ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΚΛΙΝΙΚΗ ΒΙΟΧΗΜΕΙΑ-ΜΟΡΙΑΚΗ ΔΙΑΓΝΩΣΤΙΚΗ»

ΕΡΕΥΝΗΤΙΚΗ ΕΡΓΑΣΙΑ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

ΤΑΥΤΟΠΟΙΗΣΗ ΚΑΙ ΜΕΛΕΤΗ ΤΗΣ ΕΚΦΡΑΣΗΣ ΝΕΩΝ ΕΝΑΛΛΑΚΤΙΚΩΝ ΜΕΤΑΓΡΑΦΩΝ ΤΟΥ ΑΝΘΡΩΠΙΝΟΥ ΓΟΝΙΔΙΟΥ ΤΗΣ ΚΥΚΛΙΝΟΕΞΑΡΤΩΜΕΝΗΣ ΚΙΝΑΣΗΣ 4 (*CDK4*) ΣΕ ΚΑΡΚΙΝΙΚΑ ΚΥΤΤΑΡΑ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΟΛΟΓΙΩΝ ΜΑΖΙΚΗΣ ΠΑΡΑΛΛΗΛΗΣ ΑΛΛΗΛΟΥΧΗΣΗΣ

ΚΩΝΣΤΑΝΤΙΝΑ ΑΘΑΝΑΣΟΠΟΥΛΟΥ

ΒΙΟΛΟΓΟΣ

ΑΘΗΝΑ 2021

ΕΘΝΙΚΟ & ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

«ΚΛΙΝΙΚΗ ΒΙΟΧΗΜΕΙΑ-ΜΟΡΙΑΚΗ ΔΙΑΓΝΩΣΤΙΚΗ»

ΕΡΕΥΝΗΤΙΚΗ ΕΡΓΑΣΙΑ ΔΙΠΛΩΜΑΤΟΣ ΕΙΔΙΚΕΥΣΗΣ

ΤΑΥΤΟΠΟΙΗΣΗ ΚΑΙ ΜΕΛΕΤΗ ΤΗΣ ΕΚΦΡΑΣΗΣ ΝΕΩΝ ΕΝΑΛΛΑΚΤΙΚΩΝ ΜΕΤΑΓΡΑΦΩΝ ΤΟΥ ΑΝΘΡΩΠΙΝΟΥ ΓΟΝΙΔΙΟΥ ΤΗΣ ΚΥΚΛΙΝΟΕΞΑΡΤΩΜΕΝΗΣ ΚΙΝΑΣΗΣ 4 (*CDK4*) ΣΕ ΚΑΡΚΙΝΙΚΑ ΚΥΤΤΑΡΑ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΟΛΟΓΙΩΝ ΜΑΖΙΚΗΣ ΠΑΡΑΛΛΗΛΗΣ ΑΛΛΗΛΟΥΧΗΣΗΣ

ΚΩΝΣΤΑΝΤΙΝΑ ΑΘΑΝΑΣΟΠΟΥΛΟΥ

ΒΙΟΛΟΓΟΣ

ΑΘΗΝΑ 2021

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ΤΜΗΜΑ ΒΙΟΛΟΓΙΑΣ
ΤΟΜΕΑΣ ΒΙΟΧΗΜΕΙΑΣ & ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ

«**Ταυτοποίηση και μελέτη της έκφρασης νέων εναλλακτικών μεταγράφων του ανθρώπινου γονιδίου της κυκλινοεξαρτώμενης κινάσης 4 (*CDK4*) σε καρκινικά κύτταρα με χρήση μεθοδολογιών μαζικής παράλληλης αλληλούχησης**»

ΚΩΝΣΤΑΝΤΙΝΑ ΑΘΑΝΑΣΟΠΟΥΛΟΥ
ΒΙΟΛΟΓΟΣ

ΕΠΙΒΛΕΠΩΝ ΜΕΛΟΣ ΔΕΠ: Ανδρέας Σκορίλας, Καθηγητής, Τμήμα Βιολογίας, Ε.Κ.Π.Α

ΤΡΙΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

Σκορίλας Ανδρέας, Καθηγητής, Τμήμα Βιολογίας, Ε.Κ.Π.Α.
Σίδερης Διαμάντης, Αναπλ. Καθηγητής, Τμήμα Βιολογίας, Ε.Κ.Π.Α.
Κοντός Χρήστος, Επίκ. Καθηγητής, Τμήμα Βιολογίας, Ε.Κ.Π.Α.

Αθήνα 2021

ΠΡΟΛΟΓΟΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε κατά το ακαδημαϊκό έτος 2020-2021 στον Τομέα Βιοχημείας και Μοριακής Βιολογίας του Τμήματος Βιολογίας του Εθνικού και Καποδιστριακού Πανεπιστημίου Αθηνών, στο πλαίσιο του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών με τίτλο «Κλινική Βιοχημεία – Μοριακή Διαγνωστική» με επιστημονικό υπεύθυνο τον Καθηγητή Κλινικής Βιοχημείας κ. Ανδρέα Σκορίλα.

Αρχικά, θα ήθελα να ευχαριστήσω ένθερμα τον Καθηγητή κ. Ανδρέα Σκορίλα, τόσο για την εμπιστοσύνη που μου παρείχε το χρονικό διάστημα εκπόνησης της παρούσας διπλωματικής, καθώς και για την επιλογή του να συμμετέχω στην ερευνητική του ομάδα. Επίσης, η στήριξή του και η επιστημονική του καθοδήγηση έπαιξαν καθοριστικό ρόλο στην υλοποίηση της συγκεκριμένης επιστημονικής μελέτης και την επίτευξη των στόχων που είχαν τεθεί στο πλαίσιο αυτής.

Επιπλέον, θα ήθελα να ευχαριστήσω τόσο τον Αναπληρωτή Καθηγητή κ. Διαμάντη Σίδερη όσο και τον Επίκουρο Καθηγητή κ. Χρήστο Κοντό για τις συμβουλές τους και την καθοδήγησή τους.

Θα ήθελα επίσης να εκφράσω τις ευχαριστίες μου προς τον μεταδιδάκτωρ Παναγιώτη Αδαμόπουλο για την άψογη συνεργασία, την εμπιστοσύνη και την καθοδήγησή του καθόλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Στο πλαίσιο αυτό, θα ήθελα να ευχαριστήσω τους υποψήφιους διδάκτορες Παναγιώτη Τσιακανίκα και Μιχαέλα Μπότη για τις συμβουλές τους κατά τη διενέργεια των πειραμάτων της παρούσας εργασίας.

Θα ήθελα να ευχαριστήσω και όλα τα υπόλοιπα μέλη της ερευνητικής ομάδας για την υποστήριξη, την εξαιρετική συνεργασία, το πνεύμα ομαδικότητας, καθώς και το αρμονικό και ευχάριστο κλίμα που επικρατούσε στο εργαστήριο όλη αυτή την περίοδο.

Τέλος, οφείλω να ευχαριστήσω την οικογένειά μου, η οποία με στήριξε καθόλη την ακαδημαϊκή μου πορεία και εξακολουθεί να με στηρίζει μέχρι σήμερα.

Πίνακας Περιεχομένων

ΣΥΝΤΟΜΟΓΡΑΦΙΕΣ

Ampli-seq:	Amplicon-sequencing
ASDT:	Alternative Splicing Detection Tool
ATP:	Adenosine triphosphate
bp:	base pair
CDK:	Cyclin-dependent kinase
cDNA:	complementary DNA
ChIP-seq:	Chromatin immunoprecipitation-sequencing
CMOS:	Complementary Metal Oxide Semiconductor
ddNTP:	Dideoxynucleotide triphosphate
dNTP:	Deoxynucleotides triphosphate
dsDNA:	double stranded DNA
EDTA:	Ethylenediaminetetraacetic acid
emPCR:	Emulsion Polymerase Chan Reaction
EST:	Expressed Sequence Tag
gDNA:	genomic DNA
HGP:	Human Genome Project
HISAT2:	Hierarchical Indexing for Spliced Alignment of Transcripts2
HT-NGS:	High-Throughput - Next Generation Sequencing
IGV:	Integrative Genomics Viewer
ISFET:	Ion-Sensitive Field-Effect Transistor
ISP:	Ion Sphere Particle
kb:	kilobase
NGS:	Next Generation Sequencing
ONT:	Oxford Nanopore Technologies
ORF:	Open Reading Frame
PacBio:	Pacific Biosciences
PCR:	Polymerase Chain Reaction
PGM:	Personal Genome Machine
PP_i:	pyrophoshate
pre-mRNA:	premature-messenger RNA
PTC:	Premature Termination Codon
PTP:	picotiter plate
RSS:	RNA Storage Solution

RT:	Reverse Transcription
smORF:	small - Open Reading Frame
SMRT:	Single-Molecule Real-Time Sequencing
SNP:	Single Nucleotide Polymorphism
snRNP:	small nuclear Ribonucleoproteins
SNV:	Single Nucleotide Variant
SRA:	Sequence Read Archive
T_a:	Temperature annealing
TGS:	Third Generation Sequencing
T_m:	Temperature melting
UV-Vis:	Ultraviolet-Visible
WES:	Whole Exome Sequencing
WGS:	Whole Genome Sequencing,
WTS:	Whole Transcriptome Sequencing
ZMW	Zero-Mode Waveguide

1. ΕΙΣΑΓΩΓΗ

Ο κλάδος της Βιοχημείας και Μοριακής Βιολογίας αποτελεί ένα σχετικά νέο πεδίο των θετικών επιστημών, το οποίο, τις τελευταίες δεκαετίες, αναπτύχθηκε τάχιστα και επέτρεψε την ανακάλυψη των βασικών δομικών συστατικών των οργανισμών, όπως τα νουκλεϊκά οξέα και οι πρωτεΐνες. Ως ερευνητικό πεδίο, η μελέτη του DNA, άρχισε να αναπτύσσεται από τα μέσα του 20ου αιώνα με την ανακάλυψη τόσο της δομής του όσο και θεμελιωδών ιδιοτήτων του. Ωστόσο, ήδη από τον προηγούμενο αιώνα, είχαν προηγηθεί μελέτες γύρω από το DNA ύστερα από την ανακάλυψη της ύπαρξης του δεοξυριβονουκλεϊκού οξέος. Το 1869, ο βιοχημικός Johann Friedrich Miescher πιστοποίησε την ύπαρξη μιας ουσίας με συγκεκριμένη όξινη αντίδραση μέσα στους πυρήνες κυττάρων, την οποία και ονόμασε νουκλεΐνη θεωρώντας ότι πρόκειται για μια νέα πρωτεΐνη [1-3]. Επόμενα πειράματα οδήγησαν στο συμπέρασμα ότι το μόριο αυτό δεν περιλαμβάνει θείο, γεγονός που υποδεικνύει ότι δεν πρόκειται για πρωτεΐνη. Το 1889, ο Richard Altmann απομόνωσε από τα κύτταρα μια νέα ουσία την οποία ονόμασε νουκλεϊκό οξύ, καθώς συμπεριφερόταν ως οξύ στις χημικές αντιδράσεις, η οποία, λίγα χρόνια νωρίτερα, αποδείχθηκε ότι ήταν το ίδιο ακριβώς μόριο με την νουκλεΐνη που είχε απομονωθεί από τον Miescher [2, 4]. Το 1893, οι Γερμανοί βιοχημικοί Albrecht Kossel και Albert Neumann απέδειξαν ότι τέσσερις διαφορετικές αζωτούχες βάσεις αποτελούν τα βασικά δομικά στοιχεία, τα οποία συγκροτούν το DNA, και στη συνέχεια ο Kossel παρατήρησε πως η χρωματίνη, η οποία απαρτίζει τα χρωμοσώματα, αποτελείται από μόρια πρωτεϊνών, τα οποία αλληλεπιδρούν με το DNA.

Στις αρχές του 20ου αιώνα ήταν ήδη γνωστό ότι το DNA είναι τμήμα των χρωμοσωμάτων. Η θεωρία ότι τα γονίδια βρίσκονται στα πυρηνικά χρωμοσώματα των κυττάρων δημιουργήθηκε από τους Theodor H. Boveri και Walter S. Sutton, στις αρχές του 1900, ενώ την δεκαετία του 1910, ο Αμερικανός γενετιστής Thomas Hunt Morgan απέδειξε την θεωρία της κληρονομικότητας των χρωμοσωμάτων, η οποία και επιβεβαιώθηκε, στη συνέχεια, από άλλους ερευνητές μέσω της χαρτογράφησης των χρωμοσωμάτων διάφορων οργανισμών [5]. Τις επόμενες δεκαετίες δημιουργήθηκαν οι πρώτες ενδείξεις ότι το γενετικό υλικό, που υπάρχει στα χρωμοσώματα και μεταβιβάζεται από γενιά σε γενιά, είναι το DNA και όχι οι πρωτεΐνες [4]. Το 1944 οι Avery, Macleod και McCarty απέδειξαν ότι το DNA είναι

το μόριο που μεταφέρει τη γενετική πληροφορία [6], γεγονός που επαληθεύτηκε το 1952 από τα πειράματα των Hershey και Chase [7]. Σήμερα είναι ευρέως γνωστό πως το DNA είναι το γενετικό υλικό που φέρει την γενετική πληροφορία όλων των έμβιων όντων με εξαίρεση τους RNA ιούς.

Στο τέλος της δεκαετίας του 1940, ο Erwin Chargaff και οι συνεργάτες του απέδειξαν ότι, σε ένα μόριο DNA, ο αριθμός των βάσεων πουρίνης είναι ίσος με τον αριθμό των πυριμιδινών, και αντίστροφα. Πιο συγκεκριμένα, η ποσότητα της θυμίνης είναι πάντα ίση με την ποσότητα αδενίνης και η ποσότητα της γουανίνης αντιστοιχεί στην ποσότητα της κυτοσίνης. Οι παραπάνω μελέτες, σε συνδυασμό με φυσικοχημικές και x-ray κρυσταλλογραφικές μελέτες από τους Wilkins [8] τους Franklin και Gosling [9] και τους James D. Watson και Francis H. C. Crick, έθεσαν τα θεμέλια για την ανακάλυψη της δομής του DNA, το 1953 [10]. Οι Watson και Crick πρότειναν το μοντέλο της διπλής έλικας του DNA, το οποίο πληρεί τις κατάλληλες προϋποθέσεις για τον χαρακτηρισμό του μορίου ως γενετικό υλικό, στο οποίο στηρίζεται η ζωή και η βιολογική εξέλιξη των οργανισμών [10, 11]. Το πληροφοριακό περιεχόμενο του DNA βασίζεται στη γραμμική του δομή, η οποία χαρακτηρίζεται ως νουκλεοτιδική αλληλουχία, και αποτελεί τη γενετική πληροφορία. Η ανακάλυψη της δομής του DNA σηματοδότησε την αρχή της ανάπτυξης ενός νέου κλάδου της επιστήμης, της Μοριακής Βιολογίας, ο οποίος εξελίχθηκε ραγδαία ύστερα από την ανακάλυψη του μηχανισμού της αντιγραφής του DNA, της θεωρίας του γενετικού κώδικα και την συνεχιζόμενη ανάπτυξη τεχνολογιών που αφορούν στη μελέτη των γονιδιωμάτων των οργανισμών.

Μεταξύ των δεκαετιών 1960 και 1970, μετά την ανακάλυψη των περιοριστικών ενζύμων στα βακτήρια και της ικανότητάς τους να διασπούν το DNA σε συγκεκριμένες θέσεις, έγινε δυνατή η απομόνωση, η ενίσχυση με κλωνοποίηση, και η μεταφορά γονίδιων από το ένα είδος στο άλλο [12]. Επιπλέον, νέες τεχνικές κατέστησαν δυνατή την αλληλούχηση του DNA, δηλαδή την ανάλυση της πρωτοταγούς δομής του DNA που επιτυγχάνεται με τον ακριβή προσδιορισμό της σειράς των βάσεων του DNA μιας περιοχής (γονίδιο, χρωμόσωμα, γονιδίωμα). Η μοριακή βιολογία, η βιοτεχνολογία, η γενετική καθώς και πλήθος άλλων επιστημών αποτελούν βασικά πεδία, στα οποία οι τεχνικές αλληλούχησης του DNA χρησιμοποιούνται ως ερευνητικό εργαλείο για τη μελέτη του γονιδιώματος των οργανισμών και κατ’ επέκταση την κατανόηση των πολύπλοκων μηχανισμών που

διέπουν τους οργανισμούς [13]. Η αλληλούχηση του DNA επέτρεψε την ανάλυση των βιοχημικών δομών μεμονωμένων γονιδίων διάφορων οργανισμών, συμπεριλαμβανομένου του ανθρώπου, και κατά συνέπεια, την πλήρη αλληλούχηση του ανθρώπινου γονιδιώματος, δίνοντας την ευκαιρία διεύρυνσης των γενετικών παραγόντων που προκαλούν ασθένειες, όπως ο καρκίνος.

Έως σήμερα, οι τεχνολογίες αλληλούχησης, οι οποίες έχουν αναπτυχθεί, διαφέρουν μεταξύ τους ως προς τη μεθοδολογία που ακολουθείται για τον προσδιορισμό της νουκλεοτιδικής ακολουθίας, καθώς και ως προς το πληροφοριακό περιεχόμενο του αποτελέσματος (output) το οποίο παρέχουν. Οι ήδη υπάρχουσες μέθοδοι αλληλούχησης μπορούν να ταξινομηθούν σε τρεις κύριες κατηγορίες, οι οποίες είναι: α) οι βασικές μέθοδοι αλληλούχησης, κατά Sanger και κατά Maxam-Gilbert, β) η αλληλούχηση επόμενης ή / και δεύτερης γενιάς και γ) η αλληλούχηση τρίτης γενιάς, καθεμία από τις οποίες θα αναλυθεί διεξοδικά, στη συνέχεια. Για περίπου 30 χρόνια, η μέθοδος τερματισμού αλυσίδας, που αναπτύχθηκε από τον Frederick Sanger, αποτελούσε την κυρίαρχη μέθοδο αλληλούχησης και χρησιμοποιήθηκε για την αλληλούχηση του ανθρώπινου γονιδιώματος (Human Genome Project) [14]. Το 2005, παρουσιάστηκαν οι πρώτοι αλληλουχητές επόμενης γενιάς, οι οποίοι επέτρεψαν τη μαζική παράλληλη αλληλούχηση εκατοντάδων τμημάτων DNA, ενώ τα τελευταία χρόνια εμφανίστηκαν μεθοδολογίες αλληλούχησης τρίτης γενιάς, οι οποίες παρέχουν γενετικές πληροφορίες πολύ υψηλής ανάλυσης.

1.1. Βασικές μέθοδοι αλληλούχησης

Από τα μέσα της δεκαετίας του 1970 ήταν εφικτή τόσο η απομόνωση συγκεκριμένων τμημάτων του DNA, όπως γονιδίων, όσο και η ενίσχυσή τους σε πολλαπλά αντίγραφα, με σκοπό την ανάλυση τους, ενώ δεν ήταν δυνατος ο προσδιορισμός της νουκλεοτιδικής ακολουθίας του DNA. Ωστόσο, μικρά μόρια tRNA, το μήκος των οποίων δεν ξεπερνά τα 75-80 νουκλεοτίδια, είχαν προσδιοριστεί ήδη από τη δεκαετία του 1960. Η πρώτη μέθοδος αλληλούχησης νουκλεϊκών οξέων, που επέτρεψε τον άμεσο προσδιορισμό τμημάτων DNA μήκους μεταξύ 100 και 500 νουκλεοτιδίων, αναπτύχθηκε από τον Frederick Sanger, το 1975. Περίπου το 1976, πραγματοποιήθηκε η ανάπτυξη δύο νέων μεθόδων, οι οποίες προσέφεραν την δυνατότητα αποκωδικοποίησης μεγάλου αριθμού βάσεων

σε σχετικά μικρό χρονικό διάστημα. Και οι δύο μέθοδοι, η διαδικασία τερματισμού αλυσίδας, που αναπτύχθηκε από τους Sanger και Coulson, και η τεχνική της χημικής διάσπασης, που αναπτύχθηκε από τους Maxam και Gilbert, χρησιμοποιούν ραδιενεργά σημασμένα μόρια κατά μήκος του DNA, σε καθορισμένες θέσεις, οι οποίες τελικά καταλαμβάνονται από την αντίστοιχη βάση που ορίζει η πρωτοταγής αλληλουχία του DNA, με τελικό αποτέλεσμα τον προσδιορισμό της σειράς των νουκλεοτιδίων στο DNA [14, 15].

1.1.1. Χημική Μέθοδος ή Μέθοδος κατά Maxam-Gilbert

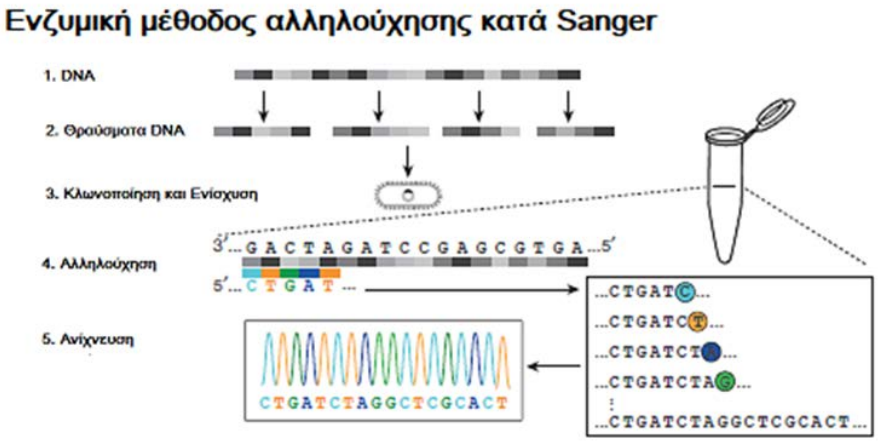
Η μέθοδος αλληλούχησης κατά Maxam-Gilbert βασίζεται στη χημική τροποποίηση των αλυσίδων DNA, που ακολουθείται από τον κατακερματισμό του DNA σε συγκεκριμένα κατάλοιπα βάσεων. Σύμφωνα με τη μέθοδο αυτή, τα μόρια DNA σημαίνονται ραδιενεργά στο 5’ άκρο και απομονώνονται θραύσματα DNA έτοιμα προς αλληλούχηση [13, 16]. Πραγματοποιούνται συνολικά τέσσερεις αντιδράσεις, οι οποίεις είναι: α) στο νουκλεοτίδιο της γουανίνης (αντίδραση G), β) στα νουκλεοτίδια των πουρινών (αντίδραση G και A), γ) στα νουκλεοτίδια των πυριμιδινών (αντίδραση C και T), και δ) στο νουκλεοτίδιο της κυτοσίνης (αντίδραση C), κάθε μία από τις οποίες γίνεται με διαφορετική χημική κατεργασία. Η διαδικασία αποτελείται από τρία βήματα: α) τη τροποποίηση της αζωτούχου βάσης, β) την αφαίρεση της τροποποιημένης βάσης από τη δεοξυριβόζη, και, τελικά, γ) τη θραύση της ακολουθίας στη συγκεκριμένη δεοξυριβόζη. Αποτέλεσμα των χημικών αυτών τροποποιήσεων είναι η δημιουργία σημασμένων θραυσμάτων από το ραδιοσημασμένο 5’ άκρο ως το σημείο της θραύσης σε κάθε μόριο. Μετά την ολοκλήρωση της επεξεργασίας, τα τμήματα DNA διαχωρίζονται με βάση το μέγεθος, μέσω ηλεκτροφόρησης, σε πήκτωμα πολυακρυλαμίδης και ανιχνεύονται μέσω αυτοραδιογραφίας, με τελικό αποτέλεσμα τον προσδιορισμό της ακολουθίας [13, 17].

