταγράφων

Συνολικά, σε όλες τις 5.354 παραλλαγές που χαρακτηρίστηκαν, το ποσοστό συμφωνίας των επιπτώσεων ήταν 82,82%. Η ασυμφωνία μεταξύ των συνόλων Sts και VSts, σχετικά με τις παραλλαγές απώλειας λειτουργίας ήταν 11,45% (Sts: 131, VSts: 116, συμφωνία: 116). Οι 15 παραλλαγές για τις οποίες δεν υπήρχε συμφωνία μεταξύ των δύο συνόλων μεταγράφων, χαρακτηρίστηκαν λανθασμένα χρησιμοποιώντας το VSts, καθώς είτε ερμηνεύθηκαν ως παραλλαγές 3' UTR, είτε ως παραλλαγές στη διαγονιδιακή περιοχή που βρίσκεται στο 3' ή 5' του γονιδίου, είτε ως παραλλαγές που βρίσκονται όντως σε εξόνιο, το οποίο όμως ανήκει σε μη κωδικό μετάγραφο. Από αυτές τις παραλλαγές, δύο εντοπίζονται στο γονίδιο *MLH1*, μία στο *RAD51D* και μία στο *TSC2*, τα οποία προδιαθέτουν στο σύνδρομο Lynch (Fostira *et al.*, 2007; Lynch & de la Chapelle, 1999), σε κληρονομικό καρκίνο του μαστού και των ωοθηκών (Apostolou & Fostira, 2013) και το σύμπλεγμα οζώδους σκλήρυνσης (Marcotte & Crino, 2006), αντίστοιχα και κληρονομούνται με αυτοσωμικό επικρατή τρόπο. Επιπλέον, δύο παραλλαγές σε καθένα από τα γονίδια *MUTYH*, *FANCL* και *ERCC2*, χαρακτηρίστηκαν επίσης λανθασμένα. Οι διαλληλικές παθογόνοι παραλλαγές σε αυτά τα γονίδια προδιαθέτουν στη *MUTYH*-σχετιζόμενη αδενωματώδη πολυποδίαση (Nielsen *et al.*, 2011), στην αναιμία Fanconi (de Winter & Joenje, 2009) και στη μελαγχρωματική ξηροδερμία (Lehmann *et al.*, 2011), αντίστοιχα. Τέλος, πέντε παθογόνοι παραλλαγές στο RECQL4 και μία στο XPA χαρακτηρίστηκαν λανθασμένα. Οι διαλληλικές παθογόνοι παραλλαγές σε αυτά τα δύο γονίδια προδιαθέτουν στα σύνδρομα Rothmund-Thomson (Kitao *et al.*, 1999) και μελαγχρωματικής ξηροδερμίας (Lehmann *et al.*, 2011), αντίστοιχα. Όλες οι προαναφερθείσες παθογόνοι παραλλαγές, εκτός από μια παραλλαγή στο γονίδιο *MUTYH*, εντοπίστηκαν σε ετεροζυγωτία. Ένας ασθενής που έφερε την παραλλαγή *MUTYH* c.960_961delGC, ήταν σύνθετος ετεροζυγώτης, καθώς έφερε και την παραλλαγή *MUTYH* c.653G>A (rs140342925) *in trans*, μια παθογόνο παρανοηματική παραλλαγή.

Μεταξύ των παραλλαγών που πιθανώς επηρεάζουν το μάτισμα, καταγράφηκε ένα ποσοστό ασυμφωνίας 6,1% (Sts: 33, VSts: 31, συμφωνία: 31). Οι δύο παραλλαγές με ασυμφωνία στο χαρακτηρισμό, που βρέθηκαν σε ετεροζυγωτία, χαρακτηρίστηκαν από το VS ως παραλλαγές στη διαγονιδιακή περιοχή που βρίσκεται στο 5' των γονιδίων *NCBP1* (NM_002486.4) και *MFSD3* (NM_138431.1), ενώ στην πραγματικότητα εντοπίζονται στους δέκτες ματίσματος του πέμπτου και δέκατου ιντρονίου των *XPA* και *RECQL4*, αντίστοιχα.

Όσον αφορά τις κωδικές παραλλαγές που δεν έχουν ως αποτέλεσμα την απώλεια λειτουργίας της πρωτεΐνης, εντοπίστηκαν 262 ασυμφωνίες μεταξύ των δύο συνόλων μεταγράφων, οι οποίες αντιστοιχούν σε ποσοστό ασυμφωνίας 9,25% (Sts: 2.828, VSts: 2.574, συμφωνία: 2.569, συνολικοί χαρακτηρισμοί: 2.831). Από αυτές, το VS χαρακτήρισε 23 ως παραλλαγές που βρίσκονται στο 3 'UTR ή σε ιντρόνιο, 228 ως παραλλαγές στη διαγονιδιακή περιοχή που βρίσκεται στο 5' ή 3' ενός άλλου γονιδίου, κι 7 ως παραλλαγές που εντοπίζονται σε εξόνιο ενός μη κωδικού μεταγράφου. Αφετέρου, χρησιμοποιώντας το Sts, δύο παρανοηματικές και μία συνώνυμη παραλλαγή χαρακτηρίστηκαν ως παραλλαγές σε περιοχή ιντρονίου. Από τις παραλλαγές που εντοπίστηκαν σε κωδικές περιοχές και για τις οποίες υπάρχει ασυμφωνία μεταξύ των δύο λογισμικών, οι 152 βρέθηκαν στη βάση ClinVar, εκ των οποίων οι 92 κατηγοριοποιούνται ως ουδέτερες/πιθανώς ουδέτερες, οι 54 κατηγοριοποιούνται ως

παραλλαγές αγνώστου σημασίας και οι 6 θεωρούνται παθογόνοι/πιθανώς παθογόνοι. Είναι αξιοσημείωτο πως όλες οι παραλλαγές αγνώστου σημασίας και οι παθογόνοι/πιθανώς παθογόνοι παραλλαγές χαρακτηρίζονται λανθασμένα από το VS. Από αυτές, πέντε και μία αφορούν επιβλαβείς παραλλαγές στα γονίδια *MUTYH* και *FANCA*, αντίστοιχα (διαλληλικές παθογόνοι παραλλαγές στο *FANCA* προδιαθέτουν στην αναιμία Fanconi (de Winter & Joenje, 2009)). Ο αριθμός των παραλλαγών που χαρακτηρίζονται ως απώλειας λειτουργίας, θέσεων ματίσματος ή παραλλαγές σε κωδικές περιοχές με τη χρήση του Sts, αλλά επισημαίνονται ως μη κωδικές με τη χρήση του VSts απεικονίζεται στο διάγραμμα Venn της Εικόνας 42.



Εικόνα 40: Διάγραμμα Venn που δείχνει τον αριθμό των παραλλαγών που χαρακτηρίζονται ως παραλλαγές απώλειας λειτουργίας, παραλλαγές που επηρεάζουν τη θέση ματίσματος ή παραλλαγές σε κωδικές περιοχές από το επιλεγμένο σύνολο μεταγράφων μας, αλλά χαρακτηρίστηκαν ως παραλλαγές σε μη κωδικές περιοχές από το VariantStudio® (18/164: 10,97% και 258/2828: 9,12 % αντίστοιχα). VEP: Επιλεγμένο σύνολο μεταγράφων που χρησιμοποιήθηκε κατά τον χαρακτηρισμό με το Variant Effect Predictor· VS: Σύνολο μεταγράφων VariantStudio®· LoF: Παραλλαγές απώλειας λειτουργίας· SpS: Παραλλαγές σε θέσεις ματίσματος· Cod: Παραλλαγές σε κωδικές περιοχές· NC: Παραλλαγές σε μη κωδικές περιοχές. Τροποποίηση από (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Figure 42: Venn diagram showing the number of variants characterized as loss-of-function variants, variants affecting a splice site, or variants in coding regions by our selected transcript set but were identified as variants in non-coding regions by VariantStudio® (18/164: 10.97% and 258/2828: 9.12% respectively).

VEP: Selected set of transcripts used during annotation using Variant Effect Predictor; LoF: Loss-of-function variants; SpS: Splice site variants; Cod: Coding variants; NC: Non-coding variants. Modification from (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Μεταξύ των μη κωδικών παραλλαγών, σημειώθηκαν 242 ασυμφωνίες, αριθμός που αντιστοιχεί σε ποσοστό ασυμφωνίας 26,33% (Sts: 2,362, VSts: 2,633, συμφωνία: 2.110, συνολικοί χαρακτηρισμοί: 2.864). Με βάση το VSts, οι 164 και 61 σχολιάστηκαν ως παραλλαγές στη διαγονιδιακή περιοχή που βρίσκεται στο 5’ ή 3’ ενός άλλου γονιδίου, αντίστοιχα, ενώ στην πραγματικότητα εντοπίζονται εντός του ιντρονίου ή στο 5’ ή 3’ UTR ενός γονιδίου που στοχεύεται από το TruSight Cancer Panel, βάσει του χαρακτηρισμού τους με τη χρήση του Sts. Επιπλέον, δεκατρείς παραλλαγές χαρακτηρίστηκαν λανθασμένα ως παραλλαγές 3’ ή 5’ UTR ενός γονιδίου ενδιαφέροντος χρησιμοποιώντας το VSts, ενώ στην πραγματικότητα ήταν ιντρονικές βάσει του χαρακτηρισμού τους με τη χρήση του Sts. Από τις προαναφερθείσες παραλλαγές, 61 αναφέρονται στη βάση δεδομένων ClinVar, εκ των οποίων οι 52 κατηγοριοποιούνται ως ουδέτερες/πιθανώς ουδέτερες, επτά ως αγνώστου σημασίας και τρεις ως παθογόνοι παραλλαγές. Από τις τελευταίες, μία και δύο εντοπίζονται στο *MUTYH* και *MLH1*, αντίστοιχα. Επιπλέον, τρεις παραλλαγές ήταν είτε παρανοηματικές είτε συνώνυμες για τα μετάγραφα που επιλέχθηκαν από το VS (σε καθένα από τα *RHBDF2*, *MUTYH* και *CDKN2A*). Αυτές οι παραλλαγές είναι ουδέτερες/πιθανώς ουδέτερες σύμφωνα με το ClinVar. Οι λανθασμένοι χαρακτηρισμοί των παθογόνων παραλλαγών από το VS συνοψίζονται στον Πίνακα 14, ενώ ο Πίνακας 15 συνοψίζει όλα τα γονίδια/περιοχές ενδιαφέροντος που χαρακτηρίστηκαν λανθασμένα, εξαιτίας του επιλεγμένου μεταγράφου, μαζί με τον αριθμό των παραλλαγών ανά γονίδιο.

Πίνακας 14: Παθογόνοι παραλλαγές με ασυμφωνία των χαρακτηρισμών τους, χρησιμοποιώντας τα δύο διαφορετικά σύνολα μεταγράφων. Τροποποίηση από (Kalfakakou*, et al.*, 2021)

Table 14: Pathogenic variants with discordant annotations, using the two different transcript sets. Modification from (Kalfakakou*, et al.*, 2021)

Γονίδιο	Θέση	Αλληλόμορφα	VS Μετάγραφο	VS CSQ	VEP Μετάγραφο	VEP CSQ	VEP HGVSc	VEP HGVSp	ClinVar Ταξινόμηση
MUTYH	1:45795006	GC/G	NM_032756.2	ds	NM_001048172.1	fs	c.1540del	p.Ala514Hisfs	-
MUTYH	1:45796890	TTCC/T	NM_032756.2	ds	NM_001048172.1	ind	c.1356_1358del	p.Glu453del	Παθ
MUTYH	1:45797228	C/T	NM_032756.2	ds	NM_001048172.1	ms, sr	c.1106G>A	p.Gly369Asp	Παθ /ΠΠαθ
MUTYH	1:45797476	GGC/G	NM_032756.2	ds	NM_001048172.1	fs	c.960_961del	p.Glu320Aspfs	-
MUTYH	1:45798117	C/T	NM_032756.2	ds	NM_001048172.1	ms	c.653G>A	p.Arg218His	Παθ /ΠΠαθ
MUTYH	1:45798458	G/A	NM_032756.2	ds	NM_001048172.1	ms	c.472C>T	p.Arg158Trp	ΠΠαθ /ΑΣ
MUTYH	1:45798475	T/C	NM_032756.2	ds	NM_001048172.1	ms	c.455A>G	p.Tyr152Cys	Παθ
MUTYH	1:45798558	TCCTATTTCCCCTA/T	NM_032756.2	ds	NM_001048172.1	int	c.423+19_423+31del	-	Παθ /ΠΠαθ
MLH1	3:37035068	GC/G	NM_014805.3	us	NM_000249.3	fs	c.31del	p.Leu11Trpfs	Παθ
MLH1	3:37035105	G/T	NM_014805.3	us	NM_000249.3	sg	c.67G>T	p.Glu23Ter	Παθ
MLH1	3:37035159	G/C	NM_014805.3	us	NM_000249.3	sr, int	c.116+5G>C	-	Παθ
MLH1	3:37038205	G/A	NM_014805.3	us	NM_000249.3	sr, int	c.207+5G>A	-	-
RECQL4	8:145737068	GC/G	NM_005309.2	ds	NM_004260.3	fs	c.3498delG	p.His1167Ilefs	-
RECQL4	8:145737690	CT/C	NM_138431.1	ds	NM_004260.3	fs	c.3073delA	p.Arg1025Glyfs	-
RECQL4	8:145737781	TG/T	NM_138431.1	ds	NM_004260.3	fs	c.3049delC	p.Gln1017Argfs	-
RECQL4	8:145738764	AC/A	NM_138431.1	ds	NM_004260.3	fs	c.2300delG	p.Gly767Valfs	-
XPA	9:100437757	AGTACAAGTCTTACG/A	NM_002486.4	ds	NM_000380.3	fs	c.772_785del	p.Arg258Tyrfs	ΠΠαθ
TSC2	16:2098654	AGAAGTTTAAGATTCTGTTGGGACTGG/A	NM_002528.5	us	NM_000548.3	fs	c.42_67del	p.Lys14Asnfs	-
FANCA	16:89807249	GAGA/G	NM_001113525.1	3’	NM_000135.2	ind	c.3788_3790del	p.Phe1263del	Παθ
RAD51D	17:33428320	C/T	NR_037714.1	nce, nc	NM_002878.3	sg	c.803G>A	p.Trp268Ter	Παθ
ERCC2	19:45855800	GC/G	NM_177417.2	ds	NM_000400.3	fs	c.2009delG	p.Gly670Alafs	-
ERCC2	19:45856550	CA/C	NM_177417.2	ds	NM_000400.3	fs	c.1707delT	p.Ile569Metfs	-

VS: VariantStudio· VEP: Variant Effect Predictor· CSQ: Επίπτωση· ds: παραλλαγή_περιοχής_3_άκρου· us: παραλλαγή_περιοχής_5_άκρου· 3’: παραλλαγή_3_άκρου_UTR· nce: παραλλαγή_εξονίου_μη_κωδικού_μεταγράφου· nc: παραλλαγή_μη_κωδικού_μεταγράφου· fs:

πλαισιοτροποποιητική_παραλλαγή· ind: απαλοιφή_εντός_πλαισίου_ανάγνωσης· ms: παρανοηματική_παραλλαγή· sr: παραλλαγή_περιοχής_ματίσματος· int παραλλαγή_ιντρονίου· sg: εισαγωγή_πρόωρου_κωδικόνιου_τερματισμού· Παθ: Παθογόνος· ΠΠαθ: Πιθανώς Παθογόνος· ΑΣ: Αγνώστου Σημασίας