1.1.2. Ενζυμική Μέθοδος ή Μέθοδος κατά Sanger

Η ενζυμική μέθοδος αλληλούχησης κατά Sanger προσφέρει τη δυνατότητα προσδιορισμού της πρωτοταγούς δομής μιας νουκλεοτιδικής αλληλουχίας με μέγιστο δυνατό μήκος προσδιορισμού τις 1000 βάσεις. Η μέθοδος στηρίζεται στην ενσωμάτωση και ανίχνευση ενός νουκλεοτιδίου στην επιμηκυνόμενη αλυσίδα του

DNA, με τη δράση ενζύμων [18]. Η αρχή μεθόδου βασίζεται στον υβριδισμό ενός συνθετικού εκκινητή στη μονόκλωνη DNA αλυσίδα. Η αντίδραση απαιτεί DNA πολυμεράση, τριφωσφορικά ολιγονουκλεοτίδια (dNTPs, -A,G,C,T-), καθώς και, σε μικρή αναλογία, φθοριοσημασμένα 2',3' διδεοξυνουκλεοτίδια (ddNTPs). Λόγω της ύπαρξης ενός υδρογόνου στον 3' άνθρακα της δεοξυριβόζης αντί ενός υδροξυλίου, η δομή των ddNTPs δεν επιτρέπει την περαιτέρω δημιουργία φωσφοδιεστερικών δεσμών στο 3' άκρο, και επομένως, η ενσωμάτωση του ddNTP σηματοδοτεί και την λήξη της αντίδρασης. Κατά τη διάρκεια της αντίδρασης, τα dNTPs βρίσκονται σε περίσσεια, και επομένως, τα ddNTPs ενσωματώνονται σε τυχαίες θέσεις στο DNA, με αποτέλεσμα την δημιουργία τμημάτων DNA διαφορετικού μήκους.

Ο διαχωρισμός των μορίων DNA, που προκύπτουν, γίνεται με τη διαδικασία της ηλεκτροφόρησης. Μετά το τέλος της διαδικασίας, σε κάθε σημασμένο μόριο DNA αναγνωρίζεται η χρωστική, επομένως και η αντίστοιχη βάση, που φέρει κάθε προϊόν στο 3' άκρο του, επιτρέποντας τον προσδιορισμό της σειράς των βάσεων στη νουκλεοτιδική αλληλουχία [19, 20]. Από την δεκαετία του 1970 έως σήμερα, η μέθοδος αλληλούχησης κατά Sanger έχει υποστεί αρκετές τροποποιήσεις και βελτιώσεις, οι οποίες συνέβαλαν όχι μόνο στην αυτοματοποίηση της ίδιας της μεθόδου, αλλά και στην ανάπτυξη νέων μεθοδολογιών αλληλούχησης του DNA, οι οποίες χρησιμοποιούνται σε ευρεία κλίμακα, τόσο για τον προσδιορισμό του γονιδιώματος πολύπλοκων οργανισμών, όπως ο άνθρωπος, όσο και σε περιπτώσεις μελέτης συγκεκριμένων χρωμοσωμικών περιοχών, όπως, για παράδειγμα, στην ανίχνευση της ύπαρξης μεταλλάξεων σε ένα γονίδιο [20].



Εικόνα 1. Τα 5 βασικά βήματα της αλληλούχησης κατά Sanger είναι: α) η επιλογή του DNA, β) η θραύση του σε μικρότερα τμήματα, γ) η ενίσχυση του δείγματος, δ) η διαδικασία αλληλούχησης, και ε) η ανίχνευση του σήματος. Το «κλειδί» της μεθοδολογίας, στην

αλληλούχηση κατά Sanger, είναι η χρήση των ddNTPs. Σημασμένα ddNTPs χρησιμοποιούνται για τον τερματισμό της σύνθεσης της DNA αλυσίδας και προκύπτει ένα σύνολο θραυσμάτων DNA, το οποίο ηλεκτροφορείται. Το οπτικό σήμα, που παράγεται, αντιστοιχεί σε συγκεκριμένη βάση και επιτυγχάνεται ο προσδιορισμός της νουκλεοτιδικής ακολουθίας (επεξεργασία εικόνας από [15]).

1.1.3. Shotgun αλληλούχηση (advanced DNA sequencing)

Οι αλληλουχητές πρώτης γενιάς, που χρησιμοποιούν την μέθοδο αλληλούχησης κατά Sanger, έχουν την ικανότητα προσδιορισμού θραυσμάτων DNA, τα οποία δεν υπερβαίνουν σε μήκος τις 1000 βάσεις [20]. Το 1979, ο Staden πρότεινε την αλληλούχηση shotgun για την ανάλυση τμημάτων DNA μεγαλύτερου μήκους [21]. Σύμφωνα με την shotgun αλληλούχηση, θραύσματα DNA, τα οποία εμφανίζουν επικαλυπτόμενα άκρα κλωνοποιούνται, έπειτα, ξεχωριστά για κάθε θραύσμα, προσδιορίζεται η νουκλεοτιδική αλληλουχία, και στη συνέχεια, πραγματοποιείται συναρμολόγηση των τμημάτων (assembly), με βάση τα επικαλυπτόμενα άκρα των τμημάτων αυτών, με τελικό αποτέλεσμα τον προσδιορισμό της ακολουθίας των βάσεων του ολικού DNA [15, 20].

1.2. Αλληλούχηση Επόμενης Γενιάς (Next-generation sequencing, NGS)

Η βελτίωση των βασικών μεθόδων αλληλούχησης, καθώς και η περαιτέρω ανάπτυξη νέων τεχνικών προσδιορισμού νουκλεοτιδικών αλληλουχιών, ενισχύθηκε με την βοήθεια της αλυσιδωτής αντίδρασης πολυμεράσης (PCR) και τις τεχνολογίες του ανασυνδυασμένου DNA [22]. Βάσει των γνωστών μεθόδων αλληλούχησης, κατασκευάστηκαν νεότεροι αλληλουχητές, όπως ο ABI PRISM από την Applied Biosystems, ο οποίος επέτρεψε την ταυτόχρονη αλληλούχηση εκατοντάδων δειγμάτων DNA, και χρησιμοποιήθηκε στο Πρόγραμμα του Ανθρώπινου γονιδιώματος (Human Genome Project, HGP) [23].

Από το 1980 έως και το 2000, αρκετές ερευνητικές ομάδες ασχολήθηκαν με την αναζήτηση νέων μεθόδων αλληλούχησης [15]. Το 2001, η ολοκλήρωση της ανάλυσης του προγράμματος του ανθρώπινου γονιδιώματος, οδήγησε στην ανάγκη εύρεσης νέων τεχνικών αλληλούχησης, με στόχο τον προσδιορισμό μεγάλων τμημάτων DNA, όπως για παράδειγμα ολόκληρων γονιδιωμάτων, με αποτέλεσμα την ελαχιστοποίηση τόσο του απαιτούμενου χρόνου όσο και του κόστους [24]. Η Αλληλούχηση Επόμενης Γενιάς (Next-generation sequencing, NGS) χαρακτηρίζεται

ως τεχνολογία υψηλής απόδοσης, η οποία επιτρέπει την ανάλυση εκατοντάδων θραυσμάτων DNA στον ίδιο χρόνο, και είναι γνωστή ως Μαζική Παράλληλη Αλληλούχηση [25]. Το βασικό χαρακτηριστικό της αλληλούχησης επόμενης γενιάς είναι η παράλληλη χρήση πολλαπλών κλώνων / εκμαγείων (templates) και ο προσδιορισμός της ακολουθίας τους, βασισμένος στις ιδιότητες της αντιγραφής του γενετικού υλικού [14, 24].

Οι πλατφόρμες αλληλούχησης επόμενης γενιάς, που έχουν αναπτυχθεί, περιλαμβάνουν βασικά κοινά βήματα, που αφορούν στην προετοιμασία του εκμαγείου, στον εμπλουτισμό του δείγματος- στόχου, και στην τελική αντίδραση αλληλούχησης, η οποία ακολουθείται από βιοπληροφορική ανάλυση των αποτελεσμάτων. Η κατασκευή NGS βιβλιοθήκης αναφέρεται στις διαδικασίες προετοιμασίας του DNA (πχ. θραύση του DNA, επιδιόρθωση των άκρων, πρόσδεση ανταπτόρων), ώστε το δείγμα να είναι κατάλληλο προς αλληλούχηση. Ο εμπλουτισμός του στόχου περιλαμβάνει αντιδράσεις όπου αυξάνουν την ειδικότητα της όλης διαδικασίας και η αντίδραση αλληλούχησης είναι το τελευταίο βήμα, το οποίο επιτρέπει τον προσδιορισμό της ακολουθίας των βάσεων στο DNA - στόχο [26]. Κάθε πλατφόρμα αλληλούχησης χρησιμοποιεί διαφορετικές τεχνικές και διαθέτει ειδικό εξοπλισμό, τα επίπεδα απόδοσης της κάθε μεθοδολογίας είναι διαφορετικά, όπως επίσης, και ο αριθμός των πειραματικών αλληλουχιών (reads), που προκύπτουν, και ως επακόλουθο, το απαιτούμενο κόστος, για κάθε αντίδραση αλληλούχησης, διαφέρει [27, 28].

Οι περισσότερες μέθοδοι αλληλούχησης του DNA, που χρησιμοποιούνται κυρίως στα εργαστήρια Βιοχημείας και Μοριακής Βιολογίας, βασίζονται στη σύνθεση νέων αλυσίδων DNA, χρησιμοποιώντας ως εκμαγείο ενισχυμένο PCR προϊόν [29]. Η διαδικασία της σύνθεσης συμπληρωματικών DNA αλυσίδων αποτελεί τη πιο κοινή μέθοδο αλληλούχησης, όπου ο ενισχυμένος κλώνος του DNA, που έχει επιλεγεί προς αλληλούχηση ύστερα από ειδική επεξεργασία, πολυμερίζεται, με τελικό στόχο την ανίχνευση σήματος ύστερα από κάθε προσθήκη, συμπληρωματικής ως προς το εκμαγείο, βάσης. Η αρχή μεθόδου, που αναφέρθηκε, αποτελεί την βασική αρχή των μεθοδολογιών που χρησιμοποιούνται στις περισσότερες πλατφόρμες αλληλούχησης επόμενης γενιάς, συμπεριλαμβανομένων των βασικότερων αντιπροσώπων Roche™, Illumina® και Ion Torrent™ [30].

1.2.1. Τεχνολογία αλληλούχησης Roche™ / 454 Πυροαλληλούχηση™

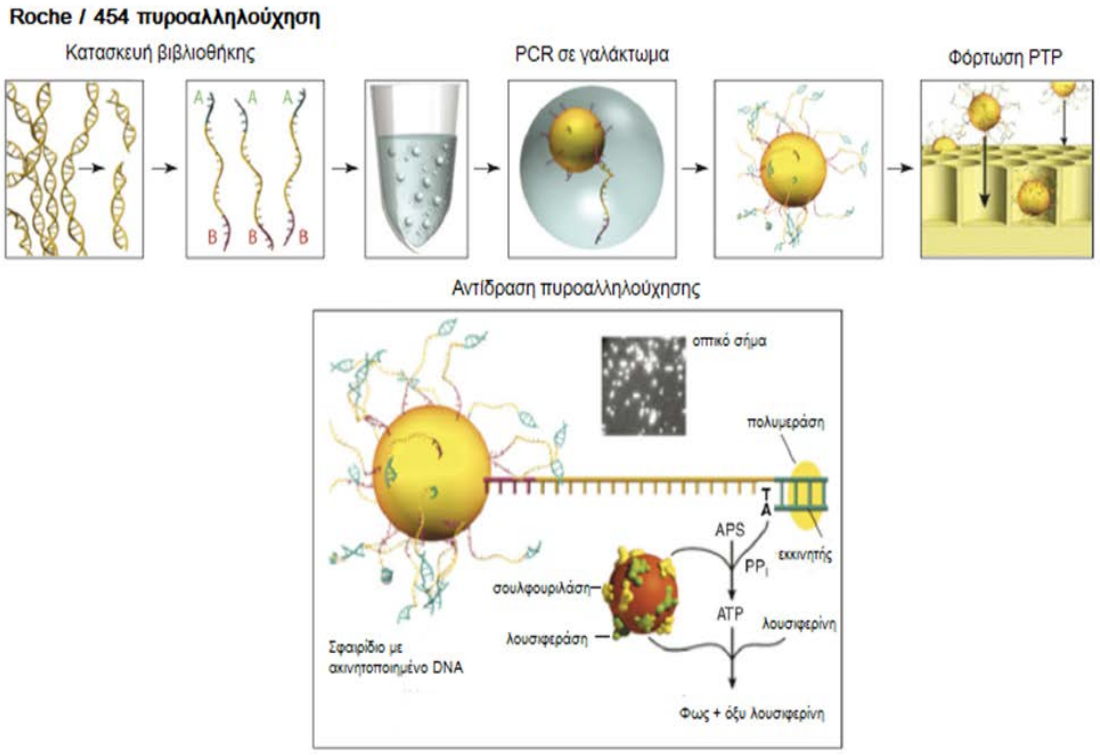
Η πυροαλληλούχηση είναι μία διαδεδομένη μέθοδος αλληλούχησης μέσω σύνθεσης, η οποία αποτελείται από ένα καταρράκτη δύο αντιδράσεων. Η πρώτη αντίδραση βασίζεται στην πρόσδεση νουκλεοτιδίου στο νεοσυντιθέμενο DNA, με ταυτόχρονη απελευθέρωση πυροφωσφορικού (PP_i), ενώ η δεύτερη περιλαμβάνει την εκπομπή ορατού φωτός, που μεταφράζεται σε σήμα στον αλληλουχητή.

Πιο αναλυτικά, αρχικά γίνεται επιλογή του DNA προς αλληλούχηση και η κατασκευή DNA βιβλιοθήκης. Γίνεται διαδοχικά, θραύση του DNA σε μικρότερα τμήματα, επεξεργασία των άκρων, ώστε τελικά να είναι κατάλληλα για την πρόσδεση προσαρμογέων (ή ανταπτόρων, adapters), με τη δράση λιγάσης. Το DNA ακινητοποιείται σε ειδικά σφαιρίδια αγαρόζης, τα οποία φέρουν ολιγονουκλεοτίδια συμπληρωματικά με την αλληλουχία των ανταπτόρων, που βρίσκεται στα άκρα του μορίου, το οποίο πρόκειται να αλληλουχηθεί. Τα τμήματα DNA ενισχύονται μέσω αλυσιδωτής αντίδρασης πολυμεράσης σε γαλάκτωμα (emulsion PCR, emPCR) μια διαδικασία που συμβαίνει σε μικροαντιδραστήρες. Κάθε θραύσμα της βιβλιοθήκης ενισχύεται στην επιφάνεια ενός σφαιριδίου που βρίσκεται σε μικροαντιδραστήρα, ο οποίος περιλαμβάνει όλα τα αντιδραστήρια που απαιτούνται για την πραγματοποίηση μιας αντίδρασης PCR. Στο τέλος της αντίδρασης, τα PCR προϊόντα που δημιουργούνται αποτελούνται από εκατομμύρια αντίγραφα του ίδιου θραύσματος, τα οποία καλύπτουν την επιφάνεια του σφαιριδίου.

Στη συνέχεια, γίνεται ανάκτηση των ενισχυμένων σφαιριδίων και ακολουθείται το βήμα του εμπλουτισμού (enrichment). Σε αυτό το στάδιο πραγματοποιείται ο διαχωρισμός των σφαιριδίων που περιέχουν τα επιτυχώς ενισχυμένα τμήματα της βιβλιοθήκης από τα σφαιρίδια, τα οποία είτε δεν έχουν ενσωματώσει κάποιο θραύσμα DNA είτε δεν έχει επιτευχθεί σε αυτά η αντίδραση PCR. Στη συνέχεια, το ενισχυμένο PCR προϊόν, που είναι ακινητοποιημένο στα σφαιρίδια, αποδιατάσσεται, με σκοπό τη δημιουργία μονόκλωνων αλυσίδων, οι οποίες θα υβριδοποιηθούν με κατάλληλο εκκινητή. Τα DNA θραύσματα είναι, πλέον, κατάλληλα ώστε να προσδιοριστούν με την μέθοδο αλληλούχησης μέσω σύνθεσης. Τα σφαιρίδια τοποθετούνται σε κατάλληλη πλάκα (picotiter plate, PTP), σχεδιασμένη ώστε να διαθέτει περισσότερα από ένα εκατομμύριο πηγάδια (wells) [31]. Κάθε ένα από τα πηγάδια αυτά μπορεί να ενσωματώσει ένα μόνο σφαιρίδιο.

Στη συνέχεια, η πλάκα επωάζεται με ένζυμα (DNA πολυμεράση, ATP σουλφουριλάση, λουσιφεράση και απυράση) και υποστρώματα (5’ φωσφοθειική αδενοσίνη και λουσιφερίνη). Διαλύματα των 4 νουκλεοτιδίων προστίθενται διαδοχικά στο ακινητοποιημένο DNA και μετά το πέρας της εκάστοτε αντίδρασης γίνεται έκπλυση αυτών στο σημείο της αντίδρασης [30].

Η πυροαλληλούχηση ξεκινάει με την προσθήκη του πρώτου διαλύματος νουκλεοτιδίων στην αντίδραση. Εάν το νουκλεοτίδιο, που προστίθεται, είναι συμπληρωματικό με την αντίστοιχη βάση, που βρίσκεται στην ίδια θέση στο εκμαγείο, τότε το νουκλεοτίδιο αυτό θα ενσωματωθεί στο νεοσυντιθέμενο κλώνο, με αποτέλεσμα την απελευθέρωση πυροφωσφορικού [32, 33]. Στη συνέχεια, το PP_i μετατρέπεται σε ATP, μια αντίδραση που καταλύεται από το ένζυμο ATP σουλφουρυλάση. Το μόριο ATP χρησιμοποιείται από την λουσιφεράση για την μετατροπή της λουσιφερίνης σε οξυλουσιφερίνη και την παραγωγή ορατού φωτός μέσω βιοφωταύγειας. Το ορατό φως, που εκπέμπεται από την αντίδραση πυροαλληλούχησης, ανιχνεύεται χρησιμοποιώντας συζευγμένα, με φορτίο, στοιχεία και, τελικά, εμφανίζεται ως κορυφή στο Pyrogram, το γράφημα που παράγεται από τον αλληλουχητή [34]. Το ποσό του εκπεμπόμενου φωτός, που δημιουργείται, είναι ανάλογο με τον αριθμό των νουκλεοτιδίων, τα οποία ενσωματώθηκαν στην αλυσίδα [35]. Εάν το νουκλεοτίδιο, που προστέθηκε, δεν είναι συμπληρωματικό με εκείνο που βρίσκεται στην αντίστοιχη θέση στο εκμαγείο, οι δύο αντιδράσεις δεν πραγματοποιούνται, και επομένως, δεν παράγεται σήμα. Τα μη ενσωματωμένα νουκλεοτίδια και το ATP αποικοδομούνται με τη δράση του ενζύμου απυράση. Μετά το τέλος της αποικοδόμησης, η διαδικασία επαναλαμβάνεται με την προσθήκη του επόμενου διαλύματος νουκλεοτιδίων [30, 33].



Εικόνα 2. Στο διάγραμμα ροής παρουσιάζονται τα βήματα, που ακολουθούνται κατά την πυροαλληλούχηση. Κατασκευάζεται βιβλιοθήκη με θραύσματα του DNA – στόχου, με τελικό βήμα την πρόσδεση ανταπτόρων στα δύο άκρα κάθε θραύσματος. Ακολουθεί η υβριδοποίηση των ολιγονουκλεοτιδίων του σφαιριδίου με το εκμαγείο και, στη συνέχεια, η PCR σε γαλάκτωμα. Επιλέγονται τα σφαιρίδια που φέρουν τις επιθυμητές αλληλουχίες DNA (enrichment) και γίνεται φόρτωμα του δείγματος σε πηγάδια, όπου τελικά, διεξάγονται οι αντιδράσεις της πυροαλληλούχησης. (Επεξεργασία από [36]).

Η πυροαλληλούχηση, μέσω του συστήματος Roche / 454™, δίνει την δυνατότητα αλληλούχησης περισσότερων του ενός εκατομμυρίου αλυσίδων DNA, και την παραγωγή πειραματικών αλληλουχιών με μήκος 400-500 bp, σε περίπου 10 ώρες [30]. Η συγκεκριμένη μέθοδος αλληλούχησης θεωρείται κατάλληλη για ανίχνευση SNPs, νησίδων CpG και ιδανική, για τη μελέτη μεθυλίωσης υποκινητών. Ωστόσο, βασικό μειονέκτημα της πυροαλληλούχησης είναι η αδυναμία του συστήματος να αναγνωρίζει επιτυχώς την ένταση του σήματος που εκπέμπεται σε κάθε αντίδραση και, τελικά, να οδηγείται σε αυξημένο αριθμό σφαλμάτων, που μπορεί να περιλαμβάνουν την λανθασμένη ενσωμάτωση κάποιας βάσης έναντι της σωστής, αλλά κυρίως την απαλοιφή ή / και την προσθήκη βάσεων, γεγονός που υποβαθμίζει την ποιότητα των πειραματικών αλληλουχιών [30].

1.2.2. Μεθοδολογία αλληλούχησης Illumina®

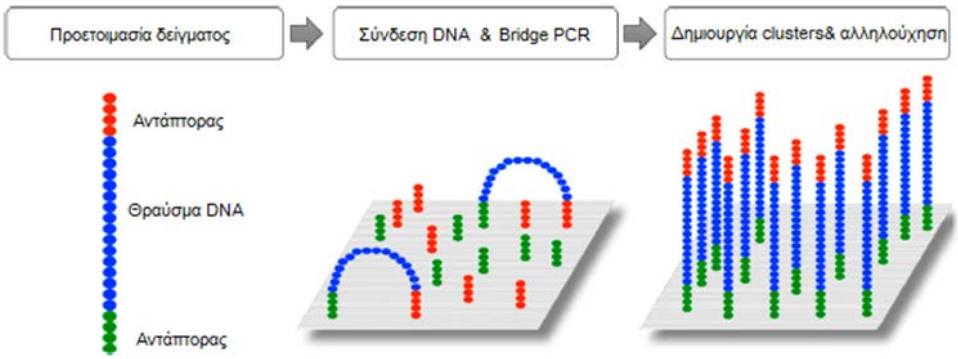
Η μεθοδολογία αλληλούχησης της Illumina® περιλαμβάνει 3 βασικά στάδια, τα οποία είναι η ενίσχυση του δείγματος, η αντίδραση αλληλούχησης και η βιοπληροφορική ανάλυση των δεδομένων. Η τεχνολογία, που χρησιμοποιεί, βασίζεται στην ανίχνευση, μέσω φθορισμού, χρωμοφόρων νουκλεοτιδίων, τα οποία συμπεριφέρονται ως ανάστροφοι τερματιστές (reversible terminators) σε κάθε αντίδραση ενίσχυσης και ενσωματώνονται σε νεοσυντιθέμενες αλυσίδες DNA, επιτρέποντας των προσδιορισμό μεμονωμένων βάσεων [37, 38]. Η βασική ιδέα της μεθοδολογίας είναι παρόμοια με την αντίστοιχη της αλληλούχησης κατά Sanger με την διαφορά ότι, στην Illumina®, κάθε βάση, που φέρει διαφορετικό φθοριόχρωμα, δίνει την δυνατότητα αναστρέψιμου τερματισμού του κλώνου, ενώ στην μέθοδο Sanger, ο πολυμερισμός χαρακτηρίζεται ως μη αναστρέψιμος [27]. Μια βασική διαφορά με τις υπόλοιπες τεχνικές είναι η διαδικασία της κλωνικής ενίσχυσης που χρησιμοποιείται, γνωστή ως ενίσχυση μέσω γέφυρας (bridge amplification), με σκοπό τον πολλαπλασιασμό του αριθμού των μορίων, που πρόκειται να αλληλουχηθούν.

Τα αρχικά βήματα της διαδικασίας είναι η απομόνωση του DNA, η θραύση του σε μικρότερα τμήματα (fragments) και, στη συνέχεια η τροποποίηση των άκρων και η πρόσδεση, σε αυτά, ανταπτόρων, οι οποίοι είναι απαραίτητοι για την αλληλούχηση και την βιοπληροφορική ανάλυση. Στη συνέχεια, διεξάγεται ο πολλαπλασιασμός των κλώνων με την βοήθεια της PCR μέσω γέφυρας [27]. Στο τέλος της αντίδρασης ενίσχυσης, κάθε cluster αποτελείται από περίπου 1000 πανομοιότυπα μόρια, τα οποία προέρχονται από ένα θραύσμα DNA [39]. Σε κάθε κύκλο, τα 4 νουκλεοτίδια προστίθενται ταυτόχρονα και είναι χημικά «μπλοκαρισμένα», έχοντας υποκαταστήσει την 3’-OH ομάδα του δακτυλίου με μια ομάδα 3’-ο-αζυδομεθυλίου. Η αλλαγή αυτή αποτρέπει την ενσωμάτωση στην αλυσίδα περισσότερων του ενός νουκλεοτιδίου τη φορά, και επομένως, κάθε ένα νουκλεοτίδιο προστίθενται, με την βοήθεια της DNA πολυμεράσης, στη σωστή θέση του νεοσυντιθέμενου μορίου, σύμφωνα με τον κανόνα της συμπληρωματικότητας του DNA [20, 27]. Στο τέλος κάθε κύκλου σύνθεσης, προσδιορίζεται η βάση, που προστέθηκε, με την χρήση H/Y, ανιχνεύοντας το σήμα φθορισμού με βάση το μήκος κύματος, και γίνεται καταγραφή του σήματος κατά μήκος του chip [40]. Τα μη ενσωματωμένα νουκλεοτίδια, οι χημικά τροποποιημένες ομάδες και τα χρωμοφόρα

απομακρύνονται, με έκπλυση, από την επιφάνεια του chip [41]. Μόλις ανιχνευθεί το φθορίζον σήμα ακολουθεί ο επόμενος κύκλος σύνθεσης έως ότου ολοκληρωθεί η πλήρης αλληλούχηση του κάθε θραύσματος [42].

Το πρώτο στάδιο για την κατασκευή μιας NGS βιβλιοθήκης (NGS library), που θα χρησιμοποιηθεί σε μια αντίδραση αλληλούχησης μέσω της μεθοδολογίας της Illumina®, είναι η θραύση του DNA σε μικρότερα τμήματα, μέσω αντιδράσεων, οι οποίες καταλύονται από τη δράση ενζύμων και οδηγούν στην διάσπαση του DNA σε τυχαίες θέσεις. Τα τμήματα που προκύπτουν, χαρακτηρίζονται από ατελή άκρα (blunt ends), τα οποία επιδιορθώνονται, ώστε, στη συνέχεια, να προσδεθούν οι αντάπτορες. Οι αντάπτορες είναι μονόκλωνες ή δίκλωνες, μικρού μήκους, αλυσίδες DNA που προσδένονται στο DNA, με φωσφοδιεστερικούς δεσμούς, μέσω της δράσης λιγάσης. Αποτελούν βασικό συστατικό για τη χημεία της αντίδρασης, καθώς οριοθετούν το DNA (insert) και, επιπλέον, μπορεί να περιλαμβάνουν ειδικές αλληλουχίες, barcodes. Ο barcode αποτελεί προέκταση του αντάπτορα, συνήθως κατά 6 bp, και επιτρέπει την εισαγωγή διαφορετικών δειγμάτων σε μια αντίδραση αλληλούχησης [43, 44].

Στη συγκεκριμένη μεθοδολογία, τα θραύσματα του DNA υβριδοποιούνται με ολιγονουκλεοτίδια, τα οποία βρίσκονται ακινητοποιημένα σε στερεή επιφάνεια (flow-cell) και αποτελούν τους εκκινητές για την αντίδραση PCR, που ακολουθεί. Μετά το στάδιο της υβριδοποίησης, το flow-cell είναι έτοιμο για να ξεκινήσει η ενίσχυση των θραυσμάτων. Η πολυμεράση συνθέτει συμπληρωματικές αλυσίδες DNA χρησιμοποιώντας ως εκμαγείο την αλυσίδα, που έχει υβριδοποιηθεί στον εκκινητή του flow-cell. Δημιουργείται ένα δίκλωνο μόριο DNA, το οποίο αποδιατάσσεται, και κάθε αλυσίδα μπορεί να προσδεθεί στο γειτονικό εκκινητή, δημιουργώντας γέφυρα, όπου στη συνέχεια ενεργοποιείται ξανά η δράση της DNA πολυμεράσης και επαναλαμβάνεται για x κύκλους αντιδράσεων. Με αυτό τον τρόπο, κάποιες αλυσίδες είναι forward ενώ οι υπόλοιπες είναι reverse πολικότητας, ενώ η διαδικασία ενίσχυσης των θραυσμάτων ονομάζεται ενίσχυση μέσω γέφυρας, και έχει ως αποτέλεσμα τη δημιουργία συστάδων (clusters) DNA [27, 45].



Εικόνα 3. Ενίσχυση μέσω Bridge PCR για την παραγωγή clusters προσδεμένων πάνω σε στερεή επιφάνεια (flow-cell). Μονόκλωνες αλυσίδες DNA φέρουν στα άκρα αλληλουχίες συμπληρωματικές με τα ολιγονουκλεοτίδια, που βρίσκονται ακινητοποιημένα πάνω στο flow-cell. Τα ολιγονουκλεοτίδια αποτελούν τους εκκινητές της PCR αντίδρασης καθώς υβριδοποιούνται με το DNA. Προστίθενται τα απαραίτητα αντιδραστήρια για την πραγματοποίηση της αντίδρασης πολυμερισμού και, τελικά, δημιουργούνται clusters από πανομοιότυπα μόρια DNA, τα οποία και θα αλληλουχηθούν, στη συνέχεια μέσω των πλατφορμών αλληλούχησης της Illumina®. (επεξεργασία από [27]).

Σε σύγκριση με τις υπόλοιπες τεχνολογίες αλληλούχησης, η Illumina® διαθέτει τον μεγαλύτερο αριθμό αλληλουχητών και, επιπλέον, είναι η πιο ευρέως διαδεδομένη μεθοδλογία αλληλούχησης που χρησιμοποιείται ως επί το πλείστων σε προγράμματα αλληλούχησης [27]. Η αλληλούχηση με reversible terminators έχει την ικανότητα προσδιορισμού μικρών τμημάτων DNA, μήκους από 50 bp έως και 300 bp (short-reads). Η αποδοτικότητα της εξαρτάται από τον τύπο της πλατφόρμας, που χρησιμοποιείται και κυμαίνεται σε ένα ευρύ φάσμα, μεταξύ χαμηλής και πολύ υψηλής ανάλυσης (small, low-throughput έως large ultra-high-throughput) [28]. Ο πρώτος αλληλουχητής, που κατασκευάστηκε χρησιμοποιώντας τη χημεία αντιδράσεων που περιγράφηκε προηγουμένως, ήταν ο Genome Analyzer® (GA®). Ο GA® αλληλουχητής χαρακτηρίζεται από το πλεονέκτημα της υψηλής ακρίβειας, κυρίως κατά στοίχιση των πειραματικών αλληλουχιών στο γονιδίωμα αναφοράς, αλλά υστερεί, καθώς παράγει πολύ μικρές πειραματικές αλληλουχίες, μέγιστου μήκους έως και 35 bp. Στη συνέχεια, νέες ανανεωμένες εκδόσεις αλληλουχητών έγιναν διαθέσιμες στην αγορά. Κύριοι εκπρόσωποι της πλατφόρμας αλληλούχησης μέσω Illumina® είναι: ο HiSeq® αλληλουχητής, ο οποίος διαθέτει την ικανότητα να παράγει πειραματικές αλληλουχίες μεγαλύτερου μήκους [27, 46, 47] και, πλέον, κατασκευάζονται ακόμα πιο εξελιγμένες πλατφόρμες

αλληλούχησης, όπως ο MiSeq®, ο NextSeq® και NovaSeq® και οι οποίοι προσφέρουν περισσότερες δυνατότητες αλληλούχησης και παράλληλα, αυξάνουν τις εφαρμογές που παρέχει η τεχνολογία αλληλούχησης μέσω της Illumina®.

1.2.3. Μεθοδολογία αλληλούχησης Ion Torrent™

Η μεθοδολογία αλληλούχησης της Ion Torrent™, η οποία αποτελεί προϊόν της εταιρείας ThermoFischer, είναι η πρώτη μέθοδος αλληλούχησης, που χαρακτηρίζεται ως 'post-light sequencing', καθώς δεν εκμεταλλεύεται τις ιδιότητες τους φθορισμού ή της βιοφωταύγειας και δεν χρησιμοποιεί τροποποιημένα νουκλεοτίδια [27, 48, 49]. Η τεχνολογία Ion Torrent™ χρησιμοποιεί τη μέθοδο αλληλούχησης μέσω σύνθεσης και βασίζεται σε ενζυματικές αντιδράσεις για την προσθήκη βάσεων, με τρόπο αντίστοιχο με την μεθοδολογία που χρησιμοποιείται κατά την πυροαλληλούχηση [14]. Ωστόσο, στις πλατφόρμες της Ion Torrent™ δεν ανιχνεύεται οπτικό σήμα, αλλά το σήμα, που παράγεται, οφείλεται στην ανίχνευση πρωτονίων (ιόντων H^+), τα οποία απελευθερώνονται κατά την ενσωμάτωση κάθε νέου dNTP στη νεοσυντιθέμενη αλυσίδα DNA [28, 50, 51]. Το αποτέλεσμα της προσθήκης ενός νέου νουκλεοτιδίου είναι η αλλαγή του pH, η οποία ανιχνεύεται από ένα σύστημα συμπληρωματικού μεταλλικού- οξειδιο- ημιαγωγού (complementary metal - oxide- semiconductor, CMOS) και ενός πεδίου ευαίσθητου σε ιόντα (Ion-Sensitive Field-Effect Transistor, ISFET) [52] (εικόνα 4). Η τεχνολογία των συγκεκριμένων αισθητήρων επιτρέπει την αυξημένη ταχύτητα αλληλούχησης κατά τη διάρκεια ανίχνευσης του σήματος [53].

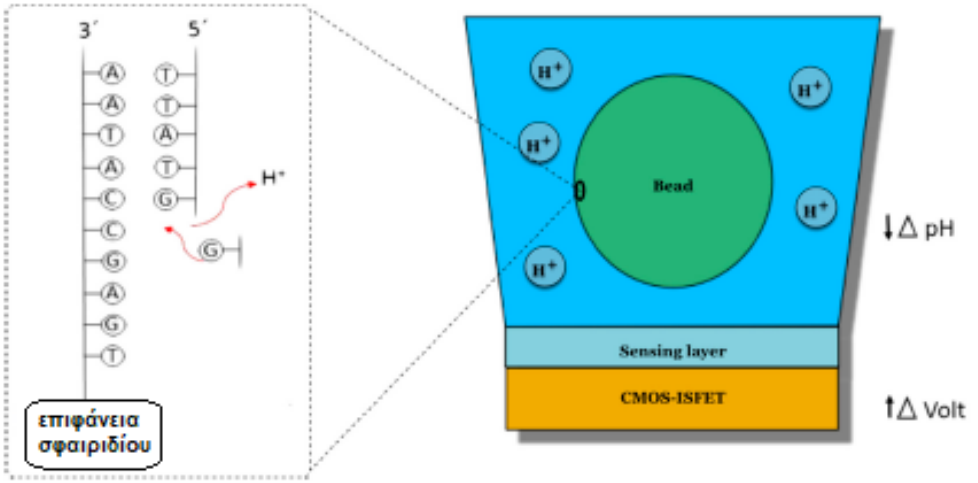
Οι αλλαγές στο pH του διαλύματος μπορούν να ανιχνευθούν ως αλλαγές δυναμικού από τον αισθητήρα. Σε κάθε κύκλο προστίθεται διάλυμα ενός εκ των 4 βάσεων στο διάλυμα της αντίδρασης και ανιχνεύεται η διαφορά δυναμικού με την ενσωμάτωση της αντίστοιχης βάσης στο νέο κλώνο DNA [50]. Εάν δεν ενσωματωθεί κανένα νουκλεοτίδιο δεν παρατηρείται αλλαγή στο δυναμικό. Όταν ενσωματωθούν δύο ίδια νουκλεοτίδια σε ένα κύκλο τότε απελευθερώνονται δύο ιόντα υδρογόνου και η διαφορά δυναμικού που παρατηρείται είναι διπλάσια και, επομένως, το σήμα πιο έντονο [43, 54]. Η αλλαγή του pH που ανιχνεύεται από το μικροσύστημα αισθητήρων βρίσκεται σχεδόν σε αναλογία με τον αριθμό των νουκλεοτιδίων που ενσωματώθηκαν, επιτρέποντας την αλληλούχηση του τμήματος DNA με

συγκεκριμένη ακρίβεια [14, 28]. Ωστόσο, μεγάλα τμήματα ομοπολυμερών του ίδιου νουκλεοτιδίου είναι μερικές φορές δύσκολο να γίνουν διακριτά [43].

Στην μεθοδολογία αλληλούχησης της Ion Torrent™, οι αντιδράσεις συμβαίνουν σε εκατομμύρια από τα πηγάδια, που καλύπτουν το chip του ημιαγωγού, το οποίο περιέχει εκατομμύρια ανιχνευτές CMOS-ISFET, οι οποίοι μετατρέπουν την χημική πληροφορία σε βάσεις νουκλεοτιδίων [43]. Η διαδικασία ξεκινάει με την θραύση του DNA σε μικρότερα τμήματα, στα άκρα των οποίων προσδένονται αντάπτορες, με τη βοήθεια λιγάσης. Τα θραύσματα του DNA συνδέονται με ειδικά σφαιρίδια, τα οποία φέρουν στην επιφάνειά τους αλληλουχίες συμπληρωματικές με τους αντάπτορες. Επόμενο στάδιο είναι η ενίσχυση των τμημάτων, που είναι προσδεδεμένα στα σφαιρίδια, με PCR σε γαλάκτωμα [55]. Η διαδικασία αυτή οδηγεί στην παραγωγή πολλαπλών αντιγράφων του ίδιου κλώνου, που βρίσκεται συνδεδεμένος σε κάθε σφαιρίδιο. Ακολουθεί επιλογή των σφαιριδίων που περιέχουν τους ενισχυμένους κλώνους με τη διαδικασία του εμπλουτισμού (enrichment), που στηρίζεται στη αλληλεπίδραση μεταξύ των μορίων βιοτίνης - στρεπταβιδίνης [56]. Τα σφαιρίδια φορτώνονται σε ειδικά chips, στα οποία θα συμβεί η αντίδραση NGS. Κάθε chip περιέχει εκατομμύρια πηγαδιών (wells), όπου καθένα από αυτά μπορεί να ενσωματώσει ένα μόνο σφαιρίδιο. Όταν τα αντιδραστήρια της αντίδρασης αλληλούχησης διαχέονται κατά μήκος του chip, το συμπληρωματικό, ως προς στην αλυσίδα-εκμαγείο, νουκλεοτίδιο, το οποίο προστίθεται κάθε φορά, ενσωματώνεται στο νέο κλώνο, με αποτέλεσμα την απελευθέρωση ενός ιόντος υδρογόνου, το οποίο ανιχνεύεται από τον CMOS-ISFET αισθητήρα του κάθε well και καταγράφεται το παραγόμενο σήμα [27, 43].

Η τεχνολογία, που διαθέτουν οι αλληλουχητές της Ion Torrent™, είναι γνωστή ως αλληλούχηση με ημιαγωγό, λόγω του συστήματος ανίχνευσης των πρωτονίων (semiconductor sequencing). Οι αλληλουχητές της Ion Torrent™ παράγουν μικρού μήκους πειραματικές αλληλουχίες, έως και 600 bp, με αρκετά μεγάλη ταχύτητα και με μικρό, σχετικά, κόστος εξοπλισμού, ενώ η απόδοση είναι μικρότερη σε σύγκριση με άλλες τεχνολογίες υψηλής απόδοσης [54, 57]. Ένας ευρέως διαδεδομένος αλληλουχητής, που διαθέτει η Ion Torrent™, είναι ο PGM™ (Ion Personal Genome Machine™), όπου αποτελεί μία αξιόπιστη πλατφόρμα αλληλούχησης και συνδυάζει απλή προετοιμασία του προς ανάλυση δείγματος, υψηλής ποιότητας πειραματικές αλληλουχίες και σχετική ευκολία κατά τη βιοπληροφορική ανάλυση των

αποτελεσμάτων. Ο συγκεκριμένος αλληλουχητής χρησιμοποιήθηκε για την διεκπεραίωση της παρούσας διπλωματικής εργασίας. Επιπλέον, υπάρχουν και άλλοι αλληλουχητές της ίδιας εταιρείας, που μοιράζονται την ίδια χημεία αντιδράσεων, αλλά διαθέτουν διαφορετικά χαρακτηριστικά, όπως το σύστημα Ion Proton™, και τα Ion S5™ και ION S5XL™ συστήματα.



Εικόνα 4. Κάθε μέθοδος αλληλούχησης χρησιμοποιεί διαφορετικές χημικές αντιδράσεις για τον προσδιορισμό του DNA. Η μεθοδολογία αλληλούχησης της Ion Torrent™ διαθέτει αισθητήρα ανίχνευσης του σήματος στην επιφάνεια κάθε πηγαδιού του chip, στο οποίο ανιχνεύεται η αλλαγή του pH μέσω της ανίχνευσης των πρωτονίων που απελευθερώνονται κατά τον πολυμερισμό του DNA (επεξεργασία από [27]).

1.2.4. Εφαρμογές της αλληλούχησης επόμενης γενιάς

Σήμερα, οι υψηλής απόδοσης τεχνολογίες αλληλούχησης επόμενης γενιάς (high-throughput NGS, HT-NGS) αποτελούν ένα από τα σημαντικότερα εργαλεία στο πεδίο της έρευνας του ανθρώπινου γονιδιώματος [24]. Οι τεχνολογίες NGS εφαρμόζονται σε ολοένα και περισσότερα ερευνητικά πεδία, τα οποία αφορούν στην *de novo* αλληλούχηση γονιδιωμάτων βακτηρίων και ιών [58, 59], την αναζήτηση γενετικών αλλαγών, με την αλληλούχηση ολόκληρου του γονιδιώματος (WGS), ή στοχευμένων περιοχών του γονιδιώματος (Targeted-seq, ampli-seq), την μελέτη των περιοχών του γονιδιώματος που κωδικοποιούν για πρωτεΐνες, δηλαδή των εξωνίων (WES), την μελέτη μη κωδικών μορίων RNA (miRNA-seq, lncRNA-seq), την κατανόηση των γενετικών μηχανισμών που καθορίζουν τις μεταβολές στην έκφραση γονιδίων, την μελέτη του μεταγραφώματος κυττάρων, ιστών αλλά και όλου

του οργανισμού, μέσω RNA-Seq [60, 61], και την διερεύνηση του τρόπου αλληλεπίδρασης DNA-πρωτεϊνών, καθώς και, των επιγενετικών αλλαγών, μέσω ChIP-Seq [62].

Πιο αναλυτικά, η αλληλούχηση ολόκληρου του γονιδιώματος (whole genome sequencing, WGS) επιτρέπει τον προσδιορισμό ολόκληρης της DNA αλληλουχίας ενός οργανισμού. Η μέθοδος WGS έχει εισαχθεί στην κλινική πράξη και χάρη στις πληροφορίες, που προσφέρει, μπορεί να χαρακτηριστεί ως ένα πολύτιμο εργαλείο, στο οποίο μπορεί να στηριχθεί η εξατομικευμένη ιατρική προκειμένου να επιτευχθεί η κατάλληλη θεραπευτική προσέγγιση [41, 63, 64]. Η μελέτη των εξωνίων (whole exome sequencing, WES) στηρίζεται στην αλληλούχηση όλων των κωδικών περιοχών, που αποτελούν περίπου το 1-2% του ανθρώπινου γονιδιώματος. Το 85% του συνόλου των μεταλλαγών του DNA συμβαίνουν σε αυτές τις μικρές περιοχές του γονιδιώματος, με αποτέλεσμα να ευθύνονται κατά κύριο λόγο για την εκδήλωση ανθρώπινων ασθενειών [65]. Με την εφαρμογή των τεχνολογιών NGS στόχος είναι ο προσδιορισμός και η ταυτοποίηση των γενετικών αλλαγών, οι οποίες ευθύνονται για αλλαγές, που παρατηρούνται στις πρωτεΐνες, και που, κατά επέκταση, οδηγούν σε ασθένειες [66].

Η στοχευμένη αλληλούχηση (Targeted-seq), με τη χρήση NGS, έχει κλινική εφαρμογή κυρίως στο πεδίο της ογκολογίας καθώς επιτρέπει τον προσδιορισμό της αλληλουχίας συγκεκριμένων περιοχών του γονιδιώματος, λειτουργώντας ως εργαλείο για την ανάλυση γνωστών ή νέων μεταλλαγών σε ένα δείγμα. Επιπλέον, μπορεί να χρησιμοποιηθεί για την μελέτη κληρονομικών ασθενειών, τα οποία χαρακτηρίζονται ως μενδελικά κληρονομούμενα νοσήματα [67]. Η χρήση της συγκεκριμένης μεθόδου αλληλούχησης προσφέρει την δυνατότητα ανίχνευσης σπάνιων μεταλλαγών, την δυνατότητα ανίχνευσης προσθήκης ή απαλοιφής βάσεων, καθώς και, έχει την ικανότητα να εντοπίζει μοναδιαίες αλλαγές σε μια μόνο βάση (single nucleotide variant, SNV) [68]. Επιπλέον, χρησιμοποιείται ο όρος Amplicon-sequencing (Ampli-seq) για την αλληλούχηση συγκεκριμένων τμημάτων DNA, τα οποία, προηγουμένως, έχουν ενισχυθεί με PCR. Η προσέγγιση αυτή χρησιμοποιείται για την γονοτύπηση μια συγκεκριμένης περιοχής ενδιαφέροντος σε ένα μεγάλο αριθμό δειγμάτων [69].

Ένα σημαντικό πλεονέκτημα, που προσφέρει η αλληλούχηση επόμενης γενιάς, είναι η δυνατότητα χαρακτηρισμού του μεταγραφώματος ενός ιστού, ενός

αναπτυξιακού σταδίου ή ενός οργανισμού. Ο όρος μεταγράφωμα αναφέρεται στο σύνολο των μορίων RNA, που παράγονται από ένα πληθυσμό κυττάρων. Η μέθοδος είναι γνωστή ως RNA-seq ή Whole Transcriptome Sequencing (WTS) και περιλαμβάνει την αλληλούχηση του cDNA, με σκοπό την απόκτηση πληροφοριών, που σχετίζονται με το περιεχόμενο του RNA στο δείγμα, και αφορά, τόσο νουκλεοτιδικές αλληλουχίες όσο και τα επίπεδα έκφρασης [30, 69]. Η αλληλούχηση επόμενης γενιάς δίνει επιπλέον, τη δυνατότητα ανακάλυψης και χαρακτηρισμού νέων μη κωδικών μορίων RNA, τα οποία έχουν σημαντικό, και κυρίως, ρυθμιστικό ρόλο σε πολλές βιολογικές διαδικασίες [70]. Τα μη κωδικά μόρια RNA μπορούν να ταυτοποιηθούν ανάλογα με το μήκος τους, είτε με προσεγγίσεις lncRNA-seq είτε μέσω miRNA-seq.

Τέλος, χρησιμοποιείται η μέθοδος ChIP-seq για την διερεύνηση των αλληλεπιδράσεων μεταξύ του DNA και των πρωτεϊνών, η οποία συνδυάζει την μαζική παράλληλη αλληλούχηση DNA με δοκιμασίες ανοσοκαθίζησης της χρωματίνης. Η διαλεύκανση των αλληλεπιδράσεων μεταξύ των δύο αυτών μορίων μπορεί να προσφέρει χρήσιμες πληροφορίες για την γονιδιακή ρύθμιση και την κατανόηση πολλών βιολογικών διαδικασιών, καθώς και, τον προσδιορισμό των σταδίων μια ασθένειας [71]. Επομένως, οι μέθοδοι RNA-seq και ChIP-seq είναι δυνατόν να χρησιμοποιηθούν συμπληρωματικά για την πλήρη κατανόηση των λειτουργιών ενός κυττάρου ή ιστού, και κατ' επέκταση, και του οργανισμού [72].