Πίνακας 15: Σύνοψη της ασυμφωνίας μεταξύ των μορίων που χαρακτηρίστηκαν βάσει των δύο συνόλων μεταγράφων. Τροποποίηση από (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Table 15: Summary of discordant molecules annotated using both transcript sets. Modification from (Kalfakakou, Konstantopoulou*, et al.*, 2021)

Συντεταγμένες γονιδίων στο TCP		**Μόριο που χαρακτηρίστηκε από το VS**				**Επιλεγμένο μετάγραφο για τον χαρακτηρισμό βάσει του VEP**				
Χρωμόσωμα	Συντεταγμένες	Όνομα	Τύπος	Μετάγραφο	Μήκος (ζβ)	Γονίδιο	Μετάγραφο	Μήκος (ζβ)	Εξόνια που επηρεάζονται	Αριθμός παραλλαγών
1	45794835-45806142	*HPDL*	γονίδιο	NM_032756.2	1803	*MUTYH*	NM_001048172.1	1794	5-16	48
1	161284047-161332984	*MPZ*	γονίδιο	NM_000530.6	1980	*SDHC*	NM_003001.3	2858	1	4
2	58386378-58468507	*VRK2*	γονίδιο	NM_006296.6	1986	*FANCL*	NM_018062.3	1738	9-14	18
3	10068098-10143614	CIDECP1	ncRNA	NR_002786.1	803	*FANCD2*	NM_033084.3	5204	1,2	3
3	14186647-14220283	*TMEM43*	γονίδιο	NM_024334.2	3343	*XPC*	NM_004628.4	3729	13-16	14
3	37034823-37107380	*EPM2AIP1*	γονίδιο	NM_014805.3	7439	*MLH1*	NM_000249.3	2662	1,2	11
3	52435029-52444366	*DNAH1*	γονίδιο	NM_015512.4	13126	*BAP1*	NM_004656.3	3717	11-17	23
7	6012870-6048756	*RSPH10B*	γονίδιο	NM_173565.3	3122	*PMS2*	NM_000535.5	2851	15	4
8	145736667-145743229	*MFSD3*	γονίδιο	NM_005309.2	1896	*RECQL4*	NM_004260.3	3840	5-22	121
9	21967751-21995300	CDKN2A-DT	lncRNA	NR_024274.1	616	*CDKN2A*	NM_058195.3	1164	1-3	13
9	35073832-35080013	*VCP*	γονίδιο	NM_007126.3	3859	*FANCG*	NM_004629.1	2649	7-14	23
9	100437191-100459639	*NCBP1*	γονίδιο	NM_002486.4	5381	*XPA*	NM_000380.3	1491	6	2
10	89622870-89731687	*KLLN*	γονίδιο	NM_001126049.1	4277	*PTEN*	NM_000314.4	5572	1	2
10	104263744-104393292	*ACTR1A*	γονίδιο	NM_005736.3	2891	*SUFU*	NM_016169.3	4994	1	6
11	61197514-61215001	*CPSF7*	γονίδιο	NM_024811.3	3764	*SDHAF2*	NM_017841.2	1227	1	2
11	64570982-64578766	*MAP4K2*	γονίδιο	NM_004579.3	2971	*MEN1*	NM_130799.2	2770	3-10	20
11	95523129-95565857	*FAM76B*	γονίδιο	NM_144664.4	3958	*CEP57*	NM_014679.4	3192	1	3
11	111957497-111990353	*NKAPD1*	γονίδιο	NM_018195.3	3757	*SDHD*	NM_003002.3	1395	1-3	9
12	58141510-58149796	*TSPAN31*	γονίδιο	NM_005981.3	1749	*CDK4*	NM_000075.3	2020	2-8	16
13	32889611-32973805	*ZAR1L*	γονίδιο	NM_001136571.1	1103	*BRCA2*	NM_000059.3	11386	1,2	2
13	48877887-49056122	RB1-DT	lncRNA	NR_046414.1	1191	*RB1*	NM_000321.2	4772	1,2	6
13	103497194-103528345	*BIVM*	γονίδιο	NM_017693.3	3873	*ERCC5*	NM_000123.3	4091	1	1
14	45605143-45670093	*FKBP3*	γονίδιο	NM_002013.3	1353	*FANCM*	NM_020937.2	7144	1,2	14
16	2097466-2138716	*NTHL1*	γονίδιο	NM_002528.5	1067	*TSC2*	NM_000548.3	5675	1-3	4
16	3631182-3661599	*NLRC3*	γονίδιο	NM_178844.2	6401	*SLX4*	NM_032444.2	7304	15	6
16	89803957-89883065	*ZNF276*	γονίδιο	NM_001113525.1	4621	*FANCA*	NM_000135.2	5460	36-38	37

17	29421945-29709134	mir4733	miRNA	NR_039886.1	76	*NF1*	NM_000267.3	12381	1	2
17	33426811-33448541	RAD51L3-RFFL	lncRNA	NR_037714.1	3904	*RAD51D*	NM_002878.3	2418	1-9	19
17	56769934-56811703	*TEX14*	γονίδιο	NM_001201457.1	4954	*RAD51C*	NM_058216.2	1337	1-3	10
17	74466973-74497872	*AANAT*	γονίδιο	NM_001166579.1	1935	*RHBDF2*	NM_001005498.3	3453	9-18	33
19	45853095-45874176	*KLC3*	γονίδιο	NM_177417.2	1793	*ERCC2*	NM_000400.3	2568	16-23	46
22	24129150-24176703	*MMP11*	γονίδιο	NM_005940.3	2276	*SMARCB1*	NM_003073.3	1717	1	1

TCP: TruSight Cancer Panel· VEP: Variant Effect Predictor· VS: VariantStudio· ζβ: ζεύγη βάσεων· lncRNA: Μεγάλο μη κωδικό RNA· ncRNA: μη κωδικό RNA· miRNA: μικρό RNA

4. ΣΥΖΗΤΗΣΗ

4.1. Βάση δεδομένων CanVaS

Η βάση δεδομένων CanVaS αποτελεί μία πηγή δεδομένων που έχει ως στόχο την καταγραφή της γενετικής ετερογένειας των ασθενών με καρκίνο στην Ελλάδα. Η CanVaS περιλαμβάνει ένα ολοκληρωμένο σύνολο σπάνιων παραλλαγών που έχουν ανιχνευθεί στο DNA γαμετικής σειράς ασθενών με καρκίνο, οι οποίες συνοδεύονται από την κατηγοριοποίησή τους βάσει των κανόνων HGVS και τη συχνότητα αλληλομόρφου στον ελληνικό πληθυσμό. Η συλλογή αποτελείται από περίπου 2.200 και 9.200 παθογόνους/πιθανώς παθογόνους και ουδέτερες/πιθανώς ουδέτερες παραλλαγές, αντίστοιχα, καθιστώντας τη βάση μία πολύτιμη πηγή γενετικών παραλλαγών σε γονίδια που προδιαθέτουν στον καρκίνο. Επιπλέον, η CanVaS περιλαμβάνει μεγάλο αριθμό παραλλαγών με άγνωστη κλινική σημασία, για τις οποίες απαιτούνται πρόσθετες πληροφορίες για να καταστεί δυνατή η κατηγοριοποίησή τους. Η εξαιρετικά λεπτομερής καταγραφή των παραλλαγών και το γεγονός πως εξετάζεται συγκεκριμένος πληθυσμός μπορούν να υποστηρίξουν την ιεράρχηση και την κατηγοριοποίηση των παραλλαγών αγνώστου κλινικής σημασίας.

Συμπληρωματικά, η CanVaS παρέχει τη γεωγραφική κατανομή της παραλλαγής στον ελλαδικό χώρο, ενώ καταγράφει τις ελληνικές ιδρυτικές παραλλαγές και την εντοπιότητά τους, επιτρέποντας τη μελέτη υποπληθυσμών και απομονωμένων ελληνικών πληθυσμών και τη διερεύνηση πιθανών ιδρυτικών φαινομένων. Τα ισχυρά ιδρυτικά φαινόμενα που παρατηρούνται σε ελληνικούς υποπληθυσμούς έχουν επισημανθεί πρόσφατα σε μια γενετική ανάλυση που πραγματοποιήθηκε σε απομονωμένους πληθυσμούς στην Κρήτη, όπου αποδείχθηκε ότι τρεις παθογόνοι παραλλαγές στα *BRCA1* και *BRCA2*, μοναδικές στον κρητικό υποπληθυσμό, και συγκεκριμένα η NM_007294.3:c.5492del στο *BRCA1* και οι NM_000059.3:c.7806-2A>T και NM_000059.3:c.6842-2675_7008-5558del στο *BRCA2*, αντιπροσωπεύουν το 48% των ταυτοποιημένων παθογόνων παραλλαγών στα *BRCA1* και *BRCA2* στους ασθενείς με καρκίνο μαστού και ωοθηκών με καταγωγή από την Κρήτη (Apostolou *et al.*, 2020).

Επιπλέον, η εις βάθος μελέτη του ελληνικού πληθυσμού και η αξιολόγηση των ιδρυτικών φαινομένων έχουν επιτρέψει τον προσδιορισμό της παθογένειας σπάνιων παραλλαγών, που περιορίζονται στους Έλληνες, όπως στην περίπτωση της παραλλαγής NM_007294.3:c.5212G>A (p. Gly1738Arg) στο γονίδιο *BRCA1.* Πρόκειται για μία παθογόνο παραλλαγή η οποία είναι ιδρυτική για τον ελληνικό πληθυσμό και έχει ανιχνευτεί σε 53 οικογένειες μέχρι σήμερα. Η παθογένειά της ήταν δύσκολο να αποδειχθεί, καθώς πρόκειται για μία παρανοηματική παραλλαγή. Ωστόσο, η επανεμφάνισή της σε Ελληνίδες ασθενείς με καρκίνο μαστού/ωοθηκών κατέστησε δυνατή την κατηγοριοποίησή της, επιτρέποντας τη σωστή διαχείριση των ασθενών που τη φέρουν, όχι μόνο στην Ελλάδα, αλλά και στην ελληνική διασπορά (Anagnostopoulos *et al.*, 2008). Αξίζει επιπλέον να αναφερθεί ότι δύο ακόμα ιδρυτικές παθογόνοι παραλλαγές στο γονίδιο *BRCA1* ελληνικών υποπληθυσμών του τύπου των μεγάλων γονιδιωματικών αναδιατάξεων έχουν χαρακτηριστεί: η NM_007294.1:c.5407-754_*8273del (απαλοιφή των εξονίων 23 και 24 του γονιδίου) με προέλευση από τα Μικρασιατικά παράλια και η NM_007294.1:c.5468-285_*4019delinsCACAG (απαλοιφή του εξονίου 24 του γονιδίου) με προέλευση από τους Έλληνες του Πόντου (Apostolou *et al.*, 2017; Konstantopoulou *et al.*, 2014; Pertesi *et al.*, 2011; Armaou *et al.*, 2009). Η CanVas διαθέτει πλούσια δεδομένα για όλες τις ιδρυτικές ελληνικές παθογόνους

παραλλαγές, συμβάλλοντας αποφασιστικά στην κλινική τους αξιολόγηση όταν εντοπίζονται εκτός Ελλάδας.

Τα γενετικά δεδομένα που καταγράφονται στη CanVaS συνοδεύονται με πληροφορίες σχετικά με το φαινότυπο των ατόμων που φέρουν την κάθε παραλλαγή. Τα φαινοτυπικά και κλινικά χαρακτηριστικά είναι ένας τύπος δεδομένων που σπανίως κοινοποιείται σε γενετικές πηγές δεδομένων. Αυτή η προσθήκη έχει ιδιαίτερη κλινική χρησιμότητα, καθώς επιτρέπει την κλινική εφαρμογή των δεδομένων μέσω της κατάλληλης κατηγοριοποίησης των παραλλαγών. Μεταξύ των δεδομένων αυτού του τύπου, περιλαμβάνονται η πρωτοπαθής διάγνωση του ασθενούς, η λεπτομερής ιστοπαθολογία του όγκου, περαιτέρω φαινοτυπικά χαρακτηριστικά που συνάδουν με διάφορα καρκινικά σύνδρομα και πληροφορίες σχετικά με το οικογενειακό ιστορικό. Αξιοσημείωτο είναι πως αυτές οι πληροφορίες παρέχονται με τυποποιημένο και δομημένο τρόπο, επιτρέποντας την αποτελεσματική και προηγμένη ανάλυσή τους. Επιπλέον, η λεπτομέρεια με την οποία καταγράφονται καθώς και η ποικιλία τους επιτρέπουν την υποστήριξη μεταγενέστερων εφαρμογών, όπως τον προσδιορισμό της φαινοτυπικής ετερογένειας, τον ακριβή προσδιορισμό της διεισδυτικότητας, την ανάλυση συσχέτισης γονότυπου-φαινότυπου, την ιατρική ακριβείας και την εξέταση της παθογένειας των παραλλαγών.

Ένα χαρακτηριστικό παράδειγμα κατηγοριοποίησης παραλλαγής με τη βοήθεια των διαθέσιμων κλινικών δεδομένων αποτελεί η παραλλαγή NM_007294.3: c.245T> C (p.Leu82Pro) του γονιδίου *BRCA1.* Η συγκεκριμένη παραλλαγή είναι μία σπάνια παρανοηματική παραλλαγή η οποία βρέθηκε σε μια νεαρή γυναίκα που έχει διαγνωστεί δύο φορές με πρωτοπαθή τριπλά αρνητικό όγκο του μαστού (Εικόνα 43) – έναν φαινότυπο που συνάδει με την παρουσία παθογόνου παραλλαγής στο *BRCA1.* Αυτές οι πληροφορίες, σε συνδυασμό με τα αποτελέσματα *in silico* εργαλείων και τα υπάρχοντα δεδομένα από λειτουργικές μελέτες που δείχνουν μια σοβαρή επίπτωση της παραλλαγής στο γονίδιο (Findlay *et al.*, 2018), συνηγορούν στην κατηγοριοποίηση της παραλλαγής ως «πιθανώς παθογόνου».

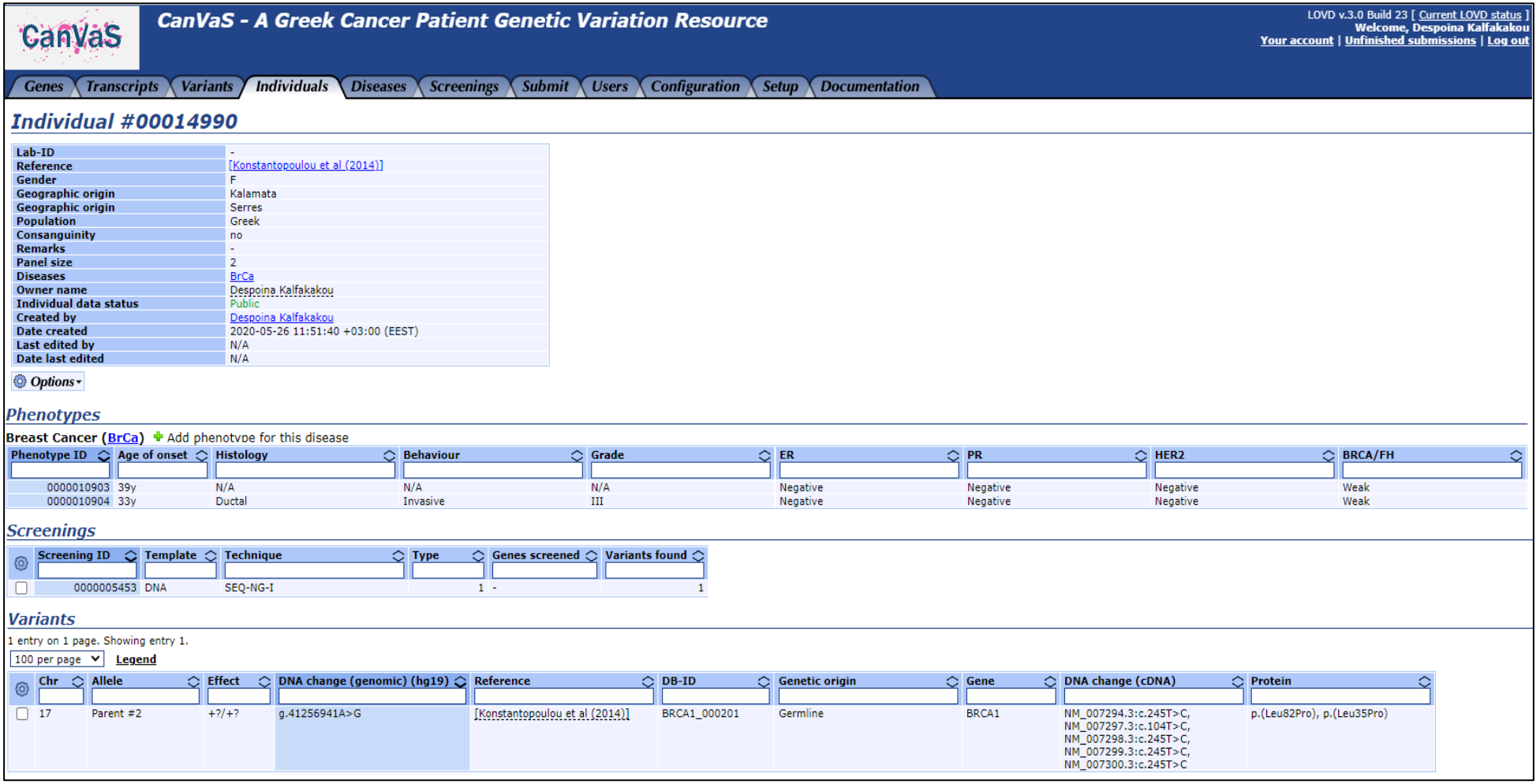
Στο παρελθόν, έχουν πραγματοποιηθεί πολλές φιλόδοξες προσπάθειες για την κατασκευή εθνικών βάσεων γενετικών δεδομένων, οι οποίες εστίαζαν σε κλινικά σχετικές παραλλαγές που έχουν βρεθεί σε γονίδια που προδιαθέτουν σε μια σειρά από γενετικές ασθένειες (Fakhro *et al.*, 2016; Charoute *et al.*, 2014; Pradhan *et al.*, 2011; Ruangrit *et al.*, 2008; van Baal *et al.*, 2007; Zlotogora *et al.*, 2007; Kleanthous *et al.*, 2006; Patrinos *et al.*, 2005; Sipila & Aula, 2002). Ανάμεσα σε αυτές περιλαμβάνεται και η Ελληνική Εθνική Βάση Δεδομένων Μεταλλάξεων (Hellenic National Mutation Database) (Patrinos *et al.*, 2005), η οποία καταγράφει παθογόνους παραλλαγές σε διάφορες γενετικές ασθένειες. Ωστόσο, η πλειονότητα αυτών των βάσεων δεν έχει ανανεωθεί, γεγονός που επισημαίνει τη δυσκολία που παρουσιάζει η συντήρηση αυτού του είδους πηγών δεδομένων. Η ανάγκη για εθνικές βάσεις γενετικών δεδομένων είναι ζωτικής σημασίας για τη μελέτη της γενετικής ετερογένειας διαφορετικών πληθυσμών. Οι βάσεις δεδομένων που εστιάζουν στην καταγραφή συγκεκριμένων ασθενειών/γονιδίων είναι πιο εύκολο να διαχειριστούν, ενώ μπορούν να υποστηρίξουν μεγαλύτερες βάσεις που καταγράφουν γενετικά δεδομένα ευρέος ενδιαφέροντος. Με αυτόν τον τρόπο, ενισχύουν τη συνεργασία και την ανταλλαγή δεδομένων εντός της επιστημονικής κοινότητας. Ένα πρώτο βήμα προς αυτήν την κατεύθυνση αποτελεί η εθνική βάση δεδομένων BRCA που εστιάζει συγκεκριμένα στον κληρονομικό καρκίνο, η οποία αποτελεί επιμέρους έργο της βραζιλιάνικης πρωτοβουλίας για την ιατρική ακριβείας (Rocha *et al.*, 2020). Στη βάση

δεδομένων BRCA καταγράφονται 180 παραλλαγές στα γονίδια *BRCA1* και *BRCA2* που έχουν βρεθεί στον πληθυσμό της Βραζιλίας.

Η ανάπτυξη της βάσης δεδομένων CanVaS στηρίχθηκε στο λογισμικό ανοιχτού κώδικα LOVD, γεγονός που επιτρέπει την εύκολη κοινή χρήση των δεδομένων, καθώς όλες οι παραλλαγές ενσωματώνονται αυτόματα στην κεντρική εγκατάσταση LOVD. Με αυτόν τον τρόπο, η βάση CanVaS χρησιμεύει ως μέρος της λύσης στη διαιρεμένη διαθεσιμότητα των γενετικών δεδομένων (den Dunnen, 2018). Ταυτόχρονα, αποτελεί ένα πρώτο παράδειγμα για άλλους ερευνητές που θέλουν να αναπτύξουν μια βάση δεδομένων για κάποιο συγκεκριμένο πληθυσμό και, ταυτόχρονα, να μοιράζονται εύκολα τα δεδομένα τους με τις ήδη καθιερωμένες πηγές γενετικών δεδομένων ευρέος ενδιαφέροντος.

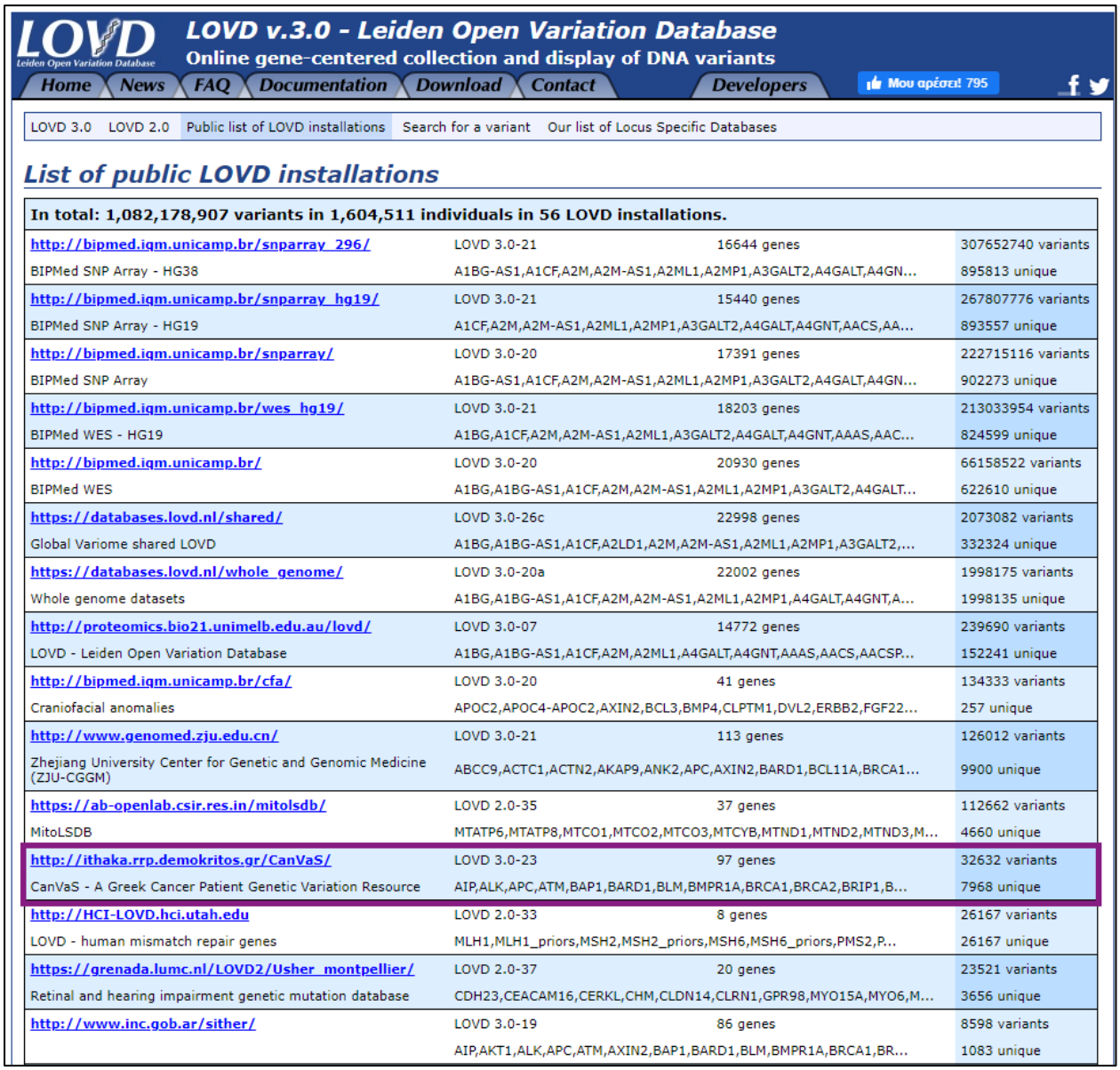
Από όσο γνωρίζουμε, η βάση δεδομένων CanVaS είναι η μεγαλύτερη βάση πληθυσμιακών δεδομένων για τον κληρονομικό καρκίνο και τη γενετική προδιάθεση στον καρκίνο, καταγράφοντας παραλλαγές σε 97 γονίδια σε συνδυασμό με κλινικά δεδομένα, ενώ είναι η μοναδική τέτοιου είδους βάση για τον Ελληνικό πληθυσμό. Επιπλέον, στη λίστα με τις δημόσιες εγκαταστάσεις LOVD, η βάση CanVaS είναι η δωδέκατη σε σειρά βάση με τις περισσότερες παραλλαγές (Εικόνα 44). Το γεγονός αυτό είναι πραγματικά εντυπωσιακό, αν αναλογιστεί κανείς πως οι παραλλαγές που καταγράφονται στη βάση CanVaS είναι μόνο οι σπάνιες παραλλαγές που έχουν εντοπιστεί σε αυτά τα 97 γονίδια προδιάθεσης στον καρκίνο.

Μέσα από τη λεπτομερή καταγραφή των συνδυασμένων γενετικών και φαινοτυπικών δεδομένων, η βάση CanVaS μπορεί να αποτελέσει μία πολύτιμη πηγή για επαγγελματίες που ασχολούνται με τη γενετική του καρκίνου. Η μακροπρόθεσμη φιλοδοξία μας είναι να οργανώσουμε μια συντονισμένη προσπάθεια μαζί με άλλα δημόσια και ιδιωτικά διαγνωστικά εργαστήρια προκειμένου να συγκεντρώσουμε τα γενετικά και φαινοτυπικά δεδομένα που συμβάλλουν στη γενετική προδιάθεση στον καρκίνο στον ελληνικό πληθυσμό, ώστε η CanVas να αποτελέσει Εθνική Βάση Αναφοράς για τον κληρονομικό καρκίνο.



Εικόνα 41: Φαινοτυπικά χαρακτηριστικά μιας γυναίκας που φέρει την παραλλαγή NM_007294.3:c.245T>C (p.Leu82Pro) στο γονίδιο *BRCA1*. Τροποποίηση από (Kalfakakou, Fostira*, et al.*, 2021).

Figure 43: Phenotypic traits of a woman carrying the *BRCA1* variant NM_007294.3:c.245T>C (p.Leu82Pro). Modification from (Kalfakakou, Fostira*, et al.*, 2021).



LOVD v.3.0 - Leiden Open Variation Database
Online gene-centered collection and display of DNA variants
Home | News | FAQ | Documentation | Download | Contact | Developers | Μου αρέσει! 795

LOVD 3.0 | LOVD 2.0 | Public list of LOVD installations | Search for a variant | Our list of Locus Specific Databases

List of public LOVD installations

In total: 1,082,178,907 variants in 1,604,511 individuals in 56 LOVD installations.			
http://bipmed.iqm.unicamp.br/snparray_296/	LOVD 3.0-21	16644 genes	307652740 variants
BIPMed SNP Array - HG38	A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GN...		895813 unique
http://bipmed.iqm.unicamp.br/snparray_hg19/	LOVD 3.0-21	15440 genes	267807776 variants
BIPMed SNP Array - HG19	A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GNT,AACS,AA...		893557 unique
http://bipmed.iqm.unicamp.br/snparray/	LOVD 3.0-20	17391 genes	222715116 variants
BIPMed SNP Array	A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GN...		902273 unique
http://bipmed.iqm.unicamp.br/wes_hg19/	LOVD 3.0-21	18203 genes	213033954 variants
BIPMed WES - HG19	A1BG,A1CF,A2M,A2M-AS1,A2ML1,A3GALT2,A4GALT,A4GNT,AAAS,AAC...		824599 unique
http://bipmed.iqm.unicamp.br/	LOVD 3.0-20	20930 genes	66158522 variants
BIPMed WES	A1BG,A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT...		622610 unique
https://databases.lovd.nl/shared/	LOVD 3.0-26c	22998 genes	2073082 variants
Global Variome shared LOVD	A1BG,A1BG-AS1,A1CF,A2LD1,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,...		332324 unique
https://databases.lovd.nl/whole_genome/	LOVD 3.0-20a	22002 genes	1998175 variants
Whole genome datasets	A1BG,A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A4GALT,A4GNT,A...		1998135 unique
http://proteomics.bio21.unimelb.edu.au/lovd/	LOVD 3.0-07	14772 genes	239690 variants
LOVD - Leiden Open Variation Database	A1BG,A1BG-AS1,A1CF,A2M,A2ML1,A4GALT,A4GNT,AAAS,AACS,AACSP...		152241 unique
http://bipmed.iqm.unicamp.br/cfa/	LOVD 3.0-20	41 genes	134333 variants
Craniofacial anomalies	APOC2,APOC4-APOC2,AXIN2,BCL3,BMP4,CLPTM1,DVL2,ERBB2,FGF22...		257 unique
http://www.genomed.zju.edu.cn/	LOVD 3.0-21	113 genes	126012 variants
Zhejiang University Center for Genetic and Genomic Medicine (ZJU-CGGM)	ABCC9,ACTC1,ACTN2,AKAP9,ANK2,APC,AXIN2,BARD1,BCL11A,BRCA1...		9900 unique
https://ab-openlab.csir.res.in/mitolsdb/	LOVD 2.0-35	37 genes	112662 variants
MitoLSDB	MTATP6,MTATP8,MTCO1,MTCO2,MTCO3,MTCYB,MTND1,MTND2,MTND3,M...		4660 unique
http://ithaka.rrp.demokritos.gr/CanVaS/	LOVD 3.0-23	97 genes	32632 variants
CanVaS - A Greek Cancer Patient Genetic Variation Resource	AIP,ALK,APC,ATM,BAP1,BARD1,BLM,BMPR1A,BRCA1,BRCA2,BRIP1,B...		7968 unique
http://HCI-LOVD.hci.utah.edu	LOVD 2.0-33	8 genes	26167 variants
LOVD - human mismatch repair genes	MLH1,MLH1_priors,MSH2,MSH2_priors,MSH6,MSH6_priors,PMS2,P...		26167 unique
https://grenada.lumc.nl/LOVD2/Usher_montpellier/	LOVD 2.0-37	20 genes	23521 variants
Retinal and hearing impairment genetic mutation database	CDH23,CEACAM16,CERKL,CHM,CLDN14,CLRN1,GPR98,MYO15A,MYO6,M...		3656 unique
http://www.inc.gob.ar/sither/	LOVD 3.0-19	86 genes	8598 variants
	AIP,AKT1,ALK,APC,ATM,AXIN2,BAP1,BARD1,BLM,BMPR1A,BRCA1,BR...		1083 unique

Εικόνα 42: Η λίστα με τις 15 πρώτες δημόσιες εγκαταστάσεις βάσεων γενετικών δεδομένων βασισμένων στο λογισμικό ανοιχτού κώδικα LOVD με τις περισσότερες παραλλαγές.