1.3. Αλληλούχηση Τρίτης Γενιάς (Third-generation sequencing, TGS)

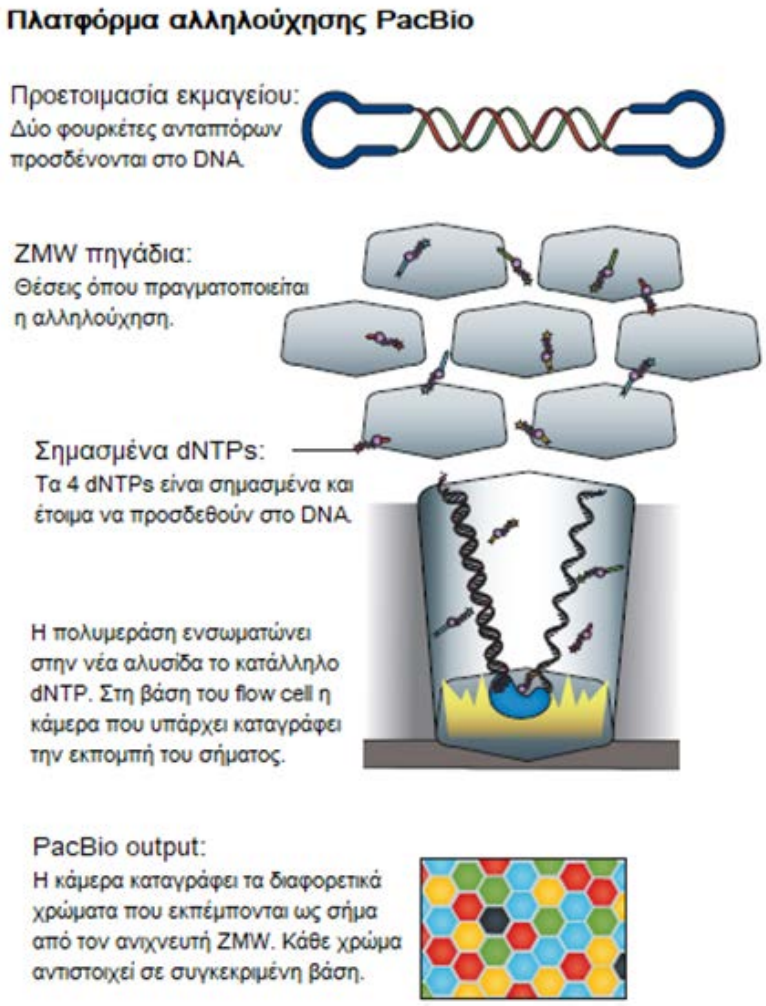
Την προηγούμενη δεκαετία, η αλληλούχηση επόμενης γενιάς υπήρξε, χωρίς αμφιβολία, ένα πανίσχυρο ερευνητικό εργαλείο, το οποίο επέτρεψε την μαζική αλληλούχηση μεγάλων περιοχών του γονιδιώματος, οδηγώντας στην ταυτοποίηση νέων μορίων, που εκφράζονται τόσο σε φυσιολογικές όσο και σε παθολογικές καταστάσεις [73]. Ωστόσο, η ανάπτυξη των νέων μεθοδολογιών αλληλούχησης τρίτης γενιάς (Third-generation sequencing, TGS) ενισχύει σημαντικά τις μελέτες για την κατανόηση της πολυπλοκότητας του ανθρώπινου γονιδιώματος [74]. Η τεχνολογία τρίτης γενιάς στηρίζεται στην αλληλούχηση ενός μόνο μορίου σε πραγματικό χρόνο (single-molecule real-time sequencing, SMRT), κατά την οποία, είναι δυνατόν να προσδιοριστούν πολύ μεγάλες αλληλουχίες DNA χωρίς να απαιτούν τη θραύση του DNA σε μικρότερα τμήματα και την ενίσχυση του μέσω

PCR [20, 75, 76]. Η τεχνολογία αλληλούχησης τρίτης γενιάς αποτελεί ένα χρήσιμο εργαλείο για την ανίχνευση επιγενετικών τροποποιήσεων και την μελέτη CpG μεθυλιώσεων σε πραγματικό χρόνο. Επιπλέον, επιτρέπει τον άμεσο προσδιορισμό του πλήρους μήκους μιας αλληλουχίας, όπως τα μετάγραφα mRNA, χωρίς να απαιτείται θραύση του DNA σε μικρότερα τμήματα κατά την προετοιμασία του δείγματος [76]. Οι δύο διαθέσιμες μεθοδολογίες αλληλούχησης τρίτης γενιάς, που χρησιμοποιούνται, είναι η Pacific Bioscience® (PacBio®) και η Nanopore Oxford Technologies®, οι οποίες ακολουθούν διαφορετικές προσεγγίσεις.

1.3.1. Μεθοδολογία αλληλούχησης PacBio®

Η μέθοδος αλληλούχησης της PacBio® αναφέρεται και ως τεχνολογία αλληλούχησης ενός μορίου σε πραγματικό χρόνο (single-molecule real-time, SMRT), χρησιμοποιεί τις ιδιότητες της σύνθεσης του DNA και επιτρέπει τον προσδιορισμό μορίων, έως και 50 kb ή ακόμα μεγαλύτερου μήκους [43]. Οι προηγούμενες μεθοδολογίες αλληλούχησης, μέσω σύνθεσης, χαρακτηρίζονται από την ιδιότητα της πολυμεράσης να προσδένεται στο μόριο DNA, που πρόκειται, να αλληλουχηθεί και να κινείται κατά μήκος αυτού συνθέτοντας την νέα συμπληρωματική αλυσίδα. Αντιθέτως, η μεθοδολογία, που χρησιμοποιείται στην αλληλούχηση SMRT, βασίζεται στην ακινητοποίηση της DNA πολυμεράσης μέσα στο ειδικά διαμορφωμένο πηγάδι του flow cell όπου θα συμβεί η αντίδραση, ενώ το DNA είναι το κινητό μόριο [27].

Στη συγκεκριμένη πλατφόρμα καθένα από τα χιλιάδες πηγάδια του flow cell διαθέτουν μια ειδική νανοτεχνολογία για την ανίχνευση του σήματος, που ονομάζεται ‘ZMW’ (zero-mode waveguide). Ο αισθητήρας ‘ZMW’ ανιχνεύει το σήμα, το οποίο δημιουργείται από την ενσωμάτωση ενός σημασμένου, με φώσφορο, νουκλεοτιδίου, καθώς η DNA πολυμεράση αντιγράφει το DNA [27]. Η ενσωμάτωση κάθε dNTP, το οποίο είναι σημασμένο με διαφορετικό φθορίζον μόριο, είναι δυνατόν να οπτικοποιηθεί με τη χρήση ενός συστήματος laser και κάμερας, όπου καταγράφουν τα εκπεμπόμενα σήματα [77].



Εικόνα 5. Διάγραμμα ροής της διαδικασίας αλληλούχησης μέσω της πλατφόρμας αλληλούχησης PacBio®. Η βασική αρχή της τεχνολογίας βασίζεται στην ενσωμάτωση συμπληρωματικών νουκλεοτιδίων σε ένα νεοσυντιθέμενο κλώνο και την ανίχνευση του σήματος, που παράγεται κατά την προσθήκη αυτή, όπως συμβαίνει και στην μεθοδολογία του NGS (επεξεργασία από [28]).

Η πλατφόρμα αλληλούχησης της PacBio® προσφέρει αρκετά πλεονεκτήματα, τα οποία δεν είναι εφικτά με τη χρήση των προηγούμενων μεθοδολογιών αλληλούχησης. Αρχικά, δίνει τη δυνατότητα αλληλούχησης ενός μόνο μορίου σε πολύ σύντομο χρονικό διάστημα και έχει, επίσης, την ικανότητα να προσδιορίζει εξαιρετικά μεγάλου μήκους αλληλουχίες, μεγαλύτερες των 10 kb, διευκολύνοντας τη *de novo* συναρμολόγηση ολόκληρων γονιδιωμάτων [77]. Επιπλέον, κατά τη διάρκεια της αλληλούχησης, ο ρυθμός με τον οποίο η DNA πολυμεράση συνθέτει τον νέο κλώνο επιτρέπει την ανίχνευση τροποποιημένων βάσεων καθώς παράγονται, παρέχοντας, επιπλέον, και κινητικά δεδομένα [78].

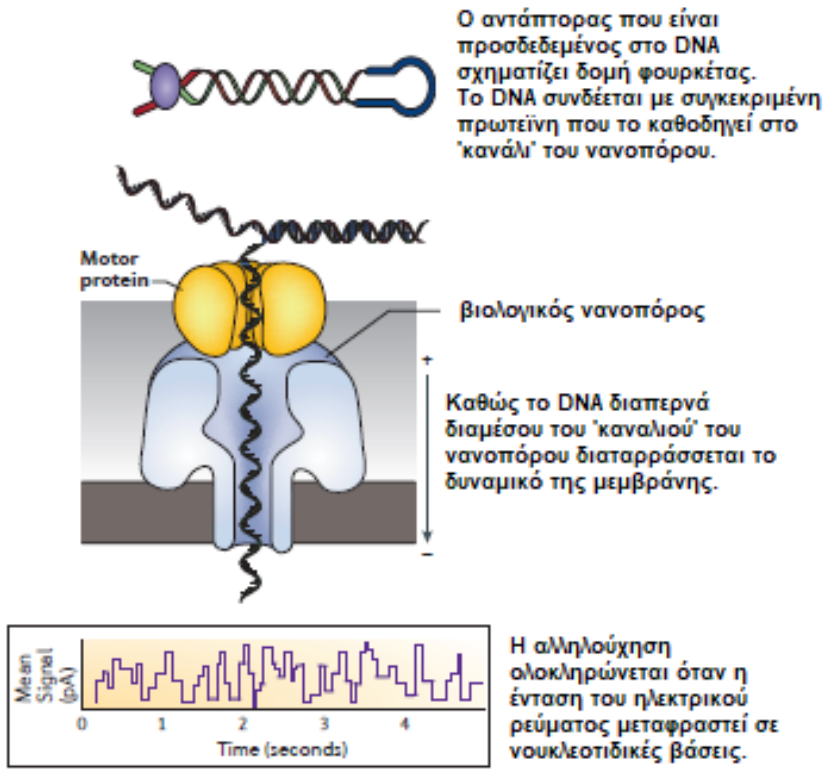
1.3.2. Μεθοδολογία αλληλούχησης Oxford Nanopore Technologies®

Οι πλατφόρμες αλληλούχησης της εταιρείας Oxford Nanopore Technologies® λειτουργούν με στόχο την αλληλούχηση ενός μορίου DNA ή RNA σε πραγματικό χρόνο. Σε αντίθεση με άλλες πλατφόρμες αλληλούχησης, οι οποίες ανιχνεύουν δευτερογενή σήματα, όπως το φως, οι χρωστικές ή το pH, η τεχνολογία της ONT ανιχνεύει απευθείας τις νουκλεοτιδικές βάσεις του DNA από ένα και μόνο μονόκλωνο μόριο [28]. Η γενική ιδέα της αλληλούχησης του DNA με την χρήση ειδικών νανοπόρων είναι αρκετά απλή και προτάθηκε, για πρώτη φορά, στο τέλος της δεκαετίας του 1990 [15]. Ωστόσο, μόλις το 2014 έγινε διαθέσιμος στην αγορά ο πρώτος αλληλουχητής, MinION™, της εταιρείας Oxford Nanopore® Technologies (ONT) [79].

Η τεχνολογία, που έχει αναπτυχθεί για την αλληλούχηση, χρησιμοποιεί πρωτεϊνικής φύσης νανοπόρους σε καθένα από τους οποίους εφαρμόζεται ηλεκτρικό πεδίο, ώστε να διευκολύνεται η διέλευση του DNA. Το σύστημα των νανοπόρων αποτελείται από νανο-αισθητήρες, οι οποίοι διαμορφώνουν ειδική δομή ‘καναλιών’, από τα οποία διαπερνά το DNA [27]. Αρχικά, το δίκλωνο μόριο DNA αποδιατάσσεται και, επομένως, στο κανάλι εισέρχεται μία μονόκλωνη αλυσίδα DNA. Μια βοηθητική πρωτεΐνη κατευθύνει την μονόκλωνη αλυσίδα DNA διευκολύνοντας την διέλευσή του διαμέσου του πόρου, που έχει σχηματιστεί. Το γεγονός αυτό οδηγεί, τελικά, στη διατάραξη του δυναμικού του καναλιού, που έχει δημιουργηθεί, η οποία ανιχνεύεται από τον αισθητήρα. Η αλλαγή του δυναμικού είναι χαρακτηριστική για κάθε αλληλουχία DNA, καθώς, κάθε μία από τις τέσσερις βάσεις διαταράσσει το κανάλι σε διαφορετικό βαθμό [14]. Ωστόσο, ενώ το προφανές είναι να υπάρχουν 1 έως 4 πιθανά σήματα (ένα χαρακτηριστικό για κάθε βάση), ο αλληλουχητής διαθέτει περισσότερα από 1000 διαφορετικά σήματα, που μπορεί να ανιχνευθούν, ένα για κάθε πιθανό μικροπολυμερές (k-mer). Επομένως, το σύστημα της Oxford Nanopore® Technologies δεν ανιχνεύει μεμονωμένες βάσεις αλλά, μικρές αλληλουχίες βάσεων (k-mers) [28]. Όπως και οι αλληλουχητές της PacBio®, έτσι και οι αλληλουχητές της ONT δίνουν την δυνατότητα αλληλούχησης πολύ μεγάλων τμημάτων DNA (> 10 kb) [80].

Το κλασικό flow cell, που χρησιμοποιούν οι περισσότεροι αλληλουχητές της ONT, όπως η συσκευή MinION™, που χρησιμοποιήθηκε στην παρούσα εργασία, αποτελείται από ένα chip, το οποίο διαθέτει 512 διαφορετικά κανάλια (νανοπόρους).

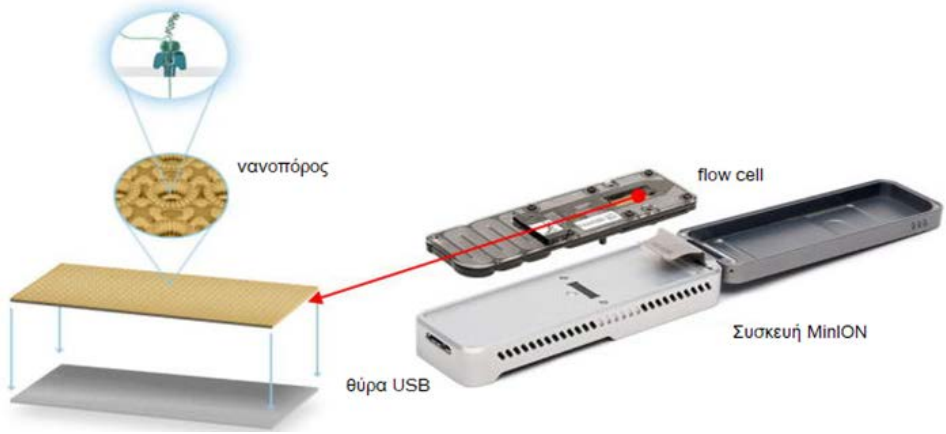
Η αλληλούχηση τρίτης γενιάς μέσω των πλατφόρμων της ONT χαρακτηρίζεται από μεγάλη ταχύτητα, καθώς τα κανάλια, που βρίσκονται στο flow cell, έχουν την δυνατότητα να προσδιορίζουν το DNA με ταχύτητα μεγαλύτερη από 500 bp ανά δευτερόλεπτο. Η πλατφόρμες της ONT μπορεί να λειτουργούν είτε με βιολογικούς νανοπόρους είτε με συνθετικά κανάλια. Το πλεονέκτημα των βιολογικών νανοπόρων είναι η εύκολη τροποποίηση, που μπορούν να υποστούν, και, επομένως, να παραχθούν πανομοιότυπα κανάλια σε μέγεθος και δομή [14].



Εικόνα 6. Η αλληλούχηση με τη χρήση νανοπόρων επιτρέπει την ανίχνευση της αλλαγής δυναμικού, που προκαλείται κατά την διέλευση του DNA στη μεμβράνη του νανοπόρου. Στη συνέχεια, γίνεται αντιστοίχιση του ηλεκτρικού σήματος με τη νουκλεοτιδική αλληλουχία του DNA.

Παρόλο που η αλληλούχηση τρίτης γενιάς μέσω της μεθοδολογίας Oxford Nanopore® Technologies δίνει την δυνατότητα παραγωγής μεγάλων πειραματικών αλληλουχιών μήκους, έως και 2 Mb, σε πολύ μικρό χρονικό διάστημα, η αποδοτικότητά της δεν είναι τόσο επιτυχής. Ειδικότερα, σημαντικό μειονέκτημα της μεθόδου είναι ο υψηλός βαθμός σφαλμάτων που συμβαίνουν κατά την

αλληλούχηση και οφείλονται κυρίως στον τύπο του flow cell και στην αναγνώριση του ηλεκτρικού σήματος. Υπάρχουν διαφορετικοί τύποι flow cells που καθορίζουν, έως ένα βαθμό, την απόδοση της αλληλούχησης. Τα R9.4 flow cells, καθώς και οι παλιότερες εκδόσεις, χρησιμοποιούν 1D χημεία [81], ενώ τα νεότερα 9.5 χρησιμοποιούν 2D. Η πιο πρόσφατη έκδοση των flow cells R10 θεωρείται ότι αυξάνει την ακρίβεια στο 99,99% [82]. Ο όρος 1D υποδηλώνει πως το εκμαγείο και η συμπληρωματικός κλώνος θα αλληλουχηθούν ως δύο ξεχωριστές αλυσίδες. Η χημεία που χρησιμοποιείται για να επιτευχθεί 2D αλληλούχηση, καθορίζεται στο στάδιο της κατασκευής της βιβλιοθήκης που θα αλληλουχηθεί. Στην 2D αλληλούχηση ο συμπληρωματικός κλώνος θα αλληλουχηθεί αμέσως μόλις ολοκληρωθεί η αλληλούχηση της πρώτης αλυσίδας με την βοήθεια ειδικών ανταπτόρων. Η 2D μέθοδος αλληλούχησης παράγει πειραματικές αλληλουχίες μεγαλύτερης ακρίβειας με μειωμένο αριθμό σφαλμάτων.



Εικόνα 7. Η φορητή συσκευή MinION™ της ONT δίνει τη δυνατότητα αλληλούχησης τμημάτων DNA εκτός εργαστηρίου, ανά πάσα στιγμή και με σχετική ευκολία. Το flow cell, στο οποίο εισάγεται το δείγμα, περιλαμβάνει σημαντικό αριθμό νανοπόρων διαμέσου των οποίων θα αλληλουχηθεί το DNA.

Το μικρό μέγεθος της MinION™ συσκευής, η ευκολία στο χειρισμό, η μικρή προετοιμασία που απαιτείται για την κατασκευή βιβλιοθήκης, που θα αλληλουχηθεί, καθώς και, το κόστος δίνουν τη δυνατότητα αλληλούχησης στο πεδίο, εκτός του συνηθισμένου, έως σήμερα, χώρου του εργαστηρίου [83, 84]. Σήμερα, είναι διαθέσιμοι στην αγορά δύο ακόμα αλληλουχητές από την ONT ο GridION™ και ο PromethION™, οι οποίοι δίνουν την δυνατότητα παράλληλης αλληλούχησης ακόμα περισσότερων μορίων DNA σε μειωμένο χρόνο, καθώς, ο GridION™ διαθέτει 5 MinION™ flow cells, ενώ ο αλληλουχητής PromethION™ περιέχει 24 (P24) ή 48 (P48)

διαφορετικά flow cells και δίνει την δυνατότητα σε έως και 144.000 (P48) ενεργά κανάλια να λειτουργούν ταυτόχρονα, προσδιορίζοντας νουκλεοτιδικές αλληλουχίες [84].

1.4. Σύγκριση των μεθοδολογιών αλληλούχησης NGS-TGS

Η Μαζική Παράλληλη Αλληλούχηση, που προσφέρει η τεχνολογία NGS, καθώς και η ανάπτυξη της τεχνολογίας αλληλούχησης ενός μόνο μορίου σε πραγματικό χρόνο, αποτελούν εξαιρετικά εργαλεία για την μελέτη του γονιδιώματος και έχουν, ήδη, συμβάλλει στην κατανόηση πολλών βιολογικών διαδικασιών που συμβαίνουν στα κύτταρα. Οι καινοτομίες των μεθοδολογιών αλληλούχησης επιτρέπουν τον προσδιορισμό του DNA με μεγάλη ακρίβεια, σε εύλογο χρονικό διάστημα και με, σχετικά, μειωμένο κόστος. Το πλήθος των πλατφόρμων αλληλούχησης που είναι διαθέσιμοι δίνει την δυνατότητα εφαρμογής των τεχνολογιών αυτών σε πολλά πεδία μελέτης, τα οποία αφορούν τους ρυθμιστικούς και λειτουργικούς μηχανισμούς που διέπουν τα κύτταρα. Ωστόσο, η κάθε μεθοδολογία έχει συγκεκριμένα πλεονεκτήματα καθώς και μειονεκτήματα με αποτέλεσμα οι πλατφόρμες αλληλούχησης να διαφοροποιούνται ανάλογα με τα χαρακτηριστικά τους, τις τεχνικές που χρησιμοποιούν και τις δυνατότητες που προσφέρουν.

Οι πλατφόρμες αλληλούχησης NGS απαιτούν μεγαλύτερη προεργασία, όσο αφορά στην προετοιμασία μιας NGS βιβλιοθήκης, σε σύγκριση με την προετοιμασία του δείγματος που πρόκειται να αλληλουχηθεί σε οποιαδήποτε TGS πλατφόρμα. Η αλληλούχηση επόμενης γενιάς απαιτεί την θραύση του DNA σε τμήματα, την πρόσδεση ανταπτόρων στα άκρα των τμημάτων και, στη συνέχεια, την ενίσχυση αυτών μέσω PCR, με σκοπό να αυξηθεί ο αριθμός των κλώνων που θα αλληλουχηθούν. Η αντίδραση PCR που πραγματοποιείται, αφενός έχει ως αποτέλεσμα την στοχευμένη αλληλούχηση του επιθυμητού DNA-στόχου, αφετέρου μπορεί να οδηγήσει στην εισαγωγή σφαλμάτων που προκύπτουν κατά την διαδικασία του πολλαπλασιασμού των κλώνων. Επιπλέον, πιθανά τμήματα του DNA, τα οποία δεν θα πολλαπλασιαστούν κατά την PCR αντίδραση, πχ λόγω αδυναμίας υβριδοποίησης των εκκινητών, θα εξαιρεθούν από την αντίδραση αλληλούχησης, με αποτέλεσμα τελικά την πιθανότητα απώλειας σημαντικών πληροφοριών.

Αντίθετα, οι μεθοδολογίες τρίτης γενιάς χρησιμοποιούν μεμονωμένα τμήματα DNA, τα οποία δεν υφίστανται περαιτέρω ενίσχυση μέσω PCR και, επομένως, δεν υπάρχει ο κίνδυνος απώλειας της πληροφορίας. Ακολούθως, η ποσότητα του DNA, που απαιτείται σε κάθε μέθοδο αλληλούχησης, διαφέρει, με τους NGS αλληλουχητές να απαιτούν σημαντικά μεγαλύτερη ποσότητα συγκριτικά με τους αλληλουχητές τρίτης γενιάς. Τέλος, οι χημικές αντιδράσεις, στις οποίες στηρίζεται κάθε μεθοδολογία, καθορίζουν τόσο τον τύπο του παραγόμενου σήματος το οποίο θα ανιχνευθεί, όσο και μια σειρά από παραμέτρους, που θα αναφερθούν στην συνέχεια.

Πίνακας 1. Συνοπτική παρουσίαση της αρχής μεθόδου κάθε μεθόδου αλληλούχησης. Στον πίνακα παρουσιάζονται οι βασικές αντιδράσεις που απαιτούνται σε κάθε μέθοδο αλληλούχησης για τον προσδιορισμό της ακολουθίας ενός τμήματος DNA.