Figure 44: The 15 public genetic database LOVD installations with the most variants.

4.2. Ροή διοχέτευσης εντολών διεργασιών VarTrace

Η ροή VarTrace σχεδιάστηκε για την ανάλυση δεδομένων που προέρχονται από τη μαζική παράλληλη αλληλούχηση DNA όγκων. Η ανάλυση της γενετικής πληροφορίας των όγκων εμπεριέχει πολύ σημαντικές δυσκολίες, λόγω της ετερογένειας του γονιδιώματος των όγκων, την περιεκτικότητα φυσιολογικών κυττάρων στα δείγματα του όγκου καθώς και των τεχνικών διατήρησης του ιστού. Επομένως, η ακριβής ανίχνευση των πραγματικών παραλλαγών είναι μία πολύπλοκη διαδικασία.

Το προγραμματιστικό εργαλείο VarTrace είναι μία πλήρως διαμορφώσιμη ροή διοχέτευσης εντολών διεργασιών για την ανάλυση δεδομένων αλληλούχησης DNA όγκου και την ανίχνευση σωματικών

παραλλαγών. Το πλεονέκτημα της ροής VarTrace είναι πως συνδυάζει διαφορετικά εργαλεία που χρησιμοποιούνται κατά κόρον κατά τη διαδικασία της βιοπληροφορικής ανάλυσης δεδομένων τέτοιου τύπου και τα εκτελεί αυτόματα. Με αυτόν τον τρόπο, διευκολύνεται η ίδια η διαδικασία της ανάλυσης δεδομένων, ενώ προάγεται η επαναληψιμότητα και η αναπαραγωγιμότητα του πειράματος. Παράλληλα, η χρήση των πολλαπλών αλγορίθμων κλήσης παραλλαγών μειώνει το ποσοστό των ψευδώς αρνητικών αποτελεσμάτων.

Η ροή VarTrace αφαιρεί τους προσαρμογείς των αναγνώσεων DNA και αντιστοιχεί τις ακολουθίες στο ανθρώπινο γονιδίωμα αναφοράς χρησιμοποιώντας το εργαλείο BWA-MEM. Η κλήση των παραλλαγών πραγματοποιείται με τα εργαλεία Mutect2, VarScan2 και Octopus, ενώ ως έξοδο ο χρήστης παίρνει την ένωση των συνόλων των παραλλαγών εξόδου των επιμέρους εργαλείων, ύστερα από εφαρμογή μιας σειράς φίλτρων βάσει της ποιότητάς τους. Η ροή μπορεί να αναλύσει δεδομένα που προέρχονται από αλληλούχηση DNA μονού ή ζεύγους άκρων, ενώ υπάρχει η επιλογή της ανάλυσης με χρήση UMI. Το πιο σημαντικό είναι πως η ροή είναι σχεδιασμένη ώστε να μπορεί να εκτελεστεί παράλληλα, γεγονός που μειώνει το χρόνο ανάλυσης κατά πολύ.

Είναι αξιοσημείωτο πως η ροή VarTrace έχει ως έξοδο, εκτός από τα βασικά αρχεία που περιλαμβάνουν βιολογικές πληροφορίες (BAM και VCF), επιπλέον δύο αρχεία για τον έλεγχο της ποιότητας των δεδομένων και ένα αρχείο με τις παραλλαγές που ανιχνεύτηκαν εμπλουτισμένες με πληροφορίες. Στην πρώτη περίπτωση, η αναφορά ποιότητας των αναγνώσεων DNA που εξάγεται από το εργαλείο FastQC περιλαμβάνει γραφήματα που παρουσιάζουν λεπτομερώς τα ποιοτικά στοιχεία των αναγνώσεων, ενώ το αρχείο ελέγχου τους βάθους ανάγνωσης επιτρέπει στον χρήστη να οπτικοποιήσει τις περιοχές με χαμηλό βάθος ανάγνωσης, ώστε να μπορέσει να τις εξετάσει μεμονωμένα και με ακρίβεια. Το τελικό αρχείο εξόδου με τις εμπλουτισμένες με πληροφορίες παραλλαγές έχει υποστεί επεξεργασία ώστε να περιλαμβάνει τις παραλλαγές που είναι πιο πιθανό να είναι πραγματικές, ενώ η μορφή του επιτρέπει την ανάλυσή του με απλά στη χρήση εργαλεία, όπως το Excel, γεγονός που επιτρέπει τη χρήση του ακόμα και από χρήστες που δε διαθέτουν εξειδικευμένες υπολογιστικές γνώσεις. Επιπλέον, το αρχείο ελέγχου του βάθους ανάγνωσης και το αρχείο εξόδου που περιλαμβάνει τις εμπλουτισμένες με πληροφορίες παραλλαγές παράγονται από προσαρμοσμένα σενάρια που αναπτύχθηκαν στο πλαίσιο της κατασκευής του VarTrace. Τα σενάρια αυτά είναι εύκολο να παραμετροποιηθούν και μπορούν να χρησιμοποιηθούν ως αυτοτελή εργαλεία σε άλλες ροές διοχέτευσης εντολών διεργασιών.

Για την αξιολόγηση της απόδοσης και της χρησιμότητάς της, η ροή VarTrace εφαρμόστηκε σε δεδομένα που προέρχονται από την αλληλούχηση των γονιδίων *BRCA1* και *BRCA2* σε 75 όγκους των ωοθηκών. Παράλληλα, τα ίδια δεδομένα αναλύθηκαν με δύο ροές διοχέτευσης εντολών διεργασιών του εμπορικά διαθέσιμου λογισμικού Biomedical Genomics Workbench (BGWB), και πιο συγκεκριμένα, την προεπιλεγμένη ροή (BGWBdef) και μία ροή που πραγματοποιεί την ανάλυση χρησιμοποιώντας τους UMI (BGWBumi). Όλες οι ανιχνευμένες παθογόνοι παραλλαγές επιβεβαιώθηκαν με τη μέθοδο Sanger.

Οι τρεις ροές διοχέτευσης εντολών διεργασιών ανίχνευσαν συνολικά 88 παθογόνους ή πιθανώς παθογόνους παραλλαγές που πληρούσαν τα κριτήρια ποιότητας, ενώ μόλις το 14,8% αυτών ανιχνεύτηκε και από τις τρεις. Σε αντίστοιχες μελέτες, οι οποίες όμως συνέκριναν μόνο αλγόριθμους κλήσης

παραλλαγών και λάμβαναν υπόψη όλες τις παραλλαγές και όχι μόνο όσες πληρούν κάποια κριτήρια, το ποσοστό συμφωνίας μεταξύ διαφορετικών αλγορίθμων ήταν 3% (Kroigard *et al.*, 2016) και 0.02%-0.5% (Q. Wang *et al.*, 2019). Μία σημαντική διαφορά της δικής μας αξιολόγησης σε σχέση με τη σύγκριση που πραγματοποίησαν οι δύο αυτές μελέτες είναι πως οι (Kroigard *et al.*, 2016) εφάρμοσαν φίλτρα σε κάποιους μόνο από αλγορίθμους κλήσεων παραλλαγών, ενώ οι (Q. Wang *et al.*, 2019) δεν εφάρμοσαν κανένα φίλτρο για τον έλεγχο της ποιότητας των παραλλαγών. Φαίνεται λοιπόν, πως η εφαρμογή φίλτρων μειώνει σημαντικά το ποσοστό των ψευδώς θετικών αποτελεσμάτων, κάτι που συμπεραίνουν και οι δύο αυτές μελέτες. Ωστόσο, θα πρέπει να πραγματοποιείται συντηρητικά, καθώς διατρέχεται ο κίνδυνος να απαλειφθούν και πραγματικές παραλλαγές οι οποίες έχουν ανιχνευτεί με μικρή συχνότητα αλληλομόρφου.

Από τις 88 παραλλαγές που ανιχνεύτηκαν συνολικά, μόνο οι 13 παραλλαγές επιβεβαιώθηκαν τελικά με Sanger. Ενώ το BGWBumi είχε το μικρότερο ποσοστό ψευδώς θετικών αποτελεσμάτων, δεν κατάφερε να ανιχνεύσει όλες τις πραγματικές παθογόνους παραλλαγές, γεγονός που έχει ιδιαίτερη βαρύτητα, δεδομένου του ότι ο συγκεκριμένος γενετικός έλεγχος στοχεύει στην ανίχνευση παθογόνων παραλλαγών με σκοπό τη χορήγηση εξατομικευμένης θεραπείας. Από την άλλη, παρόλο που το BGWBdef ανίχνευσε όλες τις παραλλαγές, είχε ένα πολύ μεγάλο ποσοστό ψευδώς θετικών ευρημάτων, η αξιολόγηση των οποίων είναι αρκετά χρονοβόρα και κοστοβόρα. Τελικά, η ροή VarTrace είχε την καλύτερη απόδοση, μιας και κατάφερε να εντοπίσει όλες τις παραλλαγές, ενώ το ποσοστό των ψευδώς θετικών ευρημάτων ήταν σημαντικά μικρότερο. (ακρίβεια=1, ανάκληση=0,5).

Δεν υπάρχουν πολλές διαθέσιμες ροές εντολών διεργασιών ανοιχτού κώδικα οι οποίες πραγματοποιούν όλα τα βήματα της ανάλυσης για την ανίχνευση παραλλαγών σε DNA όγκου. Στο παρελθόν, έχουν προκύψει δύο δημοσιεύσεις με την περιγραφή τέτοιων ροών (Liu *et al.*, 2017; do Valle *et al.*, 2016), της οποίες χρησιμοποιούνται τα BWA-mem, Picard, GATK και Mutect (την προηγούμενη έκδοση του εργαλείου Mutect2), ενώ οι (Liu *et al.*, 2017) χρησιμοποιούν επιπλέον το VarScan (την προηγούμενη έκδοση του VarScan2) για την κλήση των παραλλαγών. Ωστόσο, οι συγκεκριμένες ροές έχουν απλά περιγραφεί και αξιολογηθεί, χωρίς να παρέχεται ο κώδικάς της για χρήση, γεγονός που καθιστά τη σύγκριση με το VarTrace και την αναπαραγωγιμότητα των αποτελεσμάτων τους αδύνατη. Πρόσφατα δημοσιεύτηκε μια πιο ολοκληρωμένη μελέτη, η οποία περιγράφει την ροή iWhale (Binatti *et al.*, 2021), της οποίας ο κώδικας είναι ανοιχτός, ενώ έχει σχεδιαστεί ώστε να είναι συμβατή και να μπορεί να εκτελείται σε όλα τα λειτουργικά συστήματα. Ωστόσο, στη δημοσίευση της iWhale πραγματοποιείται ολοκληρωμένη αξιολόγηση της ροής μόνο σε προσομοιωμένα δεδομένα τα οποία δεν είναι εύκολο να μοντελοποιήσουν την πολυπλοκότητα των πραγματικών δεδομένων που έχουν παραχθεί από την αλληλούχηση DNA όγκου, ενώ κατά την αξιολόγηση με πραγματικά δεδομένα που έχουν παραχθεί με WES μελετήθηκε μόνο μία παραλλαγή που είχε επιβεβαιωθεί με ανεξάρτητη πειραματική μέθοδο.

Ένας σημαντικός περιορισμός που έχει η παρούσα μελέτη είναι πως δεν αξιολογήθηκε η ανάλυση των δεδομένων με τη χρήση των UMI, παρόλο που η ροή VarTrace υποστηρίζει αυτή την επιλογή, καθώς από της πρώτες δοκιμές του VarTrace με τη χρήση UMI φαινόταν πως της πραγματικές παραλλαγές εξαλείφονταν ως τεχνουργήματα. Επομένως, ένα επόμενο βήμα για τη βελτίωση της ροής διοχέτευσης

εντολών διεργασιών VarTrace, είναι η παραμετροποίησή της ώστε να μπορεί να υποστηρίζει κι αυτή τη λειτουργία χωρίς σφάλματα.

Τέλος, παρόλο που η ροή VarTrace είναι σχεδιασμένη ώστε να μπορεί να εκτελείται σε οποιοδήποτε υπολογιστή, στην παρούσα μορφή της είναι απαραίτητο να έχει πραγματοποιηθεί πρότερη εγκατάσταση των επιμέρους εργαλείων από το χρήστη. Γι' αυτόν το λόγο, έχει ξεκινήσει ήδη η συγγραφή της ροής σε γλώσσα CWL. Αυτή η αναβάθμιση θα επιτρέπει την εκτέλεση της ροής VarTrace σε οποιαδήποτε πλατφόρμα χωρίς να χρειάζεται η πρότερη ρύθμιση των παραμέτρων και των εργαλείων από το χρήστη. Έπειτα από τις τελευταίες προσθήκες, και δεδομένου ότι ήδη αποτελεί ένα ισχυρό εργαλείο για την ακριβή ανίχνευση σωματικών παραλλαγών βάσει και των αποτελεσμάτων της σύγκρισης με τις εμπορικά διαθέσιμες ροές, η ροή VarTrace θα είναι πλέον ακόμα πιο εύκολη στη χρήση, ενώ θα μπορεί να αποτελέσει τη βάση για την ανάπτυξη μιας εμπορικά διαθέσιμης εφαρμογής.

4.3. Αξιολόγηση εμπορικού λογισμικού για τον εμπλουτισμό παραλλαγών με πληροφορίες

Ο σκοπός της παρούσας μελέτης είναι η ανάδειξη των παγίδων που κρύβονται στη διαδικασία της βιοπληροφορικής ανάλυσης των δεδομένων Αλληλούχησης Επόμενης Γενιάς, οι οποίες μπορούν εύκολα να παραβλεφθούν και τελικά να οδηγήσουν σε παραπλανητικά αποτελέσματα. Πιο συγκεκριμένα, η μελέτη επικεντρώνεται στην αξιολόγηση του VariantStudio (VS), της εμπορικά διαθέσιμου λογισμικού το οποίο χρησιμοποιείται για τον εμπλουτισμό παραλλαγών με πληροφορίες. Το συγκεκριμένο λογισμικό είναι και κατά κάποιον τρόπο το κυρίαρχο στον τομέα αυτό, καθώς συνοδεύει την επικρατούσα πλέον πλατφόρμα Αλληλούχησης Επόμενης Γενιάς, της Illumina. Για την αξιολόγηση του λογισμικού, συγκρίθηκε ο χαρακτηρισμός των παραλλαγών ως της την επίδρασή της στην ακολουθία του γονιδίου που αποκτήθηκε χρησιμοποιώντας τα προεπιλεγμένα μετάγραφα του VS (VS transcript set – VSts), με το χαρακτηρισμό των παραλλαγών που αποκτήθηκε χρησιμοποιώντας ένα επιλεγμένο σύνολο μεταγράφων (Selected transcript set – Sts). Ο εμπλουτισμός των παραλλαγών με πληροφορίες με το Sts πραγματοποιήθηκε με τη χρήση του εργαλείου VEP της Ensembl. Το VS έχει μια προεπιλεγμένη επιλογή μεταγράφων, τα οποία της δεν μπορούν να τροποποιηθούν από τον χρήστη. Για το Sts επιλέχθηκαν χειροκίνητα τα πρωτεύοντα μετάγραφα για κάθε γονίδιο.

Η σύγκριση ανέδειξε μία σειρά από ασυμφωνίες μεταξύ των χαρακτηρισμών, οι οποίες μπορούν να αποδοθούν στην ακατάλληλη επιλογή μεταγράφων από το VS. Πιο συγκεκριμένα, το ποσοστό συμφωνίας επί του συνολικού αριθμού των χαρακτηρισμών των παραλλαγών ήταν μόλις 82,82%. Επιπλέον, το 11% των παραλλαγών απώλειας λειτουργίας της πρωτεΐνης και το 9% των παρανοηματικών παραλλαγών χαρακτηρίστηκαν ως παραλλαγές σε μη κωδικές περιοχές από το VS. Η πιθανότητα οι παραλλαγές αυτών των κατηγοριών να είναι κλινικά σημαντικές είναι μεγάλη, γεγονός που αναδεικνύει την έκταση του προβλήματος του εσφαλμένου χαρακτηρισμού των παραλλαγών.

Επομένως, ένα σημαντικό συμπέρασμα της της σύγκρισης είναι η προσοχή η οποία πρέπει να δοθεί σε ένα στάδιο της ανάλυσης δεδομένων Αλληλούχησης Επόμενης Γενιάς που θα μπορούσε εύκολα να παραβλεφθεί· την επιλογή των κατάλληλων μεταγράφων για τον εμπλουτισμό των παραλλαγών με πληροφορίες. Η σημασία της επιλογής των κατάλληλων μεταγράφων έχει αναδειχθεί εκτενώς στο

παρελθόν. Πιο συγκεκριμένα, οι (McCarthy *et al.*, 2014), συνέκριναν τα εργαλεία εμπλουτισμού παραλλαγών ανοιχτού κώδικα VEP και ANNOVAR (K. Wang *et al.*, 2010), καθώς και τα σύνολα μεταγράφων RefSeq (Pruitt *et al.*, 2012) και Ensembl (Flicek *et al.*, 2012). Στη μελέτη της, σε μια άμεση σύγκριση χρησιμοποιώντας τα ίδια μετάγραφα, το ακριβές ποσοστό αντιστοίχισης ήταν 86,5%, ενώ στην περίπτωση της σύγκρισης μεταξύ των συνόλων μεταγράφων RefSeq και Ensembl, το ακριβές ποσοστό αντιστοίχισης ήταν 85%. Σε μια μεταγενέστερη μελέτη, από της (Frankish *et al.*, 2015), όπου συγκρίθηκαν τα σύνολα γονιδίων GENCODE (Harrow *et al.*, 2012) και RefSeq, φάνηκε ότι υπήρχε ασυμφωνία μεταξύ περίπου του 30% των παραλλαγών απώλειας λειτουργίας.

Έχοντας υπόψη αυτές της μελέτες, δεν θα περίμενε κανείς τον ίδιο χαρακτηρισμό των παραλλαγών χρησιμοποιώντας διαφορετικά μετάγραφα. Ωστόσο, με τη σύγκριση αυτή επισημαίνεται η «δυσλειτουργία» της εμπορικά διαθέσιμου λογισμικού εμπλουτισμού παραλλαγών. Αυτή η «δυσλειτουργία» παίρνει άλλη διάσταση όταν η ανάλυση αφορά γονιδιακούς ελέγχους ασθενών. Εν προκειμένω, η συμπεριφορά του VS είναι απλώς της περιορισμός του λογισμικού που μόνο η Illumina μπορεί να διορθώσει. Παρ' όλα αυτά, έχοντας κατά νου το συγκεκριμένο παράδειγμα, σκοπός της μελέτης είναι η ανάδειξη της σημασίας της εις βάθος κατανόησης κάθε βήματος της ανάλυσης και των παραμέτρων, καθώς και της κατάλληλης εκπαίδευσης και κατάρτισης των ατόμων που συμμετέχουν στην ανάλυση δεδομένων Αλληλούχησης Επόμενης Γενιάς. Η επιλογή του λογισμικού που θα χρησιμοποιηθεί για ανάλυση προϋποθέτει ότι ο χρήστης θα μπορεί να αμφισβητήσει τον αλγόριθμο του λογισμικού κάνοντας σημαντικές παρατηρήσεις.

Είναι αξιοσημείωτο ότι η προεπιλεγμένη επιλογή μεταγράφων του VS άλλαξε έπειτα από αναβάθμιση του λογισμικού, από την έκδοση v2.0 στην v3.0. Η έκδοση v2.0 χαρακτήριζε της παραλλαγές χρησιμοποιώντας το μετάγραφο το οποίο υπόκειται την πιο σοβαρή συνέπεια. Επιπλέον, τον Αύγουστο του 2017 η Illumina® κυκλοφόρησε ένα ακόμα λογισμικό εμπλουτισμού πληροφοριών, το Variant Interpreter®, το οποίο ύστερα από δοκιμή, φαίνεται να πραγματοποιεί το χαρακτηρισμό των παραλλαγών χωρίς σφάλματα. Ωστόσο, το VS εξακολουθεί να είναι διαθέσιμο και συνιστάται από την εταιρεία για χρήση με τα αντίστοιχα αντιδραστήρια για αλληλούχηση DNA.

Στη μελέτη της, αν η ανάλυση πραγματοποιείτο αποκλειστικά με χρήση του λογισμικού VS v3.0, θα μπορούσε να οδηγήσει στην απώλεια κλινικά σημαντικών παραλλαγών. Δύο σημαντικά παραδείγματα είναι ο εσφαλμένος σχολιασμός παραλλαγών απώλειας λειτουργίας στα γονίδια *BRCA2* και *MLH1*, τα οποία σχετίζονται με αυξημένο διά βίου κίνδυνο ανάπτυξης πολλαπλών καρκίνων. Επιπλέον, τα άτομα που φέρουν παθογόνους παραλλαγές σε αυτά τα γονίδια, μπορούν να επωφεληθούν σημαντικά από τη χρήση στοχευμένων θεραπειών, της αναστολέων PARP (άτομα που φέρουν παθογόνο παραλλαγή στα *BRCA1* και *BRCA2*) και ανοσοθεραπείας (άτομα με σύνδρομο Lynch), ενώ μπορούν να συμμετέχουν σε πρωτόκολλα εξατομικευμένης παρακολούθησης.

Συμπερασματικά, στην παρούσα μελέτη αναδεικνύεται ένα ευάλωτο σημείο της διαδικασίας εμπλουτισμού παραλλαγών με πληροφορίες, που αν δε του δοθεί η δέουσα προσοχή μπορεί να οδηγήσει σε εσφαλμένα αποτελέσματα. Στην περίπτωση που το λογισμικό χρησιμοποιείται σε κλινικά δεδομένα, ο λανθασμένος χαρακτηρισμός των παραλλαγών μπορεί να έχει άμεσο αντίκτυπο στην κλινική διαχείριση των ασθενών. Μία καλή πρακτική για την αποφυγή τέτοιων φαινομένων θα μπορούσε να

είναι η διασταύρωση των αποτελεσμάτων με ένα ανεξάρτητο λογισμικό. Αυτό προκύπτει και από την ίδια τη μελέτη, καθώς αν δεν είχε πραγματοποιηθεί ο χαρακτηρισμός των παραλλαγών με το επιλεγμένο σύνολο δεδομένων και με το λογισμικό VEP, δε θα είχαν ανιχνευθεί τα σφάλματα κατά τη χρήση του VS. Κατ' αυτόν τον τρόπο, αν καθιερωθεί αυτή η πρακτική και από άλλα διαγνωστικά εργαστήρια, θα μπορούσε να μειωθεί το ποσοστό σφάλματος των γονιδιακών ελέγχων και κατ' επέκταση, να βελτιωθεί η φροντίδα των ασθενών.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Χάρη στις ραγδαίες εξελίξεις στους τομείς της γενετικής και της βιοτεχνολογίας, και στις λύσεις που έχει προσφέρει στα προβλήματα που έχουν εισάγει αυτές οι εξελίξεις, η επιστήμη της βιοπληροφορικής έχει εδραιώσει τη θέση της στην ιατρική ακριβείας. Στην παρούσα διατριβή μελετώνται και παρουσιάζονται κάποιες μόνο εφαρμογές της βιοπληροφορικής στη γενετική του καρκίνου.

Η βάση δεδομένων CanVaS αποτελεί την πρώτη πηγή γενετικών δεδομένων Ελλήνων ασθενών με καρκίνο. Τα δεδομένα της CanVaS προέρχονται από την ανάλυση του DNA 7.363 ατόμων. Παράλληλα, καταγράφοντας 22.089 σπάνιες παραλλαγές -οι οποίες αντιστοιχούν σε 7.968 μοναδικές παραλλαγές- σε 97 γονίδια τα οποία είναι γνωστά ή ύποπτα για προδιάθεση σε κάποιον καρκίνο ή καρκινικό σύνδρομο, η CanVaS αποτελεί τη μεγαλύτερη βάση πληθυσμιακών δεδομένων για τον κληρονομικό καρκίνο και τη γενετική προδιάθεση στον καρκίνο, απ' όσο γνωρίζουμε. Ένα επίσης μοναδικό χαρακτηριστικό της βάσης είναι η καταγραφή των λεπτομερών κλινικών χαρακτηριστικών των ατόμων που φέρουν τις παραλλαγές, η οποία επιτρέπει την προηγμένη ανάλυση των δεδομένων της βάσης. Το γεγονός ότι η CanVaS έχει δημιουργηθεί με τη χρήση του λογισμικού ανοιχτού κώδικα LOVD επιτρέπει την άμεση διασύνδεσή της με τις καθιερωμένες κεντρικές πηγές δεδομένων και την ενσωμάτωσή της σε μεγαλύτερες βάσεις γενετικών δεδομένων ευρέως ενδιαφέροντος. Έτσι, η CanVaS μπορεί να αποτελέσει το πρώτο βήμα για την ολοκλήρωση ενός πιο μακροπρόθεσμου στόχου για μια συντονισμένη προσπάθεια για τη σύνοψη και την αποσαφήνιση του ελληνικού φάσματος των παραλλαγών στον καρκίνο, σε συνεργασία με τα νοσοκομεία και τα διαγνωστικά εργαστήρια της χώρας που δραστηριοποιούνται στη γενετική διάγνωση των κληρονομικών μορφών του καρκίνου, με απώτερο σκοπό τη χρήση τους για τη θεμελίωση των κλινικών κατευθυντήριων γραμμών για τους Έλληνες ασθενείς με καρκίνο.

Η ροή διοχέτευσης εντολών διεργασιών VarTrace είναι ένα ολοκληρωμένο και πλήρως διαμορφώσιμο υπολογιστικό εργαλείο ανοιχτού κώδικα για την ανάλυση δεδομένων προερχόμενων από τον προσδιορισμό της αλληλουχίας του DNA όγκου με τη μέθοδο Αλληλούχησης Επόμενης Γενιάς. Χάρη στην εξειδικευμένη παραμετροποίησή του και τη χρήση πολλαπλών αλγορίθμων κλήσης παραλλαγών, το VarTrace καταφέρνει να ανιχνεύσει με ακρίβεια τις παραλλαγές σωματικής σειράς, ενώ παράλληλα περιορίζει την κλήση ψευδώς θετικών ευρημάτων. Το συμπέρασμα αυτό επιβεβαιώνεται και από τη σύγκριση του VarTrace με δύο εμπορικά διαθέσιμες ροές διοχέτευσης εντολών διεργασιών. Το VarTrace παρέχει σημαντικά αρχεία για την αξιολόγηση της ποιότητας του πειράματος Αλληλούχησης Επόμενης Γενιάς, ενώ το τελικό αρχείο με τον εμπλουτισμό των παραλλαγών με πληροφορίες δίνεται σε μορφή φιλική προς το χρήστη. Η αναβάθμιση του VarTrace ώστε να είναι συμβατό με όλα τα λειτουργικά συστήματα θα επιτρέπει την εκτέλεση της ροής σε οποιαδήποτε πλατφόρμα χωρίς να χρειάζεται η πρότερη ρύθμιση παραμέτρων και η εγκατάσταση των επιμέρους εργαλείων από το χρήστη.

Η αξιολόγηση του εμπορικού λογισμικού VS για τον εμπλουτισμό των παραλλαγών με πληροφορίες ανέδειξε κάποιες παγίδες που κρύβονται στη βιοπληροφορική ανάλυση των γενετικών δεδομένων, ειδικά όταν έχει εφαρμογή στην κλινική πράξη. Πιο συγκεκριμένα, σημειώθηκε η σημασία της προσεκτικής επιλογής του λογισμικού που χρησιμοποιείται κατά την ανάλυση, καθώς και της εμπειρίας και εξειδίκευσης που πρέπει να έχει ο χρήστης ώστε να μπορεί να αξιολογήσει την απόδοσή του.

Επιπλέον, ένα σημαντικό συμπέρασμα της αξιολόγησης είναι η προσοχή η οποία πρέπει να δοθεί σε ένα στάδιο της ανάλυσης δεδομένων Αλληλούχησης Επόμενης Γενιάς που θα μπορούσε εύκολα να παραβλεφθεί· την επιλογή των κατάλληλων μεταγράφων για τον εμπλουτισμό των παραλλαγών με πληροφορίες. Ο χαρακτηρισμός των παραλλαγών με το σύνολο μεταγράφων του VS λειτούργησε λανθασμένα ακόμα και για παραλλαγές απώλειας λειτουργίας οι οποίες έχουν σημαντική κλινική χρησιμότητα. Πρακτικές όπως η διασταύρωση των αποτελεσμάτων με ένα ανεξάρτητο λογισμικό θα μπορούσαν να συνεισφέρουν στον περιορισμό των σφαλμάτων κατά τον γονιδιακό έλεγχο, γεγονός που θα έχει άμεσο θετικό αντίκτυπο στην κλινική διαχείριση των ασθενών και των ατόμων υψηλού κινδύνου.

6. ΒΙΒΛΙΟΓΡΑΦΙΑ

Abdelwahab Yousef, A. J. (2017). Male Breast Cancer: Epidemiology and Risk Factors. *Semin Oncol, 44*(4), 267-272. doi:10.1053/j.seminoncol.2017.11.002

Abraham, B. J., Hnisz, D., Weintraub, A. S., Kwiatkowski, N., Li, C. H., Li, Z., . . . Young, R. A. (2017). Small genomic insertions form enhancers that misregulate oncogenes. *Nat Commun, 8*, 14385. doi:10.1038/ncomms14385

ACOG, T. A. C. o. O. a. G., SGO, T. S. o. G. O., Modesitt, S. C., Lu, K., Chen, L., & Powell, C. B. (2017). Practice Bulletin No 182: Hereditary Breast and Ovarian Cancer Syndrome. *Obstet Gynecol, 130*(3), e110-e126. doi:10.1097/AOG.0000000000002296

Anagnostopoulos, T., Pertesi, M., Konstantopoulou, I., Armaou, S., Kamakari, S., Nasioulas, G., . . . Yannoukakos, D. (2008). G1738R is a BRCA1 founder mutation in Greek breast/ovarian cancer patients: evaluation of its pathogenicity and inferences on its genealogical history. *Breast Cancer Res Treat, 110*(2), 377-385. doi:10.1007/s10549-007-9729-y

Andrews, S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Babraham Institute. Cambridge, UK.