Μεθοδολογία	**Δείγμα**	**Τεχνολογία αντίδρασης**	**Σήμα**
Αλληλούχηση Sanger	Μείγμα	PCR, Αντίδραση σύνθεσης με τερματισμό αλυσίδας	Φθορισμός
Πυροαλληλούχηση®	Ενισχυμένος κλώνος	PCR σε γαλάκτωμα, Αντίδραση σύνθεσης με διαδοχική προσθήκη βάσεων	Χημειοφωταύγεια
Illumina®	Ενισχυμένος κλώνος	PCR μέσω γέφυρας, Αντίδραση σύνθεσης με ανάστροφους τερματιστές	Φθορισμός
Ion Torrent™	Ενισχυμένος κλώνος	PCR σε γαλάκτωμα, Αντίδραση σύνθεσης με διαδοχική προσθήκη βάσεων	Αλλαγές στο pH
PacBio®	Μεμονωμένο μόριο DNA	Αντίδραση σύνθεσης χωρίς την απαίτηση ενίσχυσης του δείγματος σε πραγματικό χρόνο	Φθορισμός
Oxford Nanopore®	Μεμονωμένο μόριο DNA/RNA	Απευθείας ανίχνευση της νουκλεοτιδικής αλληλουχίας χωρίς την απαίτηση ενίσχυσης του δείγματος σε πραγματικό χρόνο	Αλλαγή δυναμικού

Ειδικότερα, οι πλατφόρμες αλληλούχησης της ONT προσφέρουν ένα νέο τύπο αλληλουχητών, που χρησιμοποιεί πρωτεΐνες νανοπόρους, και επιτρέπει την αλληλούχηση ενός μορίου DNA χωρίς να απαιτείται η θραύση του σε τμήματα και η

ενίσχυσή του μέσω PCR. Επιπλέον, η αλληλούχηση με την χρήση νανοπόρων δεν ακολουθεί την συνήθη μεθοδολογία αλληλούχησης που βασίζεται στην σύνθεση ενός νέου κλώνου DNA συμπληρωματικού ως προς το προς προσδιορισμό μόριο, αλλά, ανιχνεύει την αλληλουχία των νουκλεοτιδικών βάσεων απευθείας, μέσω διατάραξης του δυναμικού της μεμβράνης που σχηματίζεται σε κάθε κανάλι-νανοπόρο. Στον πίνακα 1 παρουσιάζονται συνοπτικά βασικά χαρακτηριστικά, που αφορούν στην μεθοδολογία, που ακολουθούν οι κυριότερες πλατφόρμες αλληλούχησης.

Ένα επόμενο σετ διαφορών μεταξύ των δύο κατηγοριών αλληλούχησης, που χρησιμοποιούνται, αφορά στα μετρικά χαρακτηριστικά της κάθε πλατφόρμας αλληλούχησης. Αρχικά, κάθε πλατφόρμα δίνει τη δυνατότητα προσδιορισμού συγκεκριμένου μήκους DNA αλληλουχίας. Ο περιορισμός αυτός έγκειται στη διαφορετική χημεία των αντιδράσεων, που χρησιμοποιείται από κάθε πλατφόρμα, καθώς και, στα χαρακτηριστικά του κάθε αλληλουχητή. Συγκεκριμένα, η αρχή μεθόδου στις πλατφόρμες αλληλούχησης επόμενης γενιάς βασίζεται στην αλληλούχηση μέσω σύνθεσης, όπου απαιτούνται κατάλληλα αντιδραστήρα, καθώς και η δράση DNA πολυμεράσης. Επομένως, το μήκος του παραγόμενου προϊόντος, που αλληλουχείται σε μία αντίδραση σύνθεσης, εξαρτάται σε μεγάλο βαθμό από την ικανότητα πολυμερισμού του ενζύμου και, για τον λόγο αυτό, οι διαθέσιμοι αλληλουχητές είναι δυνατόν να προσδιορίζουν πειραματικές αλληλουχίες συγκεκριμένου μήκους. Αντίθετα, η αλληλούχηση με την χρήση των αλληλουχητών της εταιρείας ONT επιτρέπει τον προσδιορισμό πολύ μεγαλύτερων αλληλουχιών, που ξεπερνούν τις 10 kb σε μήκος, αφού ανιχνεύουν απευθείας τη διαφορά δυναμικού που παράγεται, καθώς η μονόκλωνη αλυσίδα DNA διέρχεται από το κανάλι της πρωτεΐνης του νανοπόρου, παραλείποντας τη διαδικασία σύνθεσης DNA αλυσίδας.

Επιπλέον, η χημεία της αντίδρασης, που χρησιμοποιεί κάθε αλληλουχητής, καθώς και τα ιδιαίτερα χαρακτηριστικά του, έχουν ως αποτέλεσμα την διαφοροποίηση του συνολικού χρόνου που απαιτείται για την εκτέλεση ενός πειράματος αλληλούχησης. Σημαντικό ρόλο στον συνολικό χρόνο, που απαιτείται για την αλληλούχηση μιας βιβλιοθήκης, έχουν, επίσης, η ποιότητα και η ποσότητα του δείγματος που εισάγεται στον αλληλουχητή. Ο πίνακας 2 παρουσιάζει τον εκτιμώμενο μέσο χρόνο πειραματικής αλληλούχησης, που αντιστοιχεί σε κάθε

αλληλουχητή νέας γενιάς, ο οποίος είναι συνάρτηση τόσο του τύπου του αλληλουχητή όσο και του προϊόντος, το οποίο προσδιορίζεται σε βέλτιστες συνθήκες πειραμάτων. Οι νέοι αλληλουχητές τρίτης γενιάς χαρακτηρίζονται από αυξημένη ταχύτητα, ενώ ταυτόχρονα, δίνουν την δυνατότητα αλληλούχησης πολύ μεγάλων τμημάτων πειραματικών αλληλουχιών. Παράλληλα, τόσο μεταξύ των μεθοδολογιών, όσο και μεταξύ των διαφορετικών αλληλουχητών, που είναι διαθέσιμοι, διαφέρει ο μέγιστος αριθμός πειραματικών αλληλουχιών, που μπορούν να προσδιοριστούν σε ένα πείραμα αλληλούχησης και, ως εκ τούτου, κάθε αλληλουχητής επιτρέπει την παραγωγή διαφορετικού μεγέθους αρχείων εξόδου με τα δεδομένα κάθε πειράματος αλληλούχησης.

Σύμφωνα με τα στοιχεία που παρουσιάζονται στον πίνακα 2, οι διαφορετικές προδιαγραφές κάθε αλληλουχητή καθορίζουν παραμέτρους, όπως το μέγιστο μήκος της πειραματικής αλληλουχίας, τον απαιτούμενο χρόνο, το μέγεθος των αρχείων εξόδου, καθώς και το πλήθος των αλληλουχιών, που θα προσδιοριστούν σε ένα πείραμα. Η τεχνολογία της Illumina® διαθέτει 6 βασικούς αλληλουχητές, οι διαφορές των οποίων συμβάλλουν στη διαμόρφωση των συνθηκών αλληλούχησης και καθορίζουν το παραγόμενο αποτέλεσμα. Ειδικότερα, ο NovaSeq® 6000, ο τελευταίος αλληλουχητής, που είναι διαθέσιμος από την εταιρεία, παράγει πειραματικές αλληλουχίες μήκους έως και 500 βάσεων και η αντίδραση αλληλούχησης διαρκεί λίγο λιγότερο από δύο μέρες. Επιπλέον, επιτρέπει την αλληλούχηση έως και 20 δισεκατομμυρίων πειραματικών αλληλουχιών και μπορεί να παράγει αρχεία εξόδου έως και 6000 Gb. Αντίθετα, οι τρεις κύριοι αλληλουχητές της Ion Torrent™ έχουν την δυνατότητα να παράγουν πειραματικές αλληλουχίες έως και 600 bp και ο χρόνος του κάθε πειράματος αλληλούχησης δεν ξεπερνά τις 24 ώρες. Ωστόσο, το μέγιστο πλήθος των πειραματικών αλληλουχιών, που μπορούν να προσδιοριστούν, είναι 130 εκατομμύρια σε ένα πείραμα αλληλούχησης που χρησιμοποιεί τον αλληλουχητή Ion S5®, με μέγιστο μέγεθος αρχείων εξόδου τα 50 Gb.

Οι αλληλουχητές τρίτης γενιάς της εταιρείας ONT, δίνουν την δυνατότητα αλληλούχησης εξαιρετικά μεγάλων τμημάτων DNA, ακόμα και ολόκληρων γονιδιωμάτων, σε ένα μόνο πείραμα αλληλούχησης. Ο μέγιστος χρόνος αλληλούχησης είναι 3 ημέρες και καθορίζεται από την αντοχή που έχουν οι πρωτεΐνες/ νανοπόροι να διατηρούν τη δομή καναλιού και να ανιχνεύουν την

μεταβολή του ρεύματος κατά την διέλευση του DNA. Τα αρχεία εξόδου περιέχουν εκατομμύρια πειραματικές αλληλουχίες, ο μέγιστος αριθμός των οποίων μεταβάλλεται από τις πειραματικές συνθήκες. Επομένως, το μεγάλο πλεονέκτημα των αλληλουχητών τρίτης γενιάς είναι η ικανότητά τους να προσδιορίζουν πάρα πολύ μεγάλες σε μήκος νουκλεοτιδικές αλληλουχίες χωρίς να απαιτείται η θραύση του DNA σε μικρότερα τμήματα.

Πίνακας 2. Σύγκριση των διαφορών που παρουσιάζονται μεταξύ των αλληλουχητών στις ευρέως χρησιμοποιούμενες πλατφόρμες αλληλούχησης δεύτερης και τρίτης γενιάς. Οι πληροφορίες, που εμφανίζονται, αναφέρονται στις βέλτιστες συνθήκες ενός πειράματος αλληλούχησης, για κάθε αλληλουχητή. Τα δεδομένα αυτά μπορεί να διαφέρουν ανάλογα με το μέγεθος του προϊόντος προς αλληλούχηση, τον τρόπο κατασκευής της βιβλιοθήκης, την επιτυχία φορτώματος του chip κ.α.

	Αλληλουχητής	**Μέγιστο μήκος πειραματικής αλληλουχίας**	**Χρόνος πειραματικής αλληλούχησης**	**Μέγιστο μέγεθος αρχείων εξόδου**	**Πειραματικές αλληλουχίες/ πείραμα αλληλούχησης**
Illumina®	**iSeq100®**	2 × 150 bp	9.5-19 h	1.2 Gb	4 εκ.
	MiniSeq®	2 × 150 bp	4-24 h	7.5 Gb	25 εκ.
	MiSeq®	2 × 300 bp	4-55 h	15 Gb	25 εκ.
	NextSeq 550®	2 × 150 bp	12-30 h	120 Gb	400 εκ.
	NextSeq 1000®	2 × 150 bp	11-48 h	330 Gb	1.1 δις.
	NextSeq 2000®	2 × 150 bp	11-48 h	330 Gb	1.1 δις.
	NovaSeq 6000®	2 × 250 bp	13-44 h	6000 Gb	20 δις.
Ion Torrent™	**PGM™**	200 bp	4.4 h	1 Gb	5.5 εκ.
	Ion Proton™	200 bp	2.5 h	15 Gb	80 εκ.
	Ion S5™	600 bp	3-22 h	50 Gb	130 εκ.
Nanopore®	**MinION™**	> 4 Mb	1 min - 72 h	50 Gb	Μεταβάλλεται ανάλογα με τις πειραματικές συνθήκες
	GridION™	> 4 Mb	1 min - 72 h	50 Gb	
	PromithION™	> 4 Mb	1 min - 72 h	300 Gb	

Πίνακας 3. Συνοπτική περιγραφή των κυριότερων εφαρμογών, που προσφέρουν οι πλατφόρμες αλληλούχησης νέας και τρίτης γενιάς. Στην δεύτερη στήλη παρουσιάζεται το μέγιστο μήκος της πειραματικής αλληλουχίας που μπορεί να προσδιοριστεί με την χρήση της κάθε πλατφόρμας.

	Αλληλουχητής	**Εφαρμογές**
Illumina®	**iSeq100®**	Small WGS, Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων
	MiniSeq®	Small WGS, Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική
	MiSeq®	Small WGS, Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq
	NextSeq 550®	Small WGS, WES, RNA-seq, προφίλ έκφρασης μεμωνομένων κυττάρων (scRNA-seq), Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq, ανίχνευση μεθυλιώσεων, αναλύσεις υγρής βιοψίας
	NextSeq 1000® **NextSeq 2000®**	Small WGS, WES, RNA-seq, προφίλ έκφρασης μεμωνομένων κυττάρων (scRNA-seq), Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq, ανίχνευση μεθυλιώσεων, αναλύσεις υγρής βιοψίας
	NovaSeq 6000®	Large WGS, Small WGS, WES, RNA-seq, προφίλ έκφρασης μεμωνομένων κυττάρων (scRNA-seq), Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq, ανίχνευση μεθυλιώσεων, αναλύσεις υγρής βιοψίας
Ion Torrent™	**PGM™**	WGS, WES, RNA-seq, Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq, *de novo* αλληλούχηση, αναλύσεις υγρής βιοψίας
	Ion Proton™	WGS, WES, RNA-seq, αναλύσεις miRNA & μικρών RNAs, Ampli-seq, Targeted-seq, ChIP-seq, αλληλούχηση του προφίλ έκφρασης γονιδίων, *de novo* αλληλούχηση, μεταγενωμική, επιγενετική
	Ion S5™	WGS, WES, RNA-seq, προφίλ έκφρασης μεμωνομένων κυττάρων (scRNA-seq), Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq, ανίχνευση μεθυλιώσεων, αναλύσεις υγρής βιοψίας, *de novo* αλληλούχηση
Nanopore®	**MinION™**	WGS, WES, RNA-seq, προφίλ έκφρασης μεμωνομένων κυττάρων (scRNA-seq), Targeted-seq, αναλύσεις miRNA & μικρών RNAs, στοχευμένη ανάλυση του προφίλ έκφρασης γονιδίων, μεταγενωμική, ChIP-seq, ανίχνευση μεθυλιώσεων, αναλύσεις υγρής βιοψίας, *de novo* αλληλούχηση, επιγενετικές τροποποιήσεις
	GridION™	
	PromithION™	

Κάθε σύστημα αλληλούχησης προσφέρει επιμέρους πλεονεκτήματα για συγκεκριμένες εφαρμογές. Για την πληρέστερη παρουσίασή τους αλλά και τη διευκόλυνση των επί μέρους συγκρίσεων οι κυριότερες εφαρμογές, που προσφέρουν οι πλατφόρμες αλληλούχησης επόμενης και τρίτης γενιάς, συνοψίζονται στον πίνακα 3.

1.5. Η διαδικασία της ωρίμανσης του mRNA

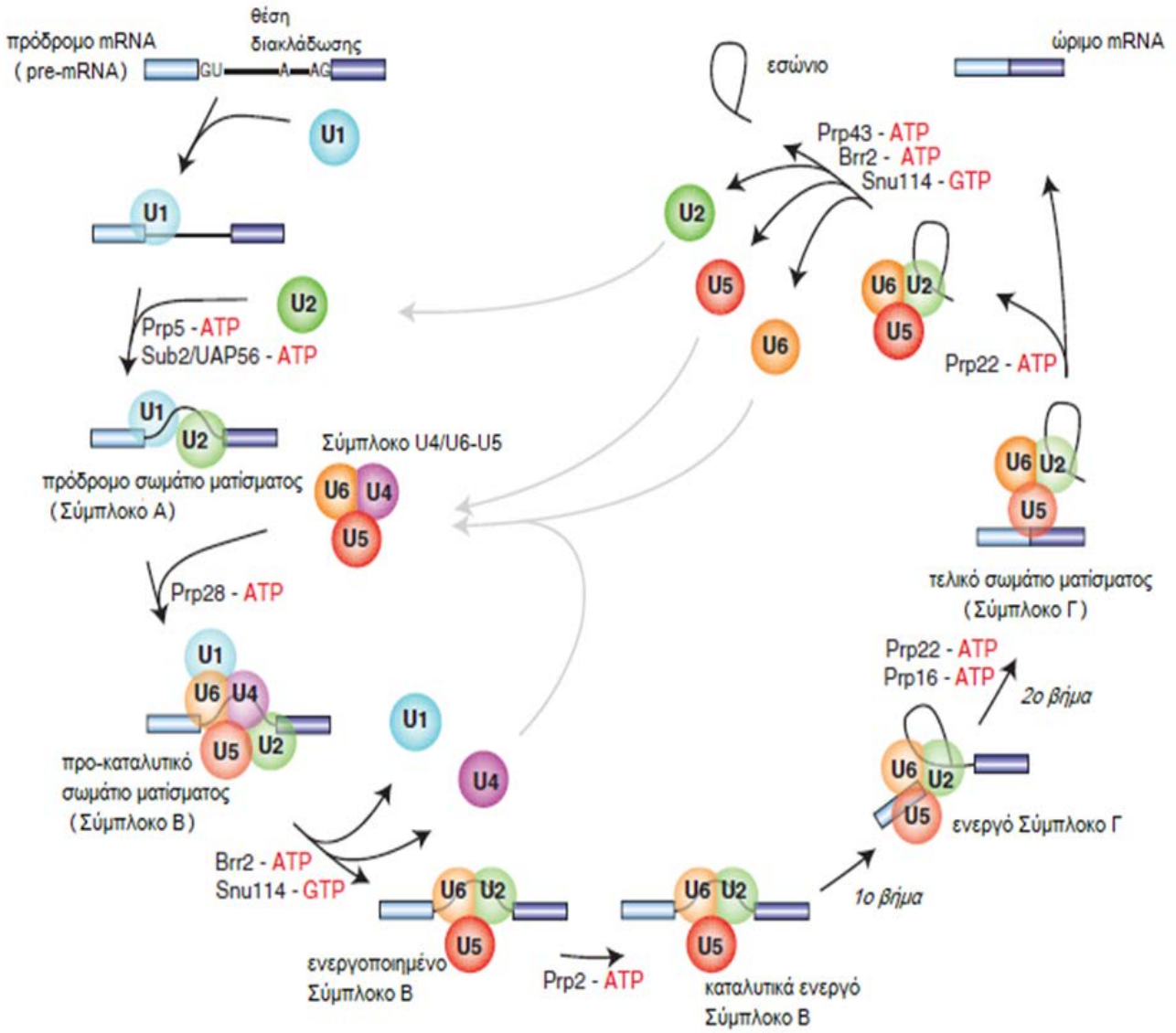
Σύμφωνα με το «Κεντρικό Δόγμα» της Μοριακής Βιολογίας, η γενετική πληροφορία μεταφέρεται από το DNA προς το RNA και, τελικά, στην πρωτεΐνη [85]. Το πρώτο βήμα της έκφρασης των περισσότερων γονιδίων είναι η παραγωγή των μεταφορικών μορίων RNA (mRNAs), τα οποία είναι μόρια κλειδιά για την μεταφορά της κωδικής πληροφορίας στο κυτταρόπλασμα, από το DNA, που βρίσκεται στον πυρήνα, και, επιπλέον, είναι υπεύθυνα για την σύνθεση των πρωτεϊνών από τα ριβοσώματα. Η παρουσία πυρηνικής μεμβράνης στα ευκαρυωτικά κύτταρα οδηγεί στην παραγωγή πρόδρομων μορίων mRNA (pre-mRNAs), κατά τη μεταγραφή του DNA. Τα πρόδρομα mRNAs υποβάλλονται σε διαδικασίες επεξεργασίας του RNA, ώστε να δημιουργηθούν ώριμα μόρια mRNA, τα οποία είναι λειτουργικά. Κατά τη διάρκεια των διαδικασιών αυτών, τα παραγόμενα mRNA μετάγραφα υφίστανται συγκεκριμένες τροποποιήσεις, οι οποίες περιλαμβάνουν τον μηχανισμό κάλυψης του 5’ άκρου, την συρραφή του mRNA και την προσθήκη της πολυαδενυλιωμένης ουράς στο 3’ άκρο κάθε μεταγράφου [86]. Η ωρίμανση (ή συρραφή ή μάτισμα) του mRNA αποτελεί βασικό μηχανισμό κατά τα στάδια της έκφρασης ενός γονιδίου, ο οποίος καταλύεται από τη δράση ειδικών σωματίων συρραφής (spliceosomes).

Η συρραφή του mRNA ανακαλύφθηκε, στο τέλος της δεκαετίας του 1970, στο RNA αδενοϊών, οι οποίοι μόλυναν κύτταρα θηλαστικών, αλλά και σε άλλα ευκαρυωτικά γονίδια, όπως για παράδειγμα τα γονίδια των ανοσοσφαιρινών [87, 88]. Κατά τη διαδικασία της ωρίμανσης, οι μη κωδικές περιοχές του γονιδίου, οι οποίες εμπεριέχονται στο πρόδρομο μόριο mRNA και είναι γνωστές ως εσώνια, αποκόπτονται από την αλληλουχία και απομακρύνονται, ενώ, στη συνέχεια, τα εξώνια, τα οποία αποτελούν τις κωδικές περιοχές, ενώνονται μεταξύ τους και σχηματίζουν ένα ώριμο μόριο mRNA [89]. Η αφαίρεση των εσωνίων βασίζεται στην αναγνώριση καλά συντηρημένων αλληλουχιών, οι οποίες ονομάζονται θέσεις

ματίσματος και βρίσκονται στο 5’ και 3’ άκρο κάθε εσωνίου, όπου και περιέχουν τα δινουκλεοτίδια GU και AG, αντίστοιχα.

Η διαδικασία της ωρίμανσης του mRNA συμβαίνει σε πολλά στάδια και οι αντιδράσεις, που λαμβάνουν χώρα, απαιτούν τη δράση διαφορετικών ριβονουκλεοπρωτεϊνικών ενζύμων (snRNPs), τα οποία δημιουργούνται από σύμπλοκα πέντε μικρών πυρηνικών RNAs (U1, U2, U4, U5, U6) με πυρηνικές πρωτεΐνες. Αρχικά, το 5’ άκρο του πρώτου εσωνίου του πρώιμου mRNA αποκόπτεται, καθώς η ριβονουκλεοπρωτεΐνη U1 προσδένεται στο εσώνιο μέσω συμπληρωματικότητας του U1 snRNA με το εσώνιο του mRNA. Στο εσώνιο υπάρχει η θέση της διακλάδωσης, η οποία φέρει μια συντηρημένη αδενίνη, επιτρέποντας την πρόσδεση του 5’ άκρου του εσωνίου, μέσω της ένωσης της αδενίνης με την γουανίνη, που βρίσκεται στο κομμένο άκρο, με αποτέλεσμα την δημιουργία θηλιάς. Το δεύτερο στάδιο, περιλαμβάνει τη δράση της ριβονουκλεοπρωτεΐνης U2 αλλά, και του συμπλόκου U4/ U6, οι οποίες ενισχύουν τη δομή της θηλιάς, διατηρώντας σταθερούς τους δεσμούς που έχουν αναπτυχθεί στο σημείο της διακλάδωσης. Στο επόμενο στάδιο, οι U2, U5 και U6 πρωτεΐνες σχηματίζουν σύμπλοκο ώστε το 3’ άκρο του πρώτου εξωνίου να έρθει σε επαφή με το 5’ άκρο του επόμενου εξωνίου, τα οποία και θα ενωθούν. Η ριβονουκλεοπρωτεΐνη U5 έχει διπλό ρόλο, ο οποίος βασίζεται στην καταλυτική ιδιότητα του ενζύμου να αποκόπτει το 3’ άκρο του εσωνίου και, στη συνέχεια, να ενισχύει την ένωση αυτού με το 5’ άκρο. Τελευταία αντίδραση είναι η συρραφή μεταξύ των εξωνίων με την δράση της U6 και, τελικά, την απελευθέρωση της θηλιάς, στην οποία είναι προσδεμένες οι U2, U5 και U6 ριβονουκλεοπρωτεΐνες. Η διαδικασία επαναλαμβάνεται προκειμένου να αποκοπούν όλα τα εσώνια που περιέχονται σε κάθε πρόδρομο μόριο mRNA [89, 90].

Η διαδικασίες ματίσματος του mRNA συμβαίνουν χάρη στη δημιουργία του σωματίου συρραφής, το οποίο είναι ένα πρωτεϊνικό σύμπλοκο αποτελούμενο από τις snRNPs και άλλες πρωτεΐνες. Το βασικό σωμάτιο ματίσματος, που δημιουργείται, ονομάζεται μείζον ή U2 σωμάτιο συρραφής και καταλύει τις διαδικασίες αφαίρεσης των περισσότερων πυρηνικών εσωνίων [89, 91], ενώ υπάρχει και ένας ακόμα σπάνιος τύπος σωματίου συρραφής, το U12, το οποίο οφείλεται για το μάτισμα ενός μικρού ποσοστού εσωνίων και αφορά στη δράση της πρωτεΐνης U12 [92].

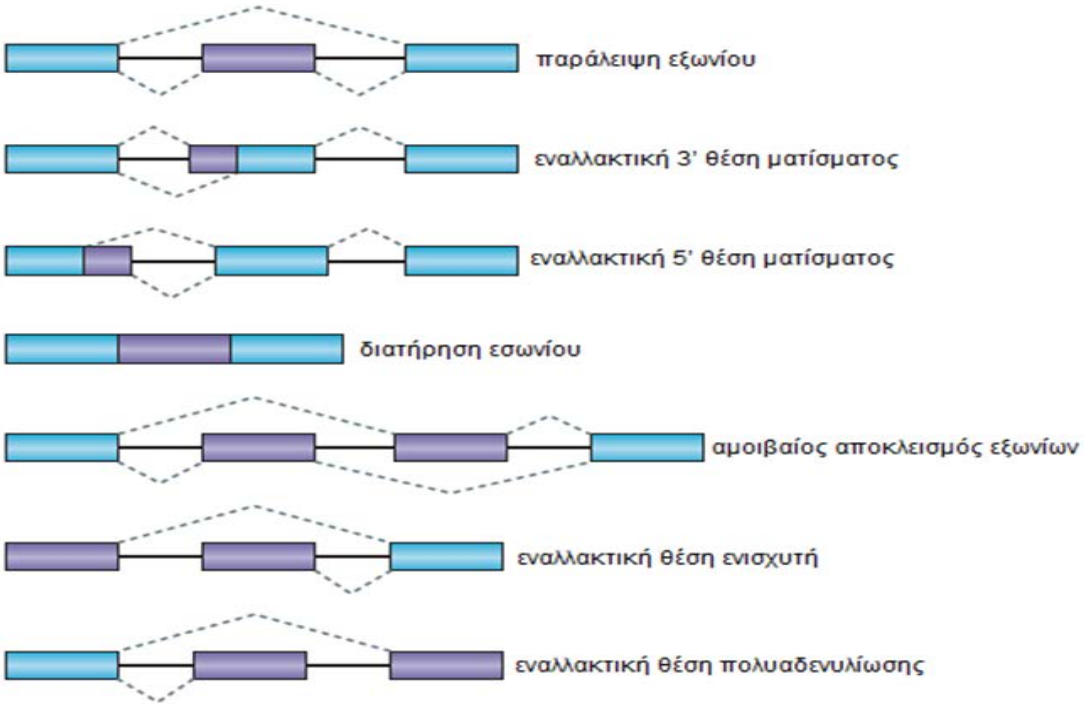


Εικόνα 8. Σχηματική αναπαράσταση της διαδικασίας ματίσματος του πρόδρομου mRNA. Κατά την διαδικασία της ωρίμανσης ριβονουκλεοπρωτεϊνικά μόρια συνδυάζονται μεταξύ τους και δημιουργούν ενεργά σύμπλοκα που καταλύουν την αντίδραση αποκοπής των εσωνίων και την αντίδραση συρραφής των εξωνίων. Τελικά, δημιουργείται το ώριμο mRNA, το οποίο είναι έτοιμο να μεταφραστεί από τα ριβοσώματα στο κυτταρόπλασμα. Επιπλέον, απελευθερώνονται τα εσώνια με τη δομή θηλιάς, ενώ αποδομείται το σύμπλοκο των snRNPs, ώστε να συγκροτηθούν εκ νέου τα αρχικά σωμάτια. (Επεξεργασία από [89]).

1.6. Εναλλακτικό μάτισμα

Στις περισσότερες περιπτώσεις, κατά την ωρίμανση του mRNA των ευκαρυωτικών κυττάρων, δεν παράγεται ένα ώριμο μόριο mRNA αλλά προκύπτουν πολλαπλά μετάγραφα, τα οποία προέρχονται από διαφορετικούς συνδυασμούς συρραφής των εξωνίων, μια διαδικασία που είναι γνωστή ως εναλλακτικό μάτισμα και έχει σαν αποτέλεσμα την παραγωγή ποικίλων πρωτεϊνικών ισομορφών από ένα γονίδιο. Η επιλογή των περιοχών του πρόδρομου mRNA, οι οποίες θα αφαιρεθούν κατά το μάτισμα, καθώς και των περιοχών που θα σχηματίσουν το νέο μετάγραφο, εξαρτάται από ρυθμιστικά στοιχεία του κυττάρου, όπως μη κωδικά RNA και ρυθμιστικές πρωτεΐνες [93]. Ο μηχανισμός του εναλλακτικού ματίσματος οδηγεί στη δημιουργία εναλλακτικών μεταγράφων, τα οποία μπορεί να διαφέρουν μεταξύ τους στο 5’ και 3’ άκρο των αμετάφραστων περιοχών ή στην κωδική αλληλουχία, που φέρουν, διαμέσου τεσσάρων βασικών κατηγοριών εναλλακτικού ματίσματος, οι οποίες είναι: η παράλειψη ενός εξωνίου, η εναλλακτική 5’ θέση ματίσματος, η εναλλακτική 3’ θέση ματίσματος και η διατήρηση εσωνίου [94]. Επιπλέον, στους τύπους εναλλακτικού ματίσματος εντάσσονται και πιο σύνθετα γεγονότα, όπως ο αμοιβαίος αποκλεισμός εξωνίων, η εναλλακτική θέση ενισχυτή και η εναλλακτική θέση πολυαδενυλίωσης [95]. Στην εικόνα 9 παρουσιάζονται, τόσο οι τέσσερεις κύριοι τύποι εναλλακτικού ματίσματος, όσο και οι σπανιότερες περιπτώσεις που οδηγούν στην παραγωγή πολλαπλών μεταγράφων.

Η δυνατότητα των ευκαρυωτικών κυττάρων να χρησιμοποιούν εναλλακτικούς τρόπους ωρίμανσης των πρόδρομων mRNA είναι υψίστης σημασίας, καθώς καταργεί τη θεώρηση ότι η πολυπλοκότητα ενός οργανισμού είναι ανάλογη του αριθμού των γονιδίων του και επιτρέπει στους ανώτερους οργανισμούς την αύξηση της ποικιλομορφίας των μεταγράφων και, κατά συνέπεια, των πρωτεϊνών, που συντίθενται από αυτά. Ωστόσο, οι διαφορές μεταξύ των μεταγράφων ενδέχεται να επηρεάζουν τη σταθερότητα του μορίου, τον εντοπισμό του στο κύτταρο ή την ικανότητα μετάφρασής του και, επομένως, δεν χαρακτηρίζονται όλα τα εναλλακτικά μετάγραφα ενός γονιδίου από την ικανότητα παραγωγής λειτουργικών πρωτεϊνών. Υπάρχουν περιπτώσεις όπου τα εναλλακτικά μετάγραφα, που δημιουργούνται, είναι μη κωδικά και δεν κωδικοποιούν για πρωτεϊνικές ισομορφές, ωστόσο, έχουν ρυθμιστικό ρόλο κατά την γονιδιακή έκφραση [94].



Εικόνα 9. Παρουσίαση των τύπων εναλλακτικού ματίσματος. Οι διαφορετικοί συνδυασμοί των εξωνίων ή η διατήρηση τμημάτων των εσωνίων οδηγεί στην παραγωγή εναλλακτικών μεταγράφων τα οποία προέρχονται από το ίδιο γονίδιο. Η ποικιλία των μεταγράφων που παρατηρείται ευθύνεται για την πολυπλοκότητα των ανώτερων οργανισμών παρά το μικρό αριθμό γονιδίων που διαθέτουν.

Την τελευταία δεκαετία, οι μηχανισμοί επεξεργασίας του RNA, και, ειδικότερα, το εναλλακτικό μάτισμα, το οποίο οδηγεί στην παραγωγή εναλλακτικών mRNA μεταγράφων από ένα μόνο πρόδρομο μόριο mRNA, μελετώνται εκτενώς, και είναι πλέον γνωστό ότι περίπου το 95% των ανθρώπινων γονιδίων εμπλέκονται σε γεγονότα εναλλακτικής συρραφής [96, 97]. Το εναλλακτικό μάτισμα επιτελείται υπό τον έλεγχο ρυθμιστικών συμπλόκων αποτελούμενων από RNA και πρωτεΐνες, των οποίων τα επίπεδα έκφρασης μεταβάλλονται ανάλογα με τον τύπο και την κατάσταση του ιστού [98]. Επιπλέον, η εκδήλωση ανθρώπινων ασθενειών, όπως ο καρκίνος έχει συσχετιστεί με εναλλακτικές μορφές μεταγράφων, και επομένως, η μελέτη του εναλλακτικού τρόπου ωρίμανσης των πρόδρομων μορίων mRNA είναι άκρως σημαντική στην κατανόηση της καρκινογένεσης [99, 100]. Η ταυτοποίηση νέων εναλλακτικών μεταγράφων γονιδίων, τα οποία έχουν συσχετιστεί με την εκδήλωση κακοηθειών, αποτελεί ένα εξαιρετικά ενδιαφέρον πεδίο μελέτης στη Μοριακή Βιολογία, καθώς η παρουσία τους μπορεί να αποτελεί διαγνωστικό, προγνωστικό χαρακτήρα ή / και θεραπευτικό στόχο [91, 101].

1.7. Το γονίδιο *CDK4*

Τα γονίδια των κυκλινο-εξαρτώμενων κινασών (Cyclin-dependent kinases, *CDKs*) είναι υπεύθυνα για τη σύνθεση πρωτεϊνών, που ανήκουν στην οικογένεια των πρωτεϊνικών κινασών, η δράση των οποίων εξαρτάται από την παρουσία κυκλινών. Τα μόρια αυτά χαρακτηρίζονται ως πρωτεϊνικές κινάσες σερίνης / θρεονίνης και η ενεργοποίηση τους απαιτεί τη σύνδεσή τους με κυκλίνες, με αποτέλεσμα τη δημιουργία συμπλόκων, τα οποία εμπλέκονται σε βασικές κυτταρικές διεργασίες, όπως η ρύθμιση του κυτταρικού κύκλου [102-104]. Μέλος της οικογένειας των *CDK* μορίων είναι η κυκλινο-εξαρτώμενη κινάση 4 (Cyclin-dependent kinase 4, *CDK4*). Το ανθρώπινο *CDK4* γονίδιο, το οποίο απαντάται και με τις ονομασίες *CMM3, PSK-J3*, βρίσκεται στο μεγάλο βραχίονα του χρωμοσώματος 12 και συγκεκριμένα στη περιοχή 12q14.1. Έως σήμερα, έχει ταυτοποιηθεί ένα μόνο mRNA μετάγραφο του γονιδίου (GenBank® accession number: NM_000075.4), το οποίο αποτελείται από 8 εξώνια και κωδικοποιεί την κυκλινο-εξαρτώμενη πρωτεϊνική κινάση 4, μία πρωτεΐνη η οποία αποτελείται από 303 αμινοξέα. Το κωδικόνιο έναρξης, το οποίο σηματοδοτεί την έναρξη της πρωτεϊνοσύνθεσης, βρίσκεται στην αρχή του δεύτερου εξωνίου, ενώ το κωδικόνιο λήξης, το οποίο σηματοδοτεί το τερματισμό της σύνθεσης της αμινοξικής αλληλουχίας, βρίσκεται στην αρχή του όγδοου εξωνίου. Η CDK4 πρωτεΐνη κατέχει εξαιρετικά σημαντικό ρυθμιστικό ρόλο κατά τη κυτταρική διαίρεση και τη διαφοροποίηση των ανώτερων ευκαρυωτικών κυττάρων [105, 106].

Όσο αφορά στην πρωτεϊνική δομή της CDK4, η πρωτοταγής αμινοξική ακολουθία, που διαθέτει η πρωτεΐνη, καθορίζει το σχηματισμό των βασικών δομικών και λειτουργικών στοιχείων για την δράση του μορίου ως κινάση. Τα πρώτα 1-96 αμινοξέα της πρωτεΐνης συνιστούν την αμινοτελική περιοχή του μορίου, η οποία έχει στο χώρο δομή β-πτυχωτής επιφάνειας αποτελούμενης από 5 αλυσίδες, ενώ τα υπόλοιπα αμινοξικά κατάλοιπα (αμινοξέα 97-303) συγκροτούν τη βασική δομή α-έλικας στο καρβοξυτελικό άκρο. Το αμινοτελικό και το καρβοξυτελικό άκρο της πρωτεΐνης σχηματίζουν την τρισδιάστατη δομή του μορίου, που έχει σχήμα διπλού λοβού και αποτελεί τυπικό χαρακτηριστικό της ομάδας των CDK ενζύμων.

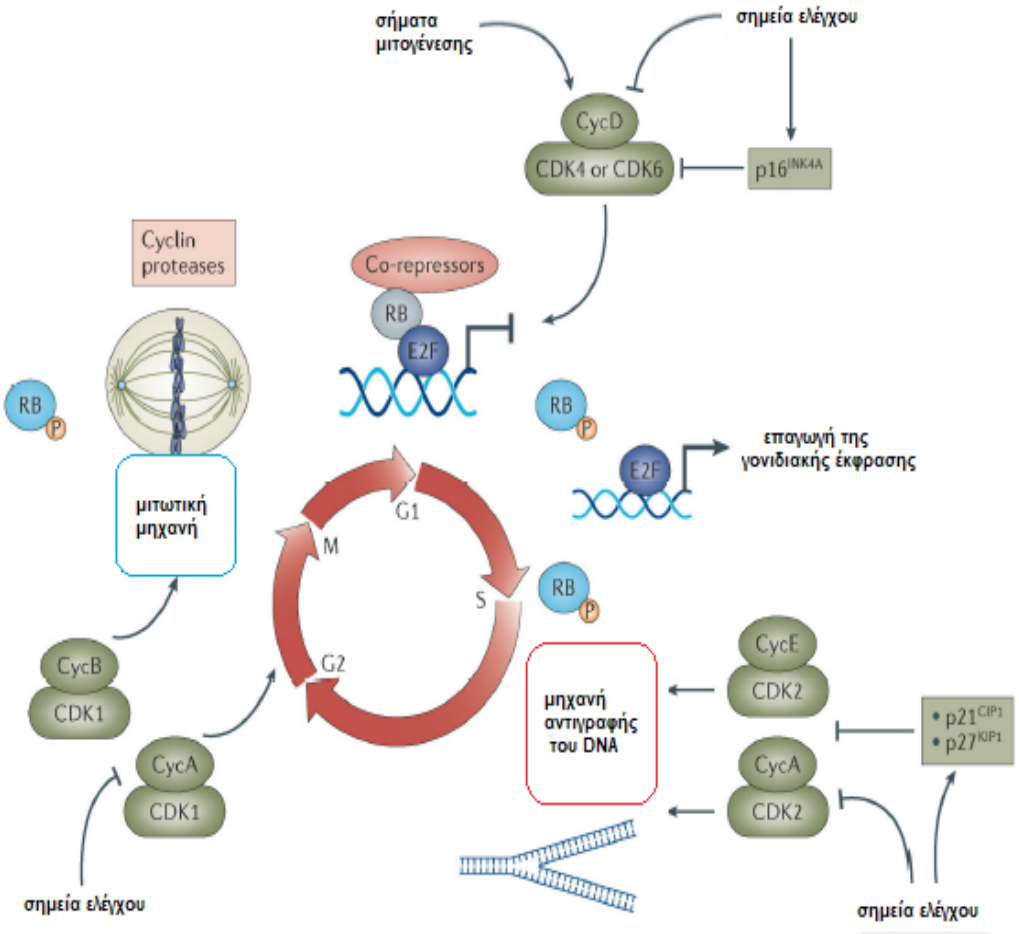
Η οικογένεια των πρωτεϊνικών κινασών επιτελεί βασικές λειτουργίες σηματοδότησης στα κύτταρα και, για αυτό το λόγο, οι περισσότερες κυκλινοεξαρτώμενες κινάσες φέρουν όμοιες συντηρημένες περιοχές στην αμινοξική

τους ακολουθία. Αναφορικά με την κυκλινο-εξαρτώμενη κινάση 4, το αμινοτελικό άκρο της πρωτεΐνης περιλαμβάνει τρεις σημαντικές επικράτειες: α) τη περιοχή πρόσδεσης της κυκλίνης, η οποία αντιστοιχεί στη πρωτεϊνική αλληλουχία PISTVRE, β) μια περιοχή πλούσια σε γλυκίνη, η οποία αποτελεί συντηρημένη περιοχή, καθώς απαντάται στα περισσότερα μέλη της οικογένειας και φέρει το χαρακτηριστικό μοτίβο GXGXXG (GVGAYG συγκεκριμένα για την CDK4), καθώς και, γ) το αμινοξύ, που βρίσκεται στη θέση 35, είναι πάντα μία λυσίνη (K35), η οποία είναι υπεύθυνη για τις αλληλεπιδράσεις μεταξύ των πρωτεϊνών, και χαρακτηρίζεται ως περιοχή δέσμευσης. Επιπλέον, η περιοχή δέσμευσης της κυκλίνης είναι υπεύθυνη για την αλληλεπίδραση της πρωτεΐνης με την κυκλίνη-D1 και ο σχηματισμός ισχυρών δεσμών μεταξύ των δύο πρωτεϊνών οδηγεί στη δημιουργία και ενεργοποίηση του συμπλόκου D1-CDK4 μέσω του οποίου διατηρείται η ομοιόσταση, η ρύθμιση σηματοδοτικών μονοπατιών και η μετάβαση του κυττάρου από τη φάση G1 στη φάση S κατά τη διάρκεια του κυτταρικού κύκλου [107-109]. Η παρουσία του μοτίβου ‘FEHV’, 4 αμινοξέων στο τέλος του αμινοτελικού άκρου, αποτελεί τη γέφυρα που συνδέει την αμινοτελικό άκρο με τη καρβοξυτελική περιοχή της πρωτεΐνης.

Από την άλλη πλευρά, οι κυριότερες επικράτειες, οι οποίες βρίσκονται στο καρβοξυτελικό άκρο, σχετίζονται με την ενεργοποίησης της πρωτεΐνης. Η παρουσία ασπαραγινικού οξέως στη θέση 140 της αλληλουχίας (D140) συνδέεται με την ικανότητα ενεργοποίησης του μορίου. Καθοριστικό ρόλο στις καταλυτικές διαδικασίες φωσφορυλίωσης, που υφίσταται η κινάση, έχει η παρουσία των μοτίβων DFG και APE, που καταλαμβάνουν τις θέσεις 158-160 και 182-184, αντίστοιχα. Η ύπαρξη των συγκεκριμένων μοτίβων αποτελεί σήματα για την ενεργοποίηση της φωσφορυλίωσης, καθώς η αμινοξική αλληλουχία ‘QMALTPVVVTLW’ βρίσκεται ενδιάμεσα στις δύο αυτές περιοχές και συγκροτεί τη θηλιά της φωσφορυλίωσης (T-loop), η οποία είναι υπεύθυνη για τον λειτουργικό ρόλο ολόκληρης της πρωτεΐνης [104, 108, 110]. Συγκεκριμένα, η ύπαρξη μια θρεονίνης στη θέση 172 (T172) της αλληλουχίας είναι υπεύθυνη για τη φωσφορυλίωση της κινάσης και την ενζυματική ενεργοποίηση του μορίου, στην οποία οφείλεται σειρά αντιδράσεων, που λαμβάνουν χώρα κατά τη κυτταρική διαίρεση.

Στην εικόνα 10, που ακολουθεί, παρουσιάζονται συνοπτικά οι βασικές αλληλεπιδράσεις της κινάσης με διαφορετικά πρωτεϊνικά μόρια του κυττάρου. Η ενεργοποίηση της CDK4 πρωτεΐνης πραγματοποιείται μέσω σηματοδοτικών

μονοπατιών, τα οποία οδηγούν την κυκλίνη D στην περιοχή δέσμευσής της στο αμινοτελικό λοβό της κινάσης, με τελικό στόχο την δημιουργία του συμπλόκου της κυκλίνης D με την πρωτεΐνη CDK4. Στη συνέχεια, ακολουθεί η φωσφορυλίωση της κινάσης στην θέση T172 και η ενεργοποίηση του συμπλόκου. Η κεντρική λειτουργία του συμπλόκου είναι η φωσφορυλίωση της πρωτεΐνης του ρετινοβλαστώματος στον πυρήνα, που έχει σαν αποτέλεσμα τον έλεγχο της γονιδιακής έκφρασης και την μετάβαση στη φάση σύνθεσης του DNA στον κυτταρικό κύκλο [111, 112]. Γεγονότα απορρύθμισης των δραστηριοτήτων της πρωτεΐνης CDK4, τα οποία αποτελούν συχνό φαινόμενο σε ένα ευρύ φάσμα κακοηθειών, υποστηρίζουν την εμπλοκή του μορίου τόσο στα στάδια δημιουργίας όγκων όσο και σε περιπτώσεις μετάστασης [103, 105, 113].



Εικόνα 10. Η ενεργοποίηση του συμπλόκου CDK4/CCND1 οδηγεί στη φωσφορυλίωση της πρωτεΐνης του ρετινοβλαστώματος και τελικά στον έλεγχο της γονιδιακής έκφρασης και του κυτταρικού κύκλου (επεξεργασία από [114]).

Συνοψίζοντας, η CDK4 κινάση είναι βασικός ρυθμιστικός παράγοντας των μεταβολών του κυτταρικού κύκλου, που επηρεάζουν την ανάπτυξη και διαφοροποίηση των κυττάρων, αλλά και τις διαδικασίες απόπτωσης και αγγειογένεσης και, για αυτό το λόγο, μεταλλάξεις του γονιδίου *CDK4,* καθώς και η παραγωγή δυσλειτουργικών πρωτεϊνών, έχουν καθοριστικό ρόλο στη ανάπτυξη ανθρώπινων κακοηθειών. Η παρουσία ενός μόνο γνωστού μεταγράφου του *CDK4,* καθώς και η κομβική σημασία της CDK4 κινάσης κατά τον κυτταρικό έλεγχο, γεννούν το ερώτημα της ύπαρξης επιπλέον εναλλακτικών μεταγράφων, που προέρχονται από εναλλακτικό μάτισμα του συγκεκριμένου γονιδίου, τα οποία ενδέχεται να διαθέτουν ουσιαστικό ρόλο στους ρυθμιστικούς μηχανισμούς των ευκαρυωτικών κυττάρων. Επομένως, η ταυτοποίηση νέων εναλλακτικών μεταγράφων του γονιδίου έχει μεγάλη σημασία, καθώς θα μπορούσε να οδηγήσει στη διαλεύκανση σημαντικών ερωτημάτων που σχετίζονται με τη κατανόηση των μηχανισμών που οδηγούν στην εκδήλωση καρκίνου.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

2.1. Βιολογικό υλικό

Στην παρούσα διπλωματική εργασία, για την μελέτη του γονιδίου *CDK4*, χρησιμοποιήθηκαν συνολικά 52 ανθρώπινες κυτταρικές σειρές, οι οποίες προέρχονται από 17 διαφορετικούς ιστούς, και οι οποίες καλλιεργήθηκαν σύμφωνα με τις οδηγίες πρωτοκόλλων του ATCC (American Type Culture Collection). Συγκεκριμένα, οι κυτταρικές σειρές, που χρησιμοποιήθηκαν ήταν: MCF-7, SK-BR-3, BT-20, MDA-MB-231, MDA-MB-468 (αδενοκαρκινώματα μαστού), BT-474, T-47D, ZR-75-1 (πορογενή αδενοκαρκινώματα μαστού), OVCAR-3, SK-OV-3, ES-2, MDAH-2774 (καρκίνος των ωοθηκών), SK-UT-1B, HeLa (αδενοκαρκινώματα του ενδομητρίου), SiHa (καρκινώματα τραχήλου της μήτρας), PC-3, DU 145, LNCaP (καρκίνος του προστάτη), T24, RT4 (καρκίνος της ουροδόχου κύστης), ACHN, 786-O, Caki-1 (καρκινώματα νεφρού), Caco-2, DLD-1, HT-29, HCT 116, SW 620, COLO 205, RKO (καρκίνος του παχέος εντέρου), AGS (γαστρικό αδενοκαρκίνωμα), HepG2, HuH-7 (ηπατοκυτταρικό αδενοκαρκίνωμα), U-87 MG, U-251 MG, D54, H4, SH-SY5Y (καρκίνος του εγκεφάλου), A549 (αδενοκαρκίνωμα του πνεύμονα), FM3, MDA-MB-435S (μελάνωμα), Raji, Daudi, U-937 (λέμφωμα), K-562, HL-60, Jurkat, REC-1, SU-DHL-1, GRANTA-519 (λευχαιμικά κύτταρα), HEK293 (φυσιολογικό εμβρυονικό νεφρό).