Antoniou, A., Pharoah, P. D., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., . . . Easton, D. F. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet, 72*(5), 1117-1130. doi:10.1086/375033

Antoniou, A. C., Casadei, S., Heikkinen, T., Barrowdale, D., Pylkas, K., Roberts, J., . . . Tischkowitz, M. (2014). Breast-cancer risk in families with mutations in PALB2. *N Engl J Med, 371*(6), 497-506. doi:10.1056/NEJMoa1400382

Apostolou, P., & Fostira, F. (2013). Hereditary breast cancer: the era of new susceptibility genes. *Biomed Res Int, 2013*, 747318. doi:10.1155/2013/747318

Apostolou, P., Fostira, F., Kouroussis, C., Kalfakakou, D., Delimitsou, A., Agelaki, S., . . . Saloustros, E. (2020). BRCA1 and BRCA2 germline testing in Cretan isolates reveals novel and strong founder effects. *Int J Cancer, 147*(5), 1334-1342. doi:10.1002/ijc.32903

Apostolou, P., Pertesi, M., Aleporou-Marinou, V., Dimitrakakis, C., Papadimitriou, C., Razis, E., . . . Fostira, F. (2017). Haplotype analysis reveals that the recurrent BRCA1 deletion of exons 23 and 24 is a Greek founder mutation. *Clin Genet, 91*(3), 482-487. doi:10.1111/cge.12824

Apostolou, P., Vratimos, A., & Fostira, F. (2015). *Breast Cancer Susceptibility Genes*: Wiley-Blackwell.

Armaou, S., Pertesi, M., Fostira, F., Thodi, G., Athanasopoulos, P. S., Kamakari, S., . . . Konstantopoulou, I. (2009). Contribution of BRCA1 germ-line mutations to breast cancer in Greece: a hospital-based study of 987 unselected breast cancer cases. *Br J Cancer, 101*(1), 32-37. doi:10.1038/sj.bjc.6605115

Arreaza, G., Qiu, P., Pang, L., Albright, A., Hong, L. Z., Marton, M. J., & Levitan, D. (2016). Pre-Analytical Considerations for Successful Next-Generation Sequencing (NGS): Challenges and Opportunities for Formalin-Fixed and Paraffin-Embedded Tumor Tissue (FFPE) Samples. *Int J Mol Sci, 17*(9). doi:10.3390/ijms17091579

Ashley, E. A. (2016). Towards precision medicine. *Nat Rev Genet, 17*(9), 507-522. doi:10.1038/nrg.2016.86

Bai, G., & Lipton, S. A. (1998). Aberrant RNA splicing in sporadic amyotrophic lateral sclerosis. *Neuron, 20*(3), 363-366. doi:10.1016/s0896-6273(00)80979-4

Barba, M., Czosnek, H., & Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses, 6*(1), 106-136. doi:10.3390/v6010106

Barbitoff, Y. A., Polev, D. E., Glotov, A. S., Serebryakova, E. A., Shcherbakova, I. V., Kiselev, A. M., . . . Predeus, A. V. (2020). Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. *Sci Rep, 10*(1), 2057. doi:10.1038/s41598-020-59026-y

Barillot, E., Calzone, L., Hupé, P., Vert, J. P., & Zinovyev, A. (2012). *Computational systems biology of cancer*: CRC Press.
Bartsch, D. K., Slater, E. P., Carrato, A., Ibrahim, I. S., Guillen-Ponce, C., Vasen, H. F., . . . Gress, T. M. (2016). Refinement of screening for familial pancreatic cancer. *Gut, 65*(8), 1314-1321. doi:10.1136/gutjnl-2015-311098
Beard, W. A., Horton, J. K., Prasad, R., & Wilson, S. H. (2019). Eukaryotic Base Excision Repair: New Approaches Shine Light on Mechanism. *Annu Rev Biochem, 88*, 137-162. doi:10.1146/annurev-biochem-013118-111315
Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed, 98*(6), 236-238. doi:10.1136/archdischild-2013-304340
Benjamin, D., Sato, T., Cibulskis, K., Getz, G., Stewart, C., & Lichtenstein, L. (2019). *Calling Somatic SNVs and Indels with Mutect2*. bioRxiV.
Bentley, A. R., Callier, S., & Rotimi, C. N. (2017). Diversity and inclusion in genomic research: why the uneven progress? *J Community Genet, 8*(4), 255-266. doi:10.1007/s12687-017-0316-6
Berger, A. H., Knudson, A. G., & Pandolfi, P. P. (2011). A continuum model for tumour suppression. *Nature, 476*(7359), 163-169. doi:10.1038/nature10275
Binatti, A., Bresolin, S., Bortoluzzi, S., & Coppe, A. (2021). iWhale: a computational pipeline based on Docker and SCons for detection and annotation of somatic variants in cancer WES data. *Brief Bioinform, 22*(3). doi:10.1093/bib/bbaa065
Borecka, M., Zemankova, P., Vocka, M., Soucek, P., Soukupova, J., Kleiblova, P., . . . Janatova, M. (2016). Mutation analysis of the PALB2 gene in unselected pancreatic cancer patients in the Czech Republic. *Cancer Genet, 209*(5), 199-204. doi:10.1016/j.cancergen.2016.03.003
BroadInstitute. (2019). Picard toolkit. *Broad Institute, GitHub repository.* Retrieved from http://broadinstitute.github.io/picard/.
Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., . . . Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res, 47*(D1), D1005-D1012. doi:10.1093/nar/gky1120
Bunz, F. (2016). *Tumor Suppressor Genes*. Dordrecht: Springer.
Caldas, C. (2012). Cancer sequencing unravels clonal evolution. *Nat Biotechnol, 30*(5), 408-410. doi:10.1038/nbt.2213
Castera, L., Harter, V., Muller, E., Krieger, S., Goardon, N., Ricou, A., . . . Vaur, D. (2018). Landscape of pathogenic variations in a panel of 34 genes and cancer risk estimation from 5131 HBOC families. *Genet Med, 20*(12), 1677-1686. doi:10.1038/s41436-018-0005-9
Chang, H. H. Y., Pannunzio, N. R., Adachi, N., & Lieber, M. R. (2017). Non-homologous DNA end joining and alternative pathways to double-strand break repair. *Nat Rev Mol Cell Biol, 18*(8), 495-506. doi:10.1038/nrm.2017.48
Charoute, H., Nahili, H., Abidi, O., Gabi, K., Rouba, H., Fakiri, M., & Barakat, A. (2014). The Moroccan Genetic Disease Database (MGDD): a database for DNA variations related to inherited disorders and disease susceptibility. *Eur J Hum Genet, 22*(3), 322-326. doi:10.1038/ejhg.2013.151
Chen, S., & Parmigiani, G. (2007). Meta-analysis of BRCA1 and BRCA2 penetrance. *J Clin Oncol, 25*(11), 1329-1333. doi:10.1200/JCO.2006.09.1066
Christmann, M., Tomicic, M. T., Roos, W. P., & Kaina, B. (2003). Mechanisms of human DNA repair: an update. *Toxicology, 193*(1-2), 3-34. doi:10.1016/s0300-483x(03)00287-7
Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., . . . Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin), 6*(2), 80-92. doi:10.4161/fly.19695

Clarke, J., Wu, H. C., Jayasinghe, L., Patel, A., Reid, S., & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol, 4*(4), 265-270. doi:10.1038/nnano.2009.12

Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., . . . Genomes Project, C. (2012). The 1000 Genomes Project: data management and community access. *Nat Methods, 9*(5), 459-462. doi:10.1038/nmeth.1974

Clayton, E. A., Khalid, S., Ban, D., Wang, L., Jordan, I. K., & McDonald, J. F. (2020). Tumor suppressor genes and allele-specific expression: mechanisms and significance. *Oncotarget, 11*(4), 462-479. doi:10.18632/oncotarget.27468

Cohen, S. A., Pritchard, C. C., & Jarvik, G. P. (2019). Lynch Syndrome: From Screening to Diagnosis to Treatment in the Era of Modern Molecular Oncology. *Annu Rev Genomics Hum Genet, 20*, 293-307. doi:10.1146/annurev-genom-083118-015406

Cooke, D. P., Wedge, D. C., & Lunter, G. (2021). A unified haplotype-based method for accurate and comprehensive variant calling. *Nat Biotechnol*. doi:10.1038/s41587-021-00861-3

Cooper, G. M., & Hausman, R. E. (2016). *The Cell: A Molecular Approach*: Sinauer Associates, Inc.

Cragun, D., Weidner, A., Tezak, A., Clouse, K., & Pal, T. (2020). Cancer risk management among female BRCA1/2, PALB2, CHEK2, and ATM carriers. *Breast Cancer Res Treat, 182*(2), 421-428. doi:10.1007/s10549-020-05699-y

Croce, C. M. (2008). Oncogenes and cancer. *N Engl J Med, 358*(5), 502-511. doi:10.1056/NEJMra072367

Dagogo-Jack, I., & Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol, 15*(2), 81-94. doi:10.1038/nrclinonc.2017.166

Daly, M. B., Pilarski, R., Yurgelun, M. B., Berry, M. P., Buys, S. S., Dickson, P., . . . Darlow, S. D. (2020). NCCN Guidelines Insights: Genetic/Familial High-Risk Assessment: Breast, Ovarian, and Pancreatic, Version 1.2020. *J Natl Compr Canc Netw, 18*(4), 380-391. doi:10.6004/jnccn.2020.0017

Daniell, J., Plazzer, J. P., Perera, A., & Macrae, F. (2018). An exploration of genotype-phenotype link between Peutz-Jeghers syndrome and STK11: a review. *Fam Cancer, 17*(3), 421-427. doi:10.1007/s10689-017-0037-3

de Winter, J. P., & Joenje, H. (2009). The genetic and molecular basis of Fanconi anemia. *Mutat Res, 668*(1-2), 11-19. doi:10.1016/j.mrfmmm.2008.11.004

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., . . . Mabee, P. (2015). Finding our way through phenotypes. *PLoS Biol, 13*(1), e1002033. doi:10.1371/journal.pbio.1002033

den Dunnen, J. T. (2018). Yet another database? *Hum Mutat, 39*(6), 755. doi:10.1002/humu.23429

den Dunnen, J. T., Dalgleish, R., Maglott, D. R., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., . . . Taschner, P. E. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat, 37*(6), 564-569. doi:10.1002/humu.22981

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., . . . Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet, 43*(5), 491-498. doi:10.1038/ng.806

Dhawan, D. (2017). *Clinical Next-Generation Sequencing: Enabling Precision Medicine*: Elsevier.

do Valle, I. F., Giampieri, E., Simonetti, G., Padella, A., Manfrini, M., Ferrari, A., . . . Castellani, G. (2016). Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data. *BMC Bioinformatics, 17*(Suppl 12), 341. doi:10.1186/s12859-016-1190-7

Easton, D. F., Pharoah, P. D., Antoniou, A. C., Tischkowitz, M., Tavtigian, S. V., Nathanson, K. L., . . . Foulkes, W. D. (2015). Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med, 372*(23), 2243-2257. doi:10.1056/NEJMsr1501341

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol, 6*(5), R44. doi:10.1186/gb-2005-6-5-r44

Ensembl. Ensembl Variation - Calculated variant consequences. https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html

Epstein, R. J. (2015). A periodic table for cancer. *Future Oncol, 11*(5), 785-800. doi:10.2217/fon.14.315

Fakhro, K. A., Staudt, M. R., Ramstetter, M. D., Robay, A., Malek, J. A., Badii, R., . . . Rodriguez-Flores, J. L. (2016). The Qatar genome: a population-specific tool for precision medicine in the Middle East. *Hum Genome Var, 3*, 16016. doi:10.1038/hgv.2016.16

Fewings, E., Larionov, A., Redman, J., Goldgraben, M. A., Scarth, J., Richardson, S., . . . Tischkowitz, M. (2018). Germline pathogenic variants in PALB2 and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without CDH1 mutation: a whole-exome sequencing study. *Lancet Gastroenterol Hepatol, 3*(7), 489-498. doi:10.1016/S2468-1253(18)30079-7

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., . . . Shendure, J. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature, 562*(7726), 217-222. doi:10.1038/s41586-018-0461-z

Fishbein, L., & Nathanson, K. L. (2012). Pheochromocytoma and paraganglioma: understanding the complexities of the genetic background. *Cancer Genet, 205*(1-2), 1-11. doi:10.1016/j.cancergen.2012.01.009

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., . . . Searle, S. M. (2012). Ensembl 2012. *Nucleic Acids Res, 40*(Database issue), D84-90. doi:10.1093/nar/gkr991

Fokkema, I. F., Taschner, P. E., Schaafsma, G. C., Celli, J., Laros, J. F., & den Dunnen, J. T. (2011). LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat, 32*(5), 557-563. doi:10.1002/humu.21438

Fostira, F., Kostantopoulou, I., Apostolou, P., Papamentzelopoulou, M. S., Papadimitriou, C., Faliakou, E., . . . Yannoukakos, D. (2020). One in three highly selected Greek patients with breast cancer carries a loss-of-function variant in a cancer susceptibility gene. *J Med Genet, 57*(1), 53-61. doi:10.1136/jmedgenet-2019-106189

Fostira, F., Mollaki, V., Lypas, G., Alexandrakis, G., Christianakis, E., Tzouvala, M., . . . Yannoukakos, D. (2018). Genetic analysis and clinical description of Greek patients with Peutz-Jeghers syndrome: Creation of a National Registry. *Cancer Genet, 220*, 19-23. doi:10.1016/j.cancergen.2017.11.004

Fostira, F., Papademitriou, C., Efremidis, A., & Yannoukakos, D. (2010). An in-frame exon-skipping MUTYH mutation is associated with early-onset colorectal cancer. *Dis Colon Rectum, 53*(8), 1197-1201. doi:10.1007/DCR.0b013e3181dcf0c1

Fostira, F., Saloustros, E., Apostolou, P., Vagena, A., Kalfakakou, D., Mauri, D., . . . Konstantopoulou, I. (2018). Germline deleterious mutations in genes other than BRCA2 are infrequent in male breast cancer. *Breast Cancer Res Treat, 169*(1), 105-113. doi:10.1007/s10549-018-4661-x

Fostira, F., Thodi, G., Konstantopoulou, I., Sandaltzopoulos, R., & Yannoukakos, D. (2007). Hereditary cancer syndromes. *J BUON, 12 Suppl 1*, S13-22.