2.2. Απομόνωση ολικού RNA

Αρχικά, πραγματοποιήθηκε λύση των κυττάρων, της κάθε κυτταρικής σειράς, σύμφωνα με τις οδηγίες του πρωτοκόλλου TRIzol®. Ακολουθώντας το συγκεκριμένο πρωτόκολλο, επιτυγχάνεται λύση των κυττάρων, ώστε, τελικά, να απελευθερωθεί το ολικό RNA, το οποίο θα απομονωθεί, στη συνέχεια, από το σύνολο των κυτταρικών στοιχείων. Το πρωτόκολλο TRIzol® βασίζεται στη χρήση του αντιδραστηρίου TRI Reagent®, το οποίο αποτελείται από διάλυμα φαινολών αναμειγμένο με διάλυμα ισοθειοκυανικής γουανιδίνης. Τα διαλύματα του αντιδραστηρίου διευκολύνουν την λύση των κυττάρων, λόγω της ικανότητάς τους να αποσταθεροποιούν την μεμβράνη των κυττάρων, αλλά, κυρίως, εμποδίζουν την ενζυματική αποικοδόμηση του RNA, καθώς λειτουργούν ως παρεμποδιστές της δράσης των RNασών, οι οποίες απελευθερώνονται κατά τη διάρκεια της κυτταρικής λύσης. Με βάση τις οδηγίες του πρωτοκόλλου, η λύση των κυττάρων γίνεται σε

ειδικούς σωλήνες φυγοκέντρησης, στους οποίους μεταφέρονται 10 mL κυττάρων, και πραγματοποιείται φυγοκέντρηση 5 λεπτών στις 1000 στροφές (1000 g). Στη συνέχεια, το υπερκείμενο απορρίπτεται και στο σωλήνα προστίθενται 1000 μL αντιδραστηρίου TRI, στο οποίο επαναδιαλυτοποιείται το υπερκείμενο με στόχο την θραύση των κυτταρικών μεμβρανών που έμειναν ακέραιες στο προηγούμενο στάδιο. Ο σωλήνας επωάζεται για 5 λεπτά σε θερμοκρασία δωματίου (25^{o} C) και πραγματοποιείται διαχωρισμός των νουκλεϊκών οξέων από τις πρωτεΐνες. Στο τέλος της αντίδρασης, ακολουθεί η απομόνωση του ολικού RNA. Επιπλέον, δίνεται η δυνατότητα αποθήκευσης του δείγματος για μικρό χρονικό διάστημα (έως 5 μέρες) στους -80^{o} C για την διασφάλιση της ακεραιότητας των μορίων RNA.

Επόμενο βήμα είναι η απομόνωση του ολικού RNA από το σύνολο των κυτταρικών στοιχείων. Η μέθοδος απαιτεί την χρήση χλωροφορμίου, ισοπροπανόλης, διαλύματος αιθανόλης 75% και κατάλληλου διαλύματος αποθήκευσης του RNA (RNA Storage Solution, RSS). Το δείγμα τοποθετείται στον πάγο, αφήνεται να ξεπαγώσει, και αμέσως μετά προστίθενται 200 mL χλωροφορμίου και ακολουθεί ανάμειξη. Το δείγμα επωάζεται για 10 λεπτά σε θερμοκρασία δωματίου και, στη συνέχεια, ακολουθεί φυγοκέντρηση στους 4^{o} C για 15 λεπτά στις 13.000 στροφές / λεπτό. Στο τέλος της φυγοκέντρησης, παρατηρείται ο σχηματισμός τριών διακριτών φάσεων στο δείγμα. Στο κατώτερο τμήμα του σωλήνα βρίσκεται η οργανική φάση, η οποία περιέχει τις πρωτεΐνες, ενώ στο επάνω τμήμα βρίσκεται η υδατική φάση, στην οποία εμπεριέχεται το RNA. Ο διαχωρισμός των δύο φάσεων διακόπτεται από την μεσόφαση, που σχηματίζεται από το DNA, που υπάρχει στο δείγμα. Το διαυγές τμήμα της υδατικής φάσης μεταφέρεται με προσοχή σε νέο σωλήνα, με τη βοήθεια πιπέτας, αποφεύγοντας την ανάδευση. Τυχόν διατάραξη του τριφασικού δείγματος οδηγεί στην ανάμειξη των φάσεων και, επομένως, των συστατικών που περιέχονται στη καθεμία, με τελικό αποτέλεσμα την αδυναμία απομόνωσης του RNA και του διαχωρισμού του από τα υπόλοιπα συστατικά του δείγματος.

Ο σωλήνας, ο οποίος περιέχει τις πρωτεΐνες και το DNA, μπορεί να αποθηκευτεί στους -80^{o} C, ενώ στον νέο σωλήνα που περιέχει το ολικό RNA των κυττάρων προστίθενται 500 μL ισοπροπανόλης, τα οποία και αναδεύονται. Στην συνέχεια, ακολουθεί επώαση για 10 λεπτά και ύστερα φυγοκέντρηση στους 4^{o} C για 8 λεπτά στα 12.000 g. Με τη βοήθεια σύριγγας αφαιρείται το υπερκείμενο και το ίζημα, στο

οποίο βρίσκεται το RNA, επαναδιαλυτοποιείται σε 1000 μL διαλύματος αιθανόλης 75% και ακολουθεί ξανά φυγοκέντρηση στους 4^{o} C για 5 λεπτά στα 12.000 g. Επαναλαμβάνεται η απομάκρυνση του υπερκείμενου και ο σωλήνας, που περιέχει το καθαρό ολικό RNA των κυττάρων, αφήνεται ανοιχτός για 2 λεπτά προκειμένου να εξατμιστούν τυχόν υπολείμματα αιθανόλης. Τέλος, προστίθεται διάλυμα RSS σε ποσότητα ανάλογη του ιζήματος (10-40 μL) και το δείγμα φυλάσσεται στους -80^{o} C.

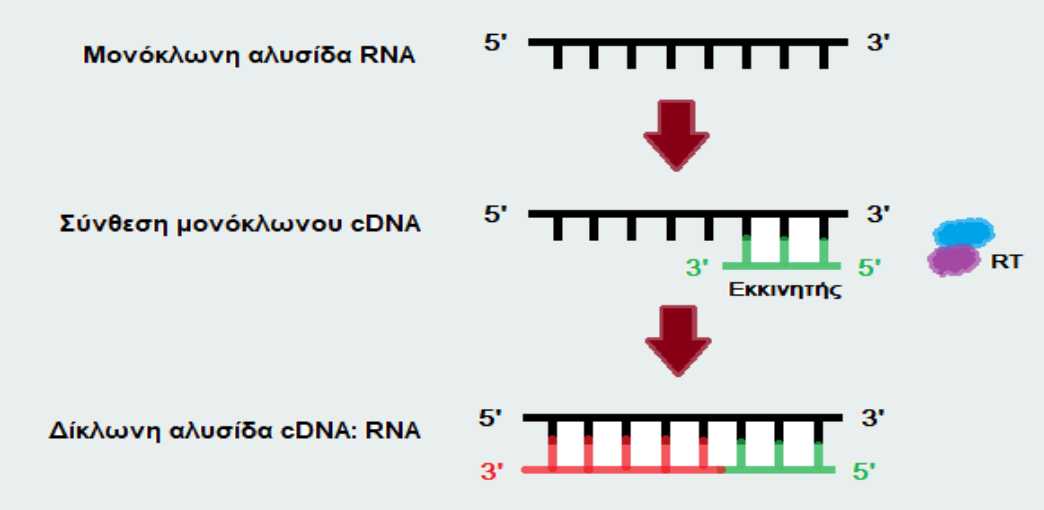
2.3. Φασματοφωτομετρικός προσδιορισμός της ποσότητας και της ποιότητας του RNA

Σε αυτό το στάδιο, δίνεται η δυνατότητα ελέγχου της ποιότητας και της ποσότητας του RNA, που απομονώθηκε, μέσω φωτομέτρησης. Για την διεκπεραίωση της παρούσας διπλωματικής εργασίας, για τον έλεγχο του δείγματος χρησιμοποιήθηκε το φωτόμετρο BioSpec-nano Micro-volume UV-Vis Spectrophotometer. Το φωτόμετρο είναι συνδεδεμένο με Η/Υ και διαθέτει αντίστοιχη εφαρμογή, στην οποία δίνονται, από τον χρήστη, πληροφορίες για το δείγμα, που πρόκειται να φωτομετρηθεί. Λαμβάνεται 1 μL δείγματος, το οποίο τοποθετείται απευθείας σε κατάλληλη θέση στο φωτόμετρο, και ακολουθεί η φωτομέτρηση στα 260 nm, το μήκος κύματος δηλαδή, στο οποίο η απορρόφηση των αζωτούχων βάσεων είναι μέγιστη, και στα 280 nm, όπου παρατηρείται μέγιστη απορρόφηση των αρωματικών αμινοξέων των πρωτεϊνών. Το λογισμικό υπολογίζει αυτόματα τη συγκέντρωση του φωτομετρούμενου δείγματος σε μg / μL. Ο λόγος των απορροφήσεων A_{260} / A_{280} χρησιμοποιείται για την εκτίμηση της καθαρότητας του δείγματος RNA και οι επιθυμητές τιμές του λόγου κυμαίνονται μεταξύ 1,8- 2,2. Τιμές εκτός του εύρους, που αναφέρθηκε, δηλώνουν την παρουσία ανεπιθύμητων προιόντων στο δείγμα που απομονώθηκε. Συγκεκριμένα, αν ο λόγος των απορροφήσεων A_{260}/A_{280} ξεπερνάει την ανώτατη οριακή τιμή (A_{260} / A_{280} > 2,2) υπάρχει ένδειξη ότι στο δείγμα εμπεριέχεται DNA. Αντίθετα, τιμές μικρότερες από 1,8 υποδηλώνουν της ύπαρξη πρωτεϊνών στο δείγμα.

2.4. Αντίστροφη μεταγραφή (Reverse transcription, RT)

Το ένζυμο αντίστροφη μεταγραφάση (Reverse Transcriptase, RT) διαθέτει ξεχωριστή ικανότητα κατάλυσης αντιδράσεων σύνθεσης, στις οποίες χρησιμοποιείται ως εκμαγείο μονόκλωνη RNA αλυσίδα, και το νεοσυντιθέμενο

συμπληρωματικό μόριο, που παράγεται, είναι DNA. Η αντίστροφη μεταγραφάση συνθέτει μονόκλωνες συμπληρωματικές αλυσίδες DNA (complementary, cDNA), με κατεύθυνση 5’ —> 3’, επιμηκύνοντας ολιγονουκλεοτιδικές αλληλουχίες εκκινητών, οι οποίοι έχουν υβριδοποιηθεί με το RNA, που χρησιμοποιείται ως εκμαγείο. Η ανάγκη σύνθεσης υβριδικών μορίων cDNA : RNA, έγκειται στο γεγονός ότι, το ολικό RNA, που απομονώθηκε, βρίσκεται σε μονόκλωνη κατάσταση και είναι πολύ ασταθές, με αποτέλεσμα να δημιουργούνται δυσκολίες κατά το χειρισμό του, σε εργαστηριακό επίπεδο, λόγω του ταχύτατου ρυθμού αποικοδόμησής του.



Εικόνα 11. Διαδικασία σύνθεσης cDNA με τη δράση της αντίστροφης μεταγραφάσης. Η αντίστροφη μεταγραφάση συνθέτει cDNA χρησιμοποιώντας ως εκμαγείο RNA μόριο με κατεύθυνση 5’-3’.

Οι εκκινητές είναι μονόκλωνες συνθετικές αλυσίδες DNA ή RNA, οι οποίοι υβριδοποιούνται μέσω συμπληρωματικότητας, με δεσμούς υδρογόνου, σε μονόκλωνα μόρια DNA ή RNA. Υπάρχουν τρεις τύποι εκκινητών που χρησιμοποιούνται στο εργαστήριο. Ο πρώτος τύπος εκκινητών αφορά τυχαία εξαμερή ολιγονουκλεοτίδια, τα οποία υβριδοποιούνται σε τυχαίες θέσεις στο μονόκλωνο RNA και επιτρέπουν την αντίστροφη μεταγραφή του συνόλου των μορίων, που απομονώθηκαν. Η χρήση αυτού του τύπου εκκινητών διευκολύνει την μελέτη κυρίως των rRNA μορίων, τα οποία αποτελούν το μεγαλύτερο ποσοστό του μεταγραφώματος (> 98%). Ο δεύτερος τύπος εκκινητών αφορά ειδικές νουκλεοτιδικές αλληλουχίες, οι οποίες είναι συμπληρωματικές με ένα μόνο γονίδιο και επιτρέπουν την επιλογή του συγκεκριμένου τμήματος από το σύνολο των μορίων, που υπάρχουν στο δείγμα. Ειδικοί εκκινητές χρησιμοποιούνται για την

μελέτη ενός γονιδίου και έχουν ως αποτέλεσμα την αύξηση της απόδοσης αυξάνοντας την ειδικότητα. Στην παρούσα διπλωματική εργασία χρησιμοποιήθηκε το τρίτο είδος εκκινητών, που είναι τα ολιγομερή θυμίνης (oligo-dT). Τα μετάγραφα, που συντίθενται από την δράση της RNA πολυμεράσης ΙΙ, φέρουν στο 3’ άκρο της αλληλουχίας τους πολυαδενυλιωμένες ουρές. Οι oligo-dT αλληλουχίες χρησιμοποιούνται για την επιλογή των mRNA, καθώς και των lncRNA μορίων, τα οποία φέρουν τις ουρές πολυαδενίνης, και επομένως, δίνουν την δυνατότητα μελέτης του μεταγραφώματος διαχωρίζοντάς το από το σύνολο του RNA των κυττάρων.

Η αντίδραση της αντίστροφης μεταγραφής πραγματοποιείται στο θερμικό κυκλοποιητή και απαιτεί τη παρουσία συγκεκριμένων αντιδραστηρίων, τα οποία προστίθενται σε δύο βήματα. Στο πρώτο βήμα, προστίθεται το δείγμα του RNA, που απομονώθηκε, το νερό, που είναι ελεύθερο από νουκλεάσες (RNase / DNase-free H_2O) για να αποφευχθεί η περίπτωση κατακερματισμού του RNA, ο oligo-dT εκκινητής, ο οποίος θα υβριδοποιηθεί με το RNA, και τα dNTPs, τα οποία θα ενσωματωθούν στο νεοσυντιθέμενο κλώνο. Στο επόμενο βήμα, προστίθενται ρυθμιστικό διάλυμα, το οποίο ρυθμίζει το pH του διαλύματος και αποτελείται από 20 mM Tris-HCl, 100 mM NaCl, 0,1 mM EDTA, 1 mM DTT, 0,01% (v/v) NP-40 και 50% (v/v) γλυκερόλη, διάλυμα DTT, το οποίο σταθεροποιεί το ένζυμο αντίστροφη μεταγραφάση, αναστολέας RNασών (RNaseOUT inhibitor), καθώς και το ένζυμο.

Η αντίστροφη μεταγραφή, που πραγματοποιήθηκε στη παρούσα διπλωματική εργασία, ακολούθησε συγκεκριμένο πρωτόκολλο, το οποίο εφαρμόζεται στο εργαστήριο Βιοχημείας του τμήματος Βιολογίας του ΕΚΠΑ, για τη σύνθεση μονόκλωνης cDNA αλυσίδας. Συνοπτικά, 2 μg του ολικού RNA, που απομονώθηκε με τη χρήση των προαναφερθέντων μεθόδων, από κάθε κυτταρική σειρά, χρησιμοποιήθηκε ως εκμαγείο για τη σύνθεση του αντίστοιχου συμπληρωματικού κλώνου. Ο oligo-dT εκκινητής, που χρησιμοποιήθηκε, σχεδιάστηκε κατάλληλα, ώστε να υβριδοποιείται στο 3’ άκρο των mRNA μεταγράφων, το οποίο φέρει την πόλυ(Α) ουρά. Η νουκλεοτιδική αλληλουχία του εκκινητή, που χρησιμοποιήθηκε, είναι η εξής: 5′ - GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTTTTTTTVN - 3′ όπου το V μπορεί να είναι οποιαδήποτε από τις τρεις αζωτούχες βάσεις αδενίνη, γουανίνη ή κυτοσίνη αλλά όχι θυμίνη, γεγονός που υποδηλώνει ότι το συγκεκριμένο νουκλεοτίδιο θα υβριδοποιηθεί ακριβώς στη τελευταία θέση πριν την έναρξη της

πολυ(Α) ουράς και το Ν αντιπροσωπεύει οποιοδήποτε από τα 4 νουκλεοτίδια. Εν συντομία, στο πρώτο στάδιο προστέθηκαν 4,5 μL διαλύματος cDNA αντιδραστηρίων, το οποίο περιείχε 2 μl ολικού RNA (2 μg), 1 μl oligo-dT εκκινητή (10 μM) και 1,5 μl καθαρό H_2O και ακολούθησε η επώασή του στο θερμικό κυκλοποιητή Veriti 96-Well Fast Thermal Cycler (Applied Biosystems™) στους 65° C για 5 λεπτά με σκοπό την αποδιάταξη των δευτεροταγών δομών του RNA. Στο τέλος της επώασης, το μείγμα των αντιδραστηρίων μεταφέρθηκε στον πάγο. Κατόπιν, προστέθηκαν 2 μl ρυθμιστικό διάλυμα, 0,25 μl DTT (100mM), 1 μl μείγμα dNTPs (10mM each), 0,25 μl (10 U) αναστολέας Rnασών (RNaseOUT inhibitor Invitrogen™, Thermo Fisher Scientific Inc.) και 1,75 μl (75 U) από το ένζυμο αντίστροφη μεταγραφάση, το οποίο στο συγκεκριμένο πείραμα ήταν η SMARTScribe™ (Takara Bio, Inc). Στο τέλος του δεύτερου σταδίου, πραγματοποιήθηκε η αντίδραση αντίστροφης μεταγραφής στο θερμικό κυκλοποιητή στους 42° C για 70 λεπτά.

Το ένζυμο SMARTScribe™, που χρησιμοποιήθηκε για την σύνθεση του cDNA, διαθέτει τις εξής ιδιότητες: 1) έχει ικανότητα πολυμερισμού πολύ μεγάλων τμημάτων RNA, μήκους μεγαλύτερου των 14,7 kb, 2) συμβάλλει στην ενίσχυση σπάνιων μεταγράφων, 3) είναι ικανή να διατηρεί την πολυπλοκότητα, που εμφανίζεται στο RNA, καθώς και, 4) είναι κατάλληλη για την ανάλυση των 5’ άκρων των μεταγράφων, αφού διαθέτει ενεργότητα τελικής τρανσφεράσης. Συγκεκριμένα, καθώς το ένζυμο φτάσει το 5’ άκρο του mRNA και ακριβώς πριν τερματιστεί η σύνθεση του συμπληρωματικού DNA η SMARTScribe™ μπορεί να προσθέσει λίγα νουκλεοτίδια (κυρίως dCTPs) στο 3’ άκρο του cDNA μορίου χωρίς την παρουσία εκμαγείου. Τέλος, η αντίδραση τερματίστηκε με αύξηση της θερμοκρασίας στους 70° C για 10 λεπτά.

2.5. Ποιοτικός έλεγχος της αντίστροφης μεταγραφής

Η αξιολόγηση της επιτυχίας της αντίδρασης αντίστροφης μεταγραφής γίνεται με την ανίχνευση γονιδίων αναφοράς. Τα γονίδια αναφοράς (housekeeping genes) είναι γονίδια, που απαιτούνται για τις βασικές λειτουργικές διεργασίες των κυττάρων ενός οργανισμού τόσο υπό φυσιολογικές όσο και υπό παθολογικές συνθήκες, και επομένως, εκφράζονται καθολικά και σταθερά σε όλα τα κύτταρα [115, 116]. Πολλά γονίδια αναφοράς έχουν χαρακτηριστεί και χρησιμοποιούνται ως δείκτες για τον

έλεγχο της ποιότητας της αντίστροφης μεταγραφής. Κύριοι αντιπρόσωποι είναι: μεταγραφικοί παράγοντες (πχ. *E2F4, BTF3*), μεταφραστικοί παράγοντες (πχ. *EIF1, EIF2A*), γονίδια που κωδικοποιούν για ριβοσωμικές πρωτεΐνες (πχ. *RPLs*) κ.α.

Στη παρούσα διπλωματική εργασία, χρησιμοποιήθηκε ως γονίδιο αναφοράς, το γονίδιο διυδροξυγενάση της τριφωσφορικής γλυκεραλδεϋδης (*GAPDH*), το οποίο χαρακτηρίζεται από υψηλή έκφραση σε πολλούς κυτταρικούς τύπους. Το γονίδιο *GAPDH* συμμετέχει σε αντιδράσεις του μεταβολισμού των κυττάρων και, επιπλέον, έχει συσχετιστεί με μη-μεταβολικές διεργασίες, όπως η μεταγραφή και η απόπτωση [117]. Ο προσδιορισμός του γονιδίου *GAPDH,* ως προϊόν της αντίστροφης μεταγραφής, που πραγματοποιήθηκε, έγινε με τη διαδικασία της ηλεκτροφόρησης για την ανίχνευση της ζώνης, που σχηματίζει ύστερα από την ενίσχυσή του μέσω της αλυσιδωτής αντίδρασης πολυμερισμού (PCR). Η παρουσία ζώνης στο πήκτωμα της ηλεκτροφόρησης υποδηλώνει την επιτυχία της αντίδρασης σύνθεσης cDNA.

2.6. Αλυσιδωτή Αντίδραση Πολυμεράσης (Polymerase Chain Reaction, PCR)

Η μέθοδος αλυσιδωτής αντίδρασης πολυμεράσης (Polymerase Chain Reaction, PCR) χρησιμοποιείται in vitro για την ενίσχυση ενός DNA-στόχου, το οποίο εμπεριέχεται σε ένα σύνολο μορίων DNA, με σκοπό την ταυτοποίησή του, τον προσδιορισμό της ακολουθίας του, τη διάγνωση γενετικών ασθενειών και την ανάλυση των αλληλομόρφων. Η αντίδραση PCR βασίζεται στη ικανότητα πολυμερισμού του ενζύμου της DNA πολυμεράσης να καταλύει την προσθήκη ενός τριφωσφορικού δεοξυριβονουκλεοτιδίου (dNTP) στο 3’ άκρο μιας DNA αλυσίδας δημιουργώντας ένα φωσφοδιεστερικό δεσμό με ταυτόχρονη απελευθέρωση ενός μορίου πυροφωσφορικού. Το ένζυμο απαιτεί την ύπαρξη μιας μονόκλωνης αλυσίδας DNA, η οποία θα χρησιμοποιηθεί ως εκμαγείο για την επιμήκυνση του συμπληρωματικού του κλώνου. Ωστόσο, η DNA πολυμεράση δεν έχει την ικανότητα σύνθεσης εκ νέου της συμπληρωματικής αλυσίδας DNA, αλλά προσθέτει dNTPs στο ελεύθερο 3’ άκρο των συμπληρωματικών κλώνων του εκμαγείου. Επομένως, για την δράση της πολυμεράσης απαιτείται η παρουσία μονόκλωνων τμημάτων DNA, τα οποία είναι συμπληρωματικά με το DNA-στόχο, ώστε να υβριδοποιηθούν σε αυτό και να επιτρέψουν, στη συνέχεια, τη πρόσδεση και τη δράση της DNA πολυμεράσης. Σε εργαστηριακό επίπεδο, τα μονόκλωνα αυτά τμήματα DNA είναι

συνθετικά ονομάζονται εκκινητές (primers) και αποτελούνται, κατά μέσο όρο, από 18-25 νουκλεοτίδια.