Foulkes, W. D. (2008). Inherited susceptibility to common cancers. *N Engl J Med, 359*(20), 2143-2153. doi:10.1056/NEJMra0802968

Frankish, A., Uszczynska, B., Ritchie, G. R., Gonzalez, J. M., Pervouchine, D., Petryszak, R., . . . Harrow, J. (2015). Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction. *BMC Genomics, 16 Suppl 8*, S2. doi:10.1186/1471-2164-16-S8-S2

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., . . . Akey, J. M. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature, 493*(7431), 216-220. doi:10.1038/nature11690

Fulda, S. (2010). Evasion of apoptosis as a cellular stress response in cancer. *Int J Cell Biol, 2010*, 370835. doi:10.1155/2010/370835
Gaffney, E. F., Riegman, P. H., Grizzle, W. E., & Watson, P. H. (2018). Factors that drive the increasing use of FFPE tissue in basic and translational cancer research. *Biotech Histochem, 93*(5), 373-386. doi:10.1080/10520295.2018.1446101
Galiatsatos, P., & Foulkes, W. D. (2006). Familial adenomatous polyposis. *Am J Gastroenterol, 101*(2), 385-398. doi:10.1111/j.1572-0241.2006.00375.x
Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2019). A brief history of bioinformatics. *Brief Bioinform, 20*(6), 1981-1996. doi:10.1093/bib/bby063
Genomes Project, C., Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., . . . Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature, 526*(7571), 68-74. doi:10.1038/nature15393
Genomics Education Program, N. (2021). Cancer Genomics. Retrieved from https://www.genomicseducation.hee.nhs.uk/cancer-genomics/.
Gilissen, C., Hoischen, A., Brunner, H. G., & Veltman, J. A. (2012). Disease gene identification strategies for exome sequencing. *Eur J Hum Genet, 20*(5), 490-497. doi:10.1038/ejhg.2011.258
Glenn, T. C. (2011). Field guide to next-generation DNA sequencers. *Mol Ecol Resour, 11*(5), 759-769. doi:10.1111/j.1755-0998.2011.03024.x
Gomez-Lopez, G., Dopazo, J., Cigudosa, J. C., Valencia, A., & Al-Shahrour, F. (2019). Precision medicine needs pioneering clinical bioinformaticians. *Brief Bioinform, 20*(3), 752-766. doi:10.1093/bib/bbx144
Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. *J Invest Dermatol, 133*(8), e11. doi:10.1038/jid.2013.248
Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, C. L., & Gelbart, W. M. (2000). *Somatic versus germinal mutation*. New York: Freeman, W. H. and Company.
Guha, T., & Malkin, D. (2017). Inherited TP53 Mutations and the Li-Fraumeni Syndrome. *Cold Spring Harb Perspect Med, 7*(4). doi:10.1101/cshperspect.a026187
Gunning, A. C., Fryer, V., Fasham, J., Crosby, A. H., Ellard, S., Baple, E. L., & Wright, C. F. (2020). Assessing performance of pathogenicity predictors using clinically relevant variant datasets. *J Med Genet*. doi:10.1136/jmedgenet-2020-107003
Gupta, S., Provenzale, D., Llor, X., Halverson, A. L., Grady, W., Chung, D. C., . . . Ogba, N. (2020). NCCN Guidelines Insights: Genetic/Familial High-Risk Assessment: Colorectal, Version 1.2020. *J Natl Compr Canc Netw*.
Half, E., Bercovich, D., & Rozen, P. (2009). Familial adenomatous polyposis. *Orphanet J Rare Dis, 4*, 22. doi:10.1186/1750-1172-4-22
Hamosh, A., Scott, A. F., Amberger, J., Bocchini, C., Valle, D., & McKusick, V. A. (2002). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res, 30*(1), 52-55. doi:10.1093/nar/30.1.52
Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell, 144*(5), 646-674. doi:10.1016/j.cell.2011.02.013
Hansford, S., Kaurah, P., Li-Chang, H., Woo, M., Senz, J., Pinheiro, H., . . . Huntsman, D. G. (2015). Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol, 1*(1), 23-32. doi:10.1001/jamaoncol.2014.168
Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., . . . Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res, 22*(9), 1760-1774. doi:10.1101/gr.135350.111
Hartl, M., & Bister, K. (2013). *Oncogenes* (2nd Edition ed.): Academic Press.
Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics, 107*(1), 1-8. doi:10.1016/j.ygeno.2015.11.003
Helena, J. M., Joubert, A. M., Grobbelaar, S., Nolte, E. M., Nel, M., Pepper, M. S., . . . Mercier, A. E. (2018). Deoxyribonucleic Acid Damage and Repair: Capitalizing on Our Understanding

of the Mechanisms of Maintaining Genomic Integrity for Therapeutic Purposes. *Int J Mol Sci, 19*(4). doi:10.3390/ijms19041148
Helleday, T. (2011). The underlying mechanism for the PARP and BRCA synthetic lethality: clearing up the misunderstandings. *Mol Oncol, 5*(4), 387-393. doi:10.1016/j.molonc.2011.07.001
Hindorff, L. A., Gillanders, E. M., & Manolio, T. A. (2011). Genetic architecture of cancer and other complex diseases: lessons learned and future directions. *Carcinogenesis, 32*(7), 945-954. doi:10.1093/carcin/bgr056
Hino, O., & Kobayashi, T. (2017). Mourning Dr. Alfred G. Knudson: the two-hit hypothesis, tumor suppressor genes, and the tuberous sclerosis complex. *Cancer Sci, 108*(1), 5-11. doi:10.1111/cas.13116
Hodson, R. (2016). Precision medicine. *Nature, 537*(7619), S49. doi:10.1038/537S49a
Hood, L. (2008). A personal journey of discovery: developing technology and changing biology. *Annu Rev Anal Chem (Palo Alto Calif), 1*, 1-43. doi:10.1146/annurev.anchem.1.031207.113113
Hood, L., & Rowen, L. (2013). The Human Genome Project: big science transforms biology and medicine. *Genome Med, 5*(9), 79. doi:10.1186/gm483
Houweling, A. C., Gijezen, L. M., Jonker, M. A., van Doorn, M. B., Oldenburg, R. A., van Spaendonck-Zwarts, K. Y., . . . Menko, F. H. (2011). Renal cancer and pneumothorax risk in Birt-Hogg-Dube syndrome; an analysis of 115 FLCN mutation carriers from 35 BHD families. *Br J Cancer, 105*(12), 1912-1919. doi:10.1038/bjc.2011.463
Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Hum Immunol*. doi:10.1016/j.humimm.2021.02.012
International HapMap, C., Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., . . . Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature, 449*(7164), 851-861. doi:10.1038/nature06258
Jalali Sefid Dashti, M., & Gamieldien, J. (2017). A practical guide to filtering and prioritizing genetic variants. *Biotechniques, 62*(1), 18-30. doi:10.2144/000114492
Jan, R., & Chaudhry, G. E. (2019). Understanding Apoptosis and Apoptotic Pathways Targeted Cancer Therapeutics. *Adv Pharm Bull, 9*(2), 205-218. doi:10.15171/apb.2019.024
Jang, H. S., Shah, N. M., Du, A. Y., Dailey, Z. Z., Pehrsson, E. C., Godoy, P. M., . . . Wang, T. (2019). Transposable elements drive widespread expression of oncogenes in human cancers. *Nat Genet, 51*(4), 611-617. doi:10.1038/s41588-019-0373-3
Janoueix-Lerosey, I., Lequin, D., Brugieres, L., Ribeiro, A., de Pontual, L., Combaret, V., . . . Delattre, O. (2008). Somatic and germline activating mutations of the ALK kinase receptor in neuroblastoma. *Nature, 455*(7215), 967-970. doi:10.1038/nature07398
Jeggo, P. A., Pearl, L. H., & Carr, A. M. (2016). DNA repair, genome stability and cancer: a historical perspective. *Nat Rev Cancer, 16*(1), 35-42. doi:10.1038/nrc.2015.4
Kaelin, W. G., Jr. (2005). The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer, 5*(9), 689-698. doi:10.1038/nrc1691
Kalfakakou, D., Fostira, F., Papathanasiou, A., Apostolou, P., Dellatola, V., Gavra, I. E., . . . Konstantopoulou, I. (2021). CanVaS: Documenting the genetic variation spectrum of Greek cancer patients. *Hum Mutat*. doi:10.1002/humu.24249
Kalfakakou, D., Konstantopoulou, I., Yannoukakos, D., & Fostira, F. (2021). Pitfalls in variant annotation for hereditary cancer diagnostics: The example of Illumina(R) VariantStudio(R). *Genomics, 113*(1 Pt 2), 748-754. doi:10.1016/j.ygeno.2020.10.005
Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alfoldi, J., Wang, Q., . . . MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature, 581*(7809), 434-443. doi:10.1038/s41586-020-2308-7
Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC Med Genomics, 8*, 37. doi:10.1186/s12920-015-0115-z

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi:10.1101/gr.229102

Kitao, S., Shimamoto, A., Goto, M., Miller, R. W., Smithson, W. A., Lindor, N. M., & Furuichi, Y. (1999). Mutations in RECQL4 cause a subset of cases of Rothmund-Thomson syndrome. *Nat Genet, 22*(1), 82-84. doi:10.1038/8788

Kleanthous, M., Patsalis, P. C., Drousiotou, A., Motazacker, M., Christodoulou, K., Cariolou, M., . . . Patrinos, G. P. (2006). The cypriot and Iranian National Mutation Frequency Databases. *Hum Mutat, 27*(6), 598-599. doi:10.1002/humu.9422

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., . . . Wilson, R. K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res, 22*(3), 568-576. doi:10.1101/gr.129684.111

Koch, L. (2020). Exploring human genomic diversity with gnomAD. *Nat Rev Genet, 21*(8), 448. doi:10.1038/s41576-020-0255-7

Konstanta, I., Fostira, F., Apostolou, P., Stratikos, E., Kalfakakou, D., Pampanos, A., . . . Yannoukakos, D. (2018). Contribution of RAD51D germline mutations in breast and ovarian cancer in Greece. *J Hum Genet, 63*(11), 1149-1158. doi:10.1038/s10038-018-0498-8

Konstantopoulou, I., Tsitlaidou, M., Fostira, F., Pertesi, M., Stavropoulou, A. V., Triantafyllidou, O., . . . Yannoukakos, D. (2014). High prevalence of BRCA1 founder mutations in Greek breast/ovarian families. *Clin Genet, 85*(1), 36-42. doi:10.1111/cge.12274

Kroigard, A. B., Thomassen, M., Laenkholm, A. V., Kruse, T. A., & Larsen, M. J. (2016). Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exome and Targeted Deep Sequencing Data. *PLoS One, 11*(3), e0151664. doi:10.1371/journal.pone.0151664

Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R., Phillips, K. A., Mooij, T. M., Roos-Blom, M. J., . . . Olsson, H. (2017). Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *JAMA, 317*(23), 2402-2416. doi:10.1001/jama.2017.7112

Kulkarni, A., & Carley, H. (2016). Advances in the recognition and management of hereditary cancer. *Br Med Bull, 120*(1), 123-138. doi:10.1093/bmb/ldw046

Kurian, A. W., Antoniou, A. C., & Domchek, S. M. (2016). Refining Breast Cancer Risk Stratification: Additional Genes, Additional Information. *Am Soc Clin Oncol Educ Book, 35*, 44-56. doi:10.14694/EDBK_158817Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., . . . International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature, 409*(6822), 860-921. doi:10.1038/35057062

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., . . . Kattman, B. L. (2020). ClinVar: improvements to accessing data. *Nucleic Acids Res, 48*(D1), D835-D844. doi:10.1093/nar/gkz972

Lee, Y. R., Chen, M., & Pandolfi, P. P. (2018). The functions and regulation of the PTEN tumour suppressor: new modes and prospects. *Nat Rev Mol Cell Biol, 19*(9), 547-562. doi:10.1038/s41580-018-0015-0

Lehmann, A. R., McGibbon, D., & Stefanini, M. (2011). Xeroderma pigmentosum. *Orphanet J Rare Dis, 6*, 70. doi:10.1186/1750-1172-6-70

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., . . . Exome Aggregation, C. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature, 536*(7616), 285-291. doi:10.1038/nature19057

Levy-Lahad, E., Catane, R., Eisenberg, S., Kaufman, B., Hornreich, G., Lishinsky, E., . . . Halle, D. (1997). Founder BRCA1 and BRCA2 mutations in Ashkenazi Jews in Israel: frequency and differential penetrance in ovarian cancer and in breast-ovarian cancer families. *Am J Hum Genet, 60*(5), 1059-1067.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform, 11*(5), 473-483. doi:10.1093/bib/bbq015
Li, M. M., Datto, M., Duncavage, E. J., Kulkarni, S., Lindeman, N. I., Roy, S., . . . Nikiforova, M. N. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. *J Mol Diagn, 19*(1), 4-23. doi:10.1016/j.jmoldx.2016.10.002
Li, Y., Zhang, J., Adikaram, P. R., Welch, J., Guan, B., Weinstein, L. S., . . . Simonds, W. F. (2020). Genotype of CDC73 germline mutation determines risk of parathyroid cancer. *Endocr Relat Cancer, 27*(9), 483-494. doi:10.1530/ERC-20-0149
Lionel, A. C., Costain, G., Monfared, N., Walker, S., Reuter, M. S., Hosseini, S. M., . . . Marshall, C. R. (2018). Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. *Genet Med, 20*(4), 435-443. doi:10.1038/gim.2017.119
Liu, Z. K., Shang, Y. K., Chen, Z. N., & Bian, H. (2017). A three-caller pipeline for variant analysis of cancer whole-exome sequencing data. *Mol Med Rep, 15*(5), 2489-2494. doi:10.3892/mmr.2017.6336
Lonser, R. R., Glenn, G. M., Walther, M., Chew, E. Y., Libutti, S. K., Linehan, W. M., & Oldfield, E. H. (2003). von Hippel-Lindau disease. *Lancet, 361*(9374), 2059-2067. doi:10.1016/S0140-6736(03)13643-4
Lynch, H. T., & de la Chapelle, A. (1999). Genetic susceptibility to non-polyposis colorectal cancer. *J Med Genet, 36*(11), 801-818.
Mai, P. L., Best, A. F., Peters, J. A., DeCastro, R. M., Khincha, P. P., Loud, J. T., . . . Savage, S. A. (2016). Risks of first and subsequent cancers among TP53 mutation carriers in the National Cancer Institute Li-Fraumeni syndrome cohort. *Cancer, 122*(23), 3673-3681. doi:10.1002/cncr.30248
Malarkey, D. E., Hoenerhoff, M. J., & Maronpot, R. R. (2018). *Carcinogenesis: Manifestation and Mechanisms*: Academic Press.
Marcotte, L., & Crino, P. B. (2006). The neurobiology of the tuberous sclerosis complex. *Neuromolecular Med, 8*(4), 531-546. doi:10.1385/NMM:8:4:531
Mardis, E. R. (2013). Next-generation sequencing platforms. *Annu Rev Anal Chem (Palo Alto Calif), 6*, 287-303. doi:10.1146/annurev-anchem-062012-092628
Marketos, S. History of Medicine. http://asclepieion.mpl.uoa.gr/parko/marketos2.htm
Mateo, J., Lord, C. J., Serra, V., Tutt, A., Balmana, J., Castroviejo-Bermejo, M., . . . de Bono, J. S. (2019). A decade of clinical development of PARP inhibitors in perspective. *Ann Oncol, 30*(9), 1437-1447. doi:10.1093/annonc/mdz192
Mavaddat, N., Michailidou, K., Dennis, J., Lush, M., Fachal, L., Lee, A., . . . Easton, D. F. (2019). Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet, 104*(1), 21-34. doi:10.1016/j.ajhg.2018.11.002
Mayer, D. K., Nekhlyudov, L., Snyder, C. F., Merrill, J. K., Wollins, D. S., & Shulman, L. N. (2014). American Society of Clinical Oncology clinical expert statement on cancer survivorship care planning. *J Oncol Pract, 10*(6), 345-351. doi:10.1200/JOP.2014.001321
McBride, K. A., Ballinger, M. L., Killick, E., Kirk, J., Tattersall, M. H., Eeles, R. A., . . . Mitchell, G. (2014). Li-Fraumeni syndrome: cancer risk assessment and clinical management. *Nat Rev Clin Oncol, 11*(5), 260-271. doi:10.1038/nrclinonc.2014.41
McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J. B., & Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Med, 6*(3), 26. doi:10.1186/gm543
McDonough, S. J., Bhagwate, A., Sun, Z., Wang, C., Zschunke, M., Gorman, J. A., . . . Cunningham, J. M. (2019). Use of FFPE-derived DNA in next generation sequencing: DNA extraction methods. *PLoS One, 14*(4), e0211400. doi:10.1371/journal.pone.0211400
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-

generation DNA sequencing data. *Genome Res, 20*(9), 1297-1303. doi:10.1101/gr.107524.110
McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., . . . Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res, 19*(9), 1527-1541. doi:10.1101/gr.091868.109
McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., . . . Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biol, 17*(1), 122. doi:10.1186/s13059-016-0974-4
Meera Krishna, B., Khan, M. A., & Khan, S. T. (2019). Next-Generation Sequencing (NGS) Platforms: An Exciting Era of Genome Sequence Analysis. In V. Tripathi, P. Kumar, P. Tripathi, A. Kishore, & M. Kamle (Eds.), *Microbial Genomics in Sustainable Agroecosystems*. Singapore: Springer.
Mello, S. S., & Attardi, L. D. (2018). Deciphering p53 signaling in tumor suppression. *Curr Opin Cell Biol, 51*, 65-72. doi:10.1016/j.ceb.2017.11.005
Menko, F. H., Maher, E. R., Schmidt, L. S., Middelton, L. A., Aittomaki, K., Tomlinson, I., . . . Linehan, W. M. (2014). Hereditary leiomyomatosis and renal cell cancer (HLRCC): renal cancer risk, surveillance and treatment. *Fam Cancer, 13*(4), 637-644. doi:10.1007/s10689-014-9735-2
Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P. A., Harshman, K., Tavtigian, S., . . . et al. (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science, 266*(5182), 66-71. doi:10.1126/science.7545954
Miller, D. T., Lee, K., Gordon, A. S., Amendola, L. M., Adelman, K., Bale, S. J., . . . Group, A. S. F. W. (2021). Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2021 update: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet Med*. doi:10.1038/s41436-021-01171-4
Mosse, Y. P., Laudenslager, M., Longo, L., Cole, K. A., Wood, A., Attiyeh, E. F., . . . Maris, J. M. (2008). Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature, 455*(7215), 930-935. doi:10.1038/nature07261
Mu, W., Lu, H. M., Chen, J., Li, S., & Elliott, A. M. (2016). Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. *J Mol Diagn, 18*(6), 923-932. doi:10.1016/j.jmoldx.2016.07.006
Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., . . . Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature, 472*(7341), 90-94. doi:10.1038/nature09807
Nielsen, M., Morreau, H., Vasen, H. F., & Hes, F. J. (2011). MUTYH-associated polyposis (MAP). *Crit Rev Oncol Hematol, 79*(1), 1-16. doi:10.1016/j.critrevonc.2010.05.011
Oliver, G. R., Hart, S. N., & Klee, E. W. (2015). Bioinformatics for clinical next generation sequencing. *Clin Chem, 61*(1), 124-135. doi:10.1373/clinchem.2014.224360
Parker, N., Schneegurt, M., Thi Tu, A. H., Forster, B., & Lister, P. (2019). Microbiology. https://opentextbc.ca/microbiologyopenstax/chapter/mutations/
Patrinos, G. P., van Baal, S., Petersen, M. B., & Papadakis, M. N. (2005). Hellenic National Mutation database: a prototype database for mutations leading to inherited disorders in the Hellenic population. *Hum Mutat, 25*(4), 327-333. doi:10.1002/humu.20157
Pertesi, M., Konstantopoulou, I., & Yannoukakos, D. (2011). Haplotype analysis of two recurrent genomic rearrangements in the BRCA1 gene suggests they are founder mutations for the Greek population. *Clin Genet, 80*(4), 375-382. doi:10.1111/j.1399-0004.2010.01532.x
Peters, U., Bien, S., & Zubair, N. (2015). Genetic architecture of colorectal cancer. *Gut, 64*(10), 1623-1636. doi:10.1136/gutjnl-2013-306705
Pierron, G. (2015). *Basis for Molecular Genetics in Cancer*: Springer.
Pilarski, R., Burt, R., Kohlman, W., Pho, L., Shannon, K. M., & Swisher, E. (2013). Cowden syndrome and the PTEN hamartoma tumor syndrome: systematic review and revised diagnostic criteria. *J Natl Cancer Inst, 105*(21), 1607-1616. doi:10.1093/jnci/djt277

Pradhan, S., Sengupta, M., Dutta, A., Bhattacharyya, K., Bag, S. K., Dutta, C., & Ray, K. (2011). Indian genetic disease database. *Nucleic Acids Res, 39*(Database issue), D933-938. doi:10.1093/nar/gkq1025

Provenzale, D., Ness, R. M., Llor, X., Weiss, J. M., Abbadessa, B., Cooper, G., . . . Ogba, N. (2020). NCCN Guidelines Insights: Colorectal Cancer Screening, Version 2.2020. *J Natl Compr Canc Netw, 18*(10), 1312-1320. doi:10.6004/jnccn.2020.0048

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res, 40*(Database issue), D130-135. doi:10.1093/nar/gkr1079

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res, 35*(Database issue), D61-65. doi:10.1093/nar/gkl842

Pulciani, S., Di Lonardo, A., Fagnani, C., & Taruscio, D. (2017). P4 Medicine versus Hippocrates. *Ann Ist Super Sanita, 53*(3), 185-191. doi:10.4415/ANN_17_03_02

Puntervoll, H. E., Yang, X. R., Vetti, H. H., Bachmann, I. M., Avril, M. F., Benfodda, M., . . . Molven, A. (2013). Melanoma prone families with CDK4 germline mutation: phenotypic profile and associations with MC1R variants. *J Med Genet, 50*(4), 264-270. doi:10.1136/jmedgenet-2012-101455

Rahner, N., & Steinke, V. (2008). Hereditary cancer syndromes. *Dtsch Arztebl Int, 105*(41), 706-714. doi:10.3238/arztebl.2008.0706

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., . . . Committee, A. L. Q. A. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med, 17*(5), 405-424. doi:10.1038/gim.2015.30

Rocha, C. S., Secolin, R., Rodrigues, M. R., Carvalho, B. S., & Lopes-Cendes, I. (2020). The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. *NPJ Genom Med, 5*, 42. doi:10.1038/s41525-020-00149-6

Rodriguez, J. M., Rodriguez-Rivas, J., Di Domenico, T., Vazquez, J., Valencia, A., & Tress, M. L. (2018). APPRIS 2017: principal isoforms for multiple gene sets. *Nucleic Acids Res, 46*(D1), D213-D217. doi:10.1093/nar/gkx997

Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H., & Poirier, G. G. (2010). PARP inhibition: PARP1 and beyond. *Nat Rev Cancer, 10*(4), 293-301. doi:10.1038/nrc2812

Rousset-Jablonski, C., & Gompel, A. (2017). Screening for familial cancer risk: Focus on breast cancer. *Maturitas, 105*, 69-77. doi:10.1016/j.maturitas.2017.08.004

Ruangrit, U., Srikummool, M., Assawamakin, A., Ngamphiw, C., Chuechote, S., Thaiprasarnsup, V., . . . Tongsima, S. (2008). Thailand mutation and variation database (ThaiMUT). *Hum Mutat, 29*(8), E68-75. doi:10.1002/humu.20787

Saleh-Gohari, N., Bryant, H. E., Schultz, N., Parker, K. M., Cassel, T. N., & Helleday, T. (2005). Spontaneous homologous recombination is induced by collapsed replication forks that are caused by endogenous DNA single-strand breaks. *Mol Cell Biol, 25*(16), 7158-7169. doi:10.1128/MCB.25.16.7158-7169.2005

Salpea, P., & Stratakis, C. A. (2014). Carney complex and McCune Albright syndrome: an overview of clinical manifestations and human molecular genetics. *Mol Cell Endocrinol, 386*(1-2), 85-91. doi:10.1016/j.mce.2013.08.022

Samadder, N. J., Jasperson, K., & Burt, R. W. (2015). Hereditary and common familial colorectal cancer: evidence for colorectal screening. *Dig Dis Sci, 60*(3), 734-747. doi:10.1007/s10620-014-3465-z

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467. doi:10.1073/pnas.74.12.5463

Shyr, D., & Liu, Q. (2013). Next generation sequencing in cancer research and clinical application. *Biol Proced Online, 15*(1), 4. doi:10.1186/1480-9222-15-4

Sipila, K., & Aula, P. (2002). Database for the mutations of the Finnish disease heritage. *Hum Mutat, 19*(1), 16-22. doi:10.1002/humu.10019
Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol, 122*(1), e59. doi:10.1002/cpmb.59
Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res, 27*(3), 491-499. doi:10.1101/gr.209601.116
Speicher, M. R., Geigl, J. B., & Tomlinson, I. P. (2010). Effect of genome-wide association studies, direct-to-consumer genetic testing, and high-speed sequencing technologies on predictive genetic counselling for cancer risk. *Lancet Oncol, 11*(9), 890-898. doi:10.1016/S1470-2045(09)70359-6
Stewart, D. R., Best, A. F., Williams, G. M., Harney, L. A., Carr, A. G., Harris, A. K., . . . Schultz, K. A. P. (2019). Neoplasm Risk Among Individuals With a Pathogenic Germline Variant in DICER1. *J Clin Oncol, 37*(8), 668-676. doi:10.1200/JCO.2018.78.4678
Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature, 458*(7239), 719-724. doi:10.1038/nature07943
Sud, A., Kinnersley, B., & Houlston, R. S. (2017). Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer, 17*(11), 692-704. doi:10.1038/nrc.2017.82
Sun, Y., Ruivenkamp, C. A., Hoffer, M. J., Vrijenhoek, T., Kriek, M., van Asperen, C. J., . . . Santen, G. W. (2015). Next-generation diagnostics: gene panel, exome, or whole genome? *Hum Mutat, 36*(6), 648-655. doi:10.1002/humu.22783
Suszynska, M., Ratajska, M., & Kozlowski, P. (2020). BRIP1, RAD51C, and RAD51D mutations are associated with high susceptibility to ovarian cancer: mutation prevalence and precise risk estimates based on a pooled analysis of ~30,000 cases. *J Ovarian Res, 13*(1), 50. doi:10.1186/s13048-020-00654-3
Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., . . . Project, N. E. S. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science, 337*(6090), 64-69. doi:10.1126/science.1219240
Thorvaldsdottir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform, 14*(2), 178-192. doi:10.1093/bib/bbs017
Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., . . . Feolo, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res, 42*(Database issue), D975-979. doi:10.1093/nar/gkt1211
Tubbs, A., & Nussenzweig, A. (2017). Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell, 168*(4), 644-656. doi:10.1016/j.cell.2017.01.002
Turajlic, S., Sottoriva, A., Graham, T., & Swanton, C. (2019). Resolving genetic heterogeneity in cancer. *Nat Rev Genet, 20*(7), 404-416. doi:10.1038/s41576-019-0114-6
Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A. P. (2008). A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic Acids Res, 36*(4), e25. doi:10.1093/nar/gkn021
van Baal, S., Kaimakis, P., Phommarinh, M., Koumbi, D., Cuppens, H., Riccardino, F., . . . Patrinos, G. P. (2007). FINDbase: a relational database recording frequencies of genetic defects leading to inherited disorders worldwide. *Nucleic Acids Res, 35*(Database issue), D690-695. doi:10.1093/nar/gkl934
van der Groep, P., van der Wall, E., & van Diest, P. J. (2011). Pathology of hereditary breast cancer. *Cell Oncol (Dordr), 34*(2), 71-88. doi:10.1007/s13402-011-0010-3
van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet, 30*(9), 418-426. doi:10.1016/j.tig.2014.07.001
van Heemst, D., den Reijer, P. M., & Westendorp, R. G. (2007). Ageing or cancer: a review on the role of caretakers and gatekeepers. *Eur J Cancer, 43*(15), 2144-2152. doi:10.1016/j.ejca.2007.07.011

van Lier, M. G., Wagner, A., Mathus-Vliegen, E. M., Kuipers, E. J., Steyerberg, E. W., & van Leerdam, M. E. (2010). High cancer risk in Peutz-Jeghers syndrome: a systematic review and surveillance recommendations. *Am J Gastroenterol, 105*(6), 1258-1264; author reply 1265. doi:10.1038/ajg.2009.725

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., . . . Zhu, X. (2001). The sequence of the human genome. *Science, 291*(5507), 1304-1351. doi:10.1126/science.1058040

Walsh, C. S. (2015). Two decades beyond BRCA1/2: Homologous recombination, hereditary cancer risk and a target for ovarian cancer therapy. *Gynecol Oncol, 137*(2), 343-350. doi:10.1016/j.ygyno.2015.02.017

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*(16), e164. doi:10.1093/nar/gkq603

Wang, Q., Kotoula, V., Hsu, P. C., Papadopoulou, K., Ho, J. W. K., Fountzilas, G., & Giannoulatou, E. (2019). Comparison of somatic variant detection algorithms using Ion Torrent targeted deep sequencing data. *BMC Med Genomics, 12*(Suppl 9), 181. doi:10.1186/s12920-019-0636-y

Weinberg, R. A. (2013). *The biology of cancer* (2nd Edition ed.). US: Garland Science.

Wetterstand, K. A. (2020). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *www.genome.gov/sequencingcostsdata*.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., . . . Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. *Nature, 378*(6559), 789-792. doi:10.1038/378789a0

Wright, W. D., Shah, S. S., & Heyer, W. D. (2018). Homologous recombination and the repair of DNA double-strand breaks. *J Biol Chem, 293*(27), 10524-10535. doi:10.1074/jbc.TM118.000372

Xue, Y., Ankala, A., Wilcox, W. R., & Hegde, M. R. (2015). Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet Med, 17*(6), 444-451. doi:10.1038/gim.2014.122

Yang, S. Y., Hsiung, C. N., Li, Y. J., Chang, G. C., Tsai, Y. H., Chen, K. Y., . . . Hsu, H. M. (2016). Fanconi anemia genes in lung adenocarcinoma- a pathway-wide study on cancer susceptibility. *J Biomed Sci, 23*, 23. doi:10.1186/s12929-016-0240-9

Yang, X., Leslie, G., Doroszuk, A., Schneider, S., Allen, J., Decker, B., . . . Tischkowitz, M. (2020). Cancer Risks Associated With Germline PALB2 Pathogenic Variants: An International Study of 524 Families. *J Clin Oncol, 38*(7), 674-685. doi:10.1200/JCO.19.01907

Young, A. D., & Gillung, J. P. (2019). Phylogenomics — principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology, 45*(2), 225-247. doi:10.1111/syen.12406

Zhen, D. B., Rabe, K. G., Gallinger, S., Syngal, S., Schwartz, A. G., Goggins, M. G., . . . Petersen, G. M. (2015). BRCA1, BRCA2, PALB2, and CDKN2A mutations in familial pancreatic cancer: a PACGENE study. *Genet Med, 17*(7), 569-577. doi:10.1038/gim.2014.153

Zlotogora, J., van Baal, S., & Patrinos, G. P. (2007). Documentation of inherited disorders and mutation frequencies in the different religious communities in Israel in the Israeli National Genetic Database. *Hum Mutat, 28*(10), 944-949. doi:10.1002/humu.20551

Φωστήρα, Φ. (2009). *Γενετική ανάλυση στον κληρονομούμενο καρκίνο του παχέος εντέρου (ΚΚΠΕ, FAP, HNPCC).* (Διδακτορική Διατριβή), Δημοκρίτειο Πανεπιστήμιο Θράκης,