Σε κάθε PCR σχεδιάζεται ένα ζεύγος εκκινητών κατάλληλο ώστε να προσδεθεί βάση συμπληρωματικότητας στα άκρα του τμήματος του DNA-στόχου. Επιπλέον, η δράση των DNA πολυμερασών, που χρησιμοποιούνται εργαστηριακά, ελέγχεται από την θερμοκρασία. Κάθε ένζυμο, που χρησιμοποιείται, χαρακτηρίζεται από τη βέλτιστη θερμοκρασία δράσης του. Η PCR χωρίζεται σε 3 βασικά στάδια, τα οποία χαρακτηρίζονται από διαφορετικό χρόνο επώασης, καθώς και διαφορετική θερμοκρασία και επαναλαμβάνονται, με σκοπό τον πολλαπλασιασμό των μορίων-στόχων. Η αντιδράσεις PCR λαμβάνουν χώρα σε ειδικούς θερμικούς κυκλοποιητές, οι οποίοι διαμορφώνουν την κατάλληλη θερμοκρασία. Αναλυτικότερα, τα στάδια της μεθόδου είναι τα εξής:

Αποδιάταξη - denaturation: Το δείγμα DNA, που πρόκειται να ενισχυθεί, αποδιατάσσεται με θέρμανση στους 95^{o} C. Στους 95^{o} C, οι δεσμοί υδρογόνου μεταξύ των συμπληρωματικών βάσεων του δίκλωνου DNA διασπώνται και, επομένως, δημιουργούνται μονόκλωνες αλυσίδες DNA, οι οποίες αποτελούν το εκμαγείο για τη σύνθεση νέων μορίων.

Υβριδοποίηση - annealing: Οι εκκινητές υβριδοποιούνται στις συμπληρωματικές αλληλουχίες, που βρίσκονται στο 3’ άκρο, του μονόκλωνου DNA-στόχου σε κατάλληλη θερμοκρασία υβριδοποίησης (T_a), η οποία εξαρτάται από τη θερμοκρασία τήξης των εκκινητών (T_m) και κυμαίνεται μεταξύ 45^{o} - 65^{o} C.

Επιμήκυνση - Elongation: Πραγματοποιείται επιμήκυνση του DNA στόχου με θέρμανση. Η θερμοκρασία στο στάδιο αυτό καθορίζεται από τη βέλτιστη θερμοκρασία του ενζύμου DNA πολυμεράση, που χρησιμοποιείται σε κάθε αντίδραση PCR, πχ. η Taq πολυμεράση, ένα ένζυμο, το οποίο χρησιμοποιείται ευρύτατα σε αντιδράσεις PCR δρα σε βέλτιστη θερμοκρασία 72^{o} C. Η πολυμεράση προσδένεται στον εκκινητή και επιμηκύνει τον νεοσυντιθέμενο κλώνο, σύμφωνα με τον κανόνα της συμπληρωματικότητας του DNA. Στο τέλος του τρίτου αυτού βήματος, και οι δύο αρχικοί κλώνοι του δείγματος DNA έχουν αντιγραφεί.

Κάθε κύκλος αντιδράσεων περιλαμβάνει τα τρία στάδια που αναφέρθηκαν παραπάνω. Ακολουθεί επανάληψη των σταδίων για 25-40 κύκλους αντιδράσεων με σκοπό την εκθετική αύξηση της αλληλουχίας, που ενισχύεται. Οι κύκλοι της αντίδρασης καθορίζονται ανάλογα με τον αριθμό αντιγράφων που επιθυμείται να

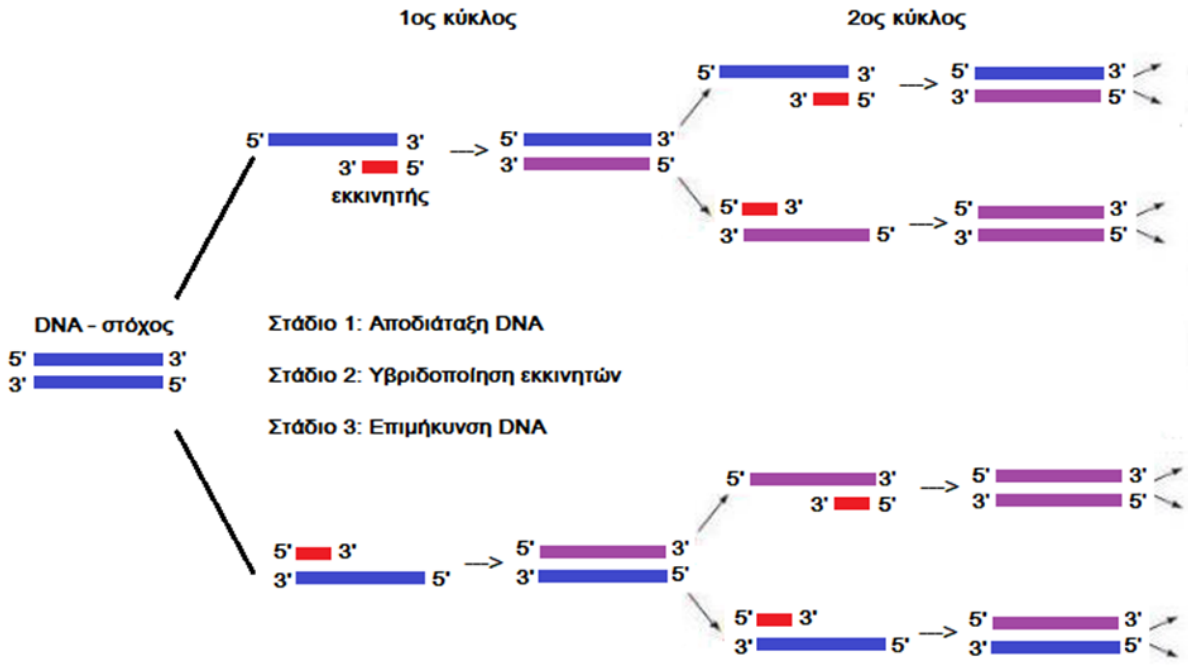
προκύψουν. Τελικό προϊόν μιας PCR αντίδρασης είναι η δημιουργία δίκλωνων μορίων DNA το πλήθος των οποίων υπολογίζεται από τον τύπο:

$$dsDNA= 2^n$$

όπου dsDNA: τα δίκλωνα μόρια DNA που αποτελούν το PCR προϊόν της αντίδρασης, 2: οι δύο αλυσίδες ενός αρχικού μορίου DNA και n: ο αριθμός των κύκλων μίας PCR αντίδρασης.

Στο τέλος των επαναλαμβανόμενων κύκλων, δίνεται ένα χρονικό διάστημα ~5 λεπτών κατά το οποίο η DNA πολυμεράση ολοκληρώνει τη σύνθεση τυχών ημιτελών αλυσίδων στην ίδια θερμοκρασία με το στάδιο της επιμήκυνση, το οποίο καλείται τελική επιμήκυνση (final elongation).

Τα προϊόντα μιας PCR αντίδρασης περιλαμβάνουν ένα πανομοιότυπο σύνολο μορίων DNA, το οποίο είναι το αποτέλεσμα της ενίσχυσης του αρχικού μορίου-στόχου, που έχει επιλεγεί και ορίσθηκε από τις θέσεις υβριδοποίησης των εκκινητών. Το προϊόν μιας PCR αντίδρασης μπορεί να αποθηκευτεί στους 4° C, καθώς πρόκειται για δίκλωνα μόρια DNA, τα οποία βρίσκονται σε μεγάλη συγκέντρωση μέσα στο διάλυμα και παραμένουν σταθερά.



Εικόνα 12. Αρχή μεθόδου αλυσιδωτής αντίδρασης πολυμεράσης. Τα τρία στάδια: Αποδιάταξη, Υβριδοποίηση, Επιμήκυνση επαναλαμβάνονται για n κύκλους αντιδράσεων.

Στην παρούσα εργασία, πραγματοποιήθηκε απλή PCR αντίδραση σε 17 διαφορετικά δείγματα, τα οποία προέρχονται από τους 17 διαφορετικούς ιστούς, που αναφέρθηκαν προηγουμένως, και ακολούθησε ηλεκτροφόρηση των προϊόντων, με σκοπό την διερεύνηση του προφίλ έκφρασης του γονιδίου *CDK4* σε αυτούς τους ιστούς.

Πιο συγκεκριμένα, η ενίσχυση του γονιδίου *CDK4,* από το σύνολο των cDNAs που δημιουργήθηκαν, πραγματοποιήθηκε μέσω μιας ειδικής αντίδρασης PCR, η οποία ονομάζεται Touchdown PCR. Η διαφορά της Touchdown PCR με την απλή PCR, η οποία περιγράφθηκε προηγουμένως, βρίσκεται στο θερμικό πρωτόκολλο που ακολουθείται κατά το στάδιο υβριδοποίησης των εκκινητών. Συγκεκριμένα, στον πρώτο κύκλο αντίδρασης η θερμοκρασία υβριδοποίησης των εκκινητών, T_a, ρυθμίζεται ώστε να είναι κατά 5^o C υψηλότερη από την θερμοκρασία τήξης των εκκινητών, T_m, ώστε να ευνοείται η πρόσδεση των εκκινητών στην αλληλουχία-στόχο, όπως ακριβώς συμβαίνει και στο πρωτόκολλο της απλής PCR. Ωστόσο, κατά την εφαρμογή Touchdown PCR, η θερμοκρασία υβριδοποίησης σε κάθε κύκλο της αντίδρασης μειώνεται σταθερά κατά 0,3^o C / κύκλο, έως ότου να εξισωθούν οι T_a και T_m. Η σταδιακή μείωση της θερμοκρασίας υβριδοποίησης οδηγεί στην αύξηση της ειδικότητας της αντίδρασης, και επομένως, και στην αύξηση της απόδοσης.

Αρχικά, έγινε ο σχεδιασμός του κατάλληλου ζεύγους εκκινητών, ώστε να υβριδοποιούνται στα άκρα του γονιδίου, και επομένως, η στόχευση να αφορά στο ολικό *CDK4* μετάγραφο και όχι ένα μικρό τμήμα αυτού. Το ζεύγος εκκινητών που επιλέχθηκε ήταν οι εκκινητές 5’ – GTGTATGGGGCCGTAGGAAC – 3’ και 5’ – AGCCACTCCATTGCTCACTC – 3’, το οποίο στοχεύει τα εξώνια 1 και 8, αντίστοιχα, δηλαδή τα εξώνια που βρίσκονται στα άκρα του μορίου και οριοθετούν το μετάγραφο *CDK4*. Ο σχεδιασμός των εκκινητών πραγματοποιήθηκε με τη χρήση του εργαλείου primer BLAST, το οποίο υπολογίζει και εμφανίζει αυτόματα τα ιδιαίτερα χαρακτηριστικά κάθε εκκινητή, όπως το σημείο τήξης (T_m), το ποσοστό GC, τον βαθμό αυτοσυμπληρωματικότητας, καθώς και τις αλληλουχίες με τις οποίες εμφανίζει συμπληρωματικότητα, και επομένως, στοχεύει. Όπως γίνεται αντιληπτό, κατά τον σχεδιασμό εκκινητών κατάλληλος θεωρείται ο εκκινητής που στοχεύει συγκεκριμένα την επιθυμητή αλληλουχία, που πρόκειται να ενισχυθεί. Αν το ζεύγος εκκινητών, που σχεδιάστηκαν, δεν είναι ειδικό ως προς την αλληλουχία-στόχο και

μπορεί να υβριδοποιηθεί και με άλλα μόρια, που βρίσκονται στο μείγμα του cDNA, υπάρχει αυξημένος κίνδυνος ενίσχυσης και των αλληλουχιών αυτών, οι οποίες θα αποτελούν ανεπιθύμητα παραπροϊόντα. Το σημείο τήξης ενός εκκινητή είναι η θερμοκρασία στην οποία το 50% των μορίων αποδιατάσσονται και επανέρχονται σε μονόκλωνη κατάσταση. Το σημείο τήξης μπορεί να υπολογιστεί από τον τύπο:

$$T_m = (\text{αριθμός G+C})\times 4°C + (\text{αριθμός A+T})\times 2°C$$

Επομένως, όσο αυξάνεται το μέγεθος του εκκινητή, αλλά και η περιεκτικότητα του σε GC, αυξάνεται και η θερμοκρασία τήξης. Τέλος, ο βαθμός αυτοσυμπληρωματικότητας αναφέρεται στην ικανότητα του εκκινητή να υβριδοποιείται με τον εαυτό του, με αποτέλεσμα την δημιουργία δομής φουρκέτας, που οδηγεί σε μειωμένη ικανότητα υβριδοποίησης των εκκινητών στο DNA- στόχο και την μείωση της απόδοσης της συνολικής αντίδρασης.

Τα αντιδραστήρια, που απαιτούνται για την πραγματοποίηση μιας PCR, είναι: νερό ελεύθερο από νουκλεάσες, ρυθμιστικό διάλυμα, ιόντα Mg^{++}, μείγμα dNTPs, το ζεύγος εκκινητών, DNA πολυμεράση και το υπόστρωμα, δηλαδή το δείγμα του DNA. Κατά ενίσχυση του *CDK4*, η DNA πολυμεράση, που χρησιμοποιήθηκε, ήταν η Taq πολυμεράση (Kapa Biosystems), η οποία συνοδεύεται από το ρυθμιστικό διάλυμα KAPA B buffer, το οποίο περιέχει Mg^{++} με τελική συγκέντρωση αντίδρασης 1,5 mM. Πιο αναλυτικά, στο σωλήνα της αντίδρασης προστέθηκαν με την σειρά οι ποσότητες των αντιδραστηρίων, οι οποίες αναγράφονται στον πίνακα 4 που ακολουθεί.

Πίνακας 4. Τα αντιδραστήρια που χρησιμοποιήθηκαν για την ενίσχυση του *CDK4* με την μέθοδο PCR. Οι όγκοι που χρησιμοποιήθηκαν, αναφέρονται σε μοναδιαία αντίδραση 25 μL.

Αντιδραστήριο	**Όγκος (μL)**
dH_2O	18,9
10X KAPA Taq ρυθμιστικό διάλυμα*	2,5
dNTPs (10 mM)	0,5
Πρόσθιος εκκινητής (10 μM)	1
Ανάστροφος εκκινητής (10 μM)	1
5 U/μL KAPA Taq πολυμεράση	0,1
Δείγμα DNA (template)	1
Σύνολο	**25**

*Στη παρούσα διπλωματική χρησιμοποιήθηκε το ρυθμιστικό διάλυμα KAPA Taq A, το οποίο περιέχει Mg^{++} σε τελική συγκέντρωση 1,5mM σε αραίωση 1X.

Πραγματοποιείται ανάδευση και σύντομη φυγοκέντρηση (spin down) του σωλήνα, που περιέχει τα αντιδραστήρια, ώστε η συνολική ποσότητα του διαλύματος να βρίσκεται στο πυθμένα του σωλήνα. Ο σωλήνας τοποθετείται στον θερμικό κυκλοποιητή, ο οποίος έχει ρυθμιστεί σύμφωνα με τις οδηγίες του πρωτοκόλλου, ώστε να επιτευχθεί εκθετική αύξηση των επιθυμητών αλληλουχιών ύστερα από την ολοκλήρωση των κύκλων της αντίδρασης πολυμερισμού. Το θερμικό πρωτόκολλο, που χρησιμοποιήθηκε, παρουσιάζεται στον πίνακα 5, που ακολουθεί.

Πίνακας 5. Το θερμικό πρωτόκολλο που ακολουθήθηκε κατά την ενίσχυση του *CDK4* μεταγράφου με τη χρήση Touchdown PCR. Κατά το στάδιο της υβριδοποίησης πραγματοποιείται σταδιακή μείωση κατά 0,3° C, σε κάθε κύκλο αντίδρασης, με σκοπό την αύξηση της απόδοσης της αντίδρασης.

Στάδια	**Θερμοκρασία (° C)**	**Χρόνος**	**Αριθμός κύκλων**
Αρχική αποδιάταξη	95	3 min	1
Αποδιάταξη	95	30 sec	30
Υβριδοποίηση	65 (ΔT_a: -0,3/ κύκλο)	30 sec	
Επιμήκυνση	72	2 min	
Τελική επιμήκυνση	72	2 min	1
Διατήρηση	4	∞	1

2.7. Ηλεκτροφόρηση σε πήκτωμα αγαρόζης

Ο ποιοτικός έλεγχος της αντίδρασης PCR πραγματοποιήθηκε με την ηλεκτροφόρηση μέρους της ποσότητας του PCR προϊόντος σε πήκτωμα αγαρόζης. Η αρχή μεθόδου της ηλεκτροφόρησης βασίζεται στην ικανότητα του DNA να κινείται προς το θετικό πόλο όταν βρεθεί σε ένα ηλεκτρικό πεδίο, καθώς το μόριο αυτό είναι αρνητικά φορτισμένο. Σε κατάλληλες συνθήκες, η αγαρόζη δημιουργεί πορώδες πήκτωμα, στο οποίο γίνεται η ηλεκτροφόρηση. Η συγκέντρωση της αγαρόζης καθορίζει το μέγεθος των πόρων και, επομένως, ρυθμίζει το μέγεθος των τμημάτων DNA, καθώς και την ταχύτητα με την οποία θα ηλεκτροφορηθούν. Η ταχύτητα με την οποία κινείται το DNA εξαρτάται από το ρυθμιστικό διάλυμα τη ηλεκτροφόρησης, το μέγεθος της εφαρμοζόμενης τάσης, την ιονική ισχύ και κυρίως από το μοριακό βάρος της αλληλουχίας που ηλεκτροφορείται, και επομένως, η τεχνική

χρησιμοποιείται για τον διαχωρισμό νουκλεϊκών οξέων διαφορετικού μεγέθους. Κατά τη διάρκεια της εφαρμογής ηλεκτρικού πεδίου στο κύκλωμα το DNA μετακινείται προς την άνοδο με τα τμήματα μικρότερου μοριακού βάρους να κινούνται ταχύτερα σε σχέση με τα τμήματα μεγαλύτερους μήκους.

Για τη δημιουργία του πηκτώματος αγαρόζης, στο οποίο θα πραγματοποιηθεί το φόρτωμα του DNA και η ηλεκτροφόρηση του, απαιτείται ρυθμιστικό διάλυμα TBE (Tris/ Borate/ EDTA), το οποίο ρυθμίζει το pH του πηκτώματος περίπου στο 8,6, κατάλληλη ποσότητα αγαρόζης,, η οποία διαμορφώνει το μέγεθος των πόρων, καθώς και βρωμιούχο αιθίδιο (EtBr). Το βρωμιούχο αιθίδιο είναι φθορίζουσα χρωστική που έχει την ικανότητα να ενσωματώνεται στη δίκλωνη αλυσίδα του DNA και με τη χρήση υπεριώδους ακτινοβολίας εμφανίζεται στο πήκτωμα η ζώνη του DNA μετά το πέρας της ηλεκτροφόρησης. Η παρουσία ζωνών DNA στο πήκτωμα επιτρέπει την αξιολόγηση του δείγματος, που ηλεκτροφορείται, καθώς μπορεί να αξιολογηθεί το μήκος κάθε τμήματος DNA (ζώνης) με τη βοήθεια πρότυπου δείγματος, το οποίο περιέχει τμήματα DNA με γνωστό μοριακό βάρος.

2.8. Καθαρισμός PCR προϊόντων

Το προϊόν κάθε αντίδρασης PCR αποτελείται από το σύνολο των μορίων DNA, τα οποία πολλαπλασιάστηκαν, ωστόσο, το διάλυμα περιέχει και τα αντιδραστήρια, τα οποία χρησιμοποιήθηκαν κατά τη διαδικασία της ενίσχυσης, όπως το ζεύγος των εκκινητών και τα dNTPs. Η περαιτέρω επεξεργασία ενός PCR προϊόντος απαιτεί τον καθαρισμό του, δηλαδή την αφαίρεση των αντιδραστηρίων της PCR που υπάρχουν στο σωλήνα της αντίδρασης. Υπάρχουν δύο ευρέως χρησιμοποιούμενες μέθοδοι καθαρισμού, οι στήλες καθαρισμού και ο καθαρισμός με σφαιρίδια. Στη παρούσα διπλωματική εργασία, χρησιμοποιήθηκε το σύστημα καθαρισμού με στήλες: NucleoSpin® Gel και PCR Clean-up kit (Macherey-Nagel GmbH & Co. KG, Duren, Germany).

Η μέθοδος καθαρισμού με τη χρήση στηλών βασίζεται στη δέσμευση του DNA πάνω στη στήλη καθαρισμού με τη χρήση του ρυθμιστικού διαλύματος δέσμευσης NTI, καθώς και μορίων τα οποία παρεμποδίζουν τη διαλυτοποίηση του DNA στο νερό. Η απομάκρυνση των αντιδραστηρίων της PCR στηρίζεται στο ρυθμιστικό διάλυμα NT3, το οποίο περιέχει αιθανόλη. Πραγματοποιείται έκπλυση του δείγματος με τη χρήση του NT3, η οποία ακολουθείται από φυγοκέντρηση του δείγματος, ώστε

να απομονωθεί το καθαρό από προσμίξεις DNA στη μεμβράνη της στήλης, ενώ τα υπόλοιπα διαλύματα συλλέγονται σε σωλήνα και απορρίπτονται. Η διαδικασία επαναλαμβάνεται και δεύτερη φορά και, στη συνέχεια, πραγματοποιείται επαναδιαλυτοποίηση του DNA σε κατάλληλο αλκαλικό ρυθμιστικό διάλυμα χαμηλής αλατότητας. Η συγκέντρωση του DNA, καθώς και η καθαρότητά του μπορούν να ελεγχθούν με φωτομέτρηση με τη χρήση του φασματοφωτομέτρου BioSpec-nano Micro-volume UV-Vis Spectrophotometer.

2.9. Αλληλούχηση Επόμενης Γενιάς (Next-generation sequencing, NGS)

Η αντίδραση αλληλούχησης νέας γενιάς, για την μελέτη του εναλλακτικού ματίσματος στο γονίδιο *CDK4,* πραγματοποιήθηκε με τη χρήση της πλατφόρμας αλληλούχησης Personal Genome Machine™ (PGM) της Ion Torrent™. Για την διεξαγωγή του πειράματος απαιτείται η κατασκευή και ποσοτικοποίηση κατάλληλης βιβλιοθήκης, η οποία αποτελεί το υπόστρωμα, για την πραγματοποίηση μιας αντίδρασης αλληλούχησης επόμενης γενιάς. Μετά την κατασκευή της βιβλιοθήκης, η διαδικασία αλληλούχησης με ημιαγωγό χωρίζεται σε τρία βασικά στάδια: α) την ενίσχυση της βιβλιοθήκης μέσω emPCR, β) τον εμπλουτισμό του εκμαγείου με τη χρήση ειδικών σφαιριδίων στρεπταβιδίνης, και γ) την αντίδραση αλληλούχησης σε μία πλατφόρμα Ion Torrent™.

2.9.1. Κατασκευή NGS βιβλιοθήκης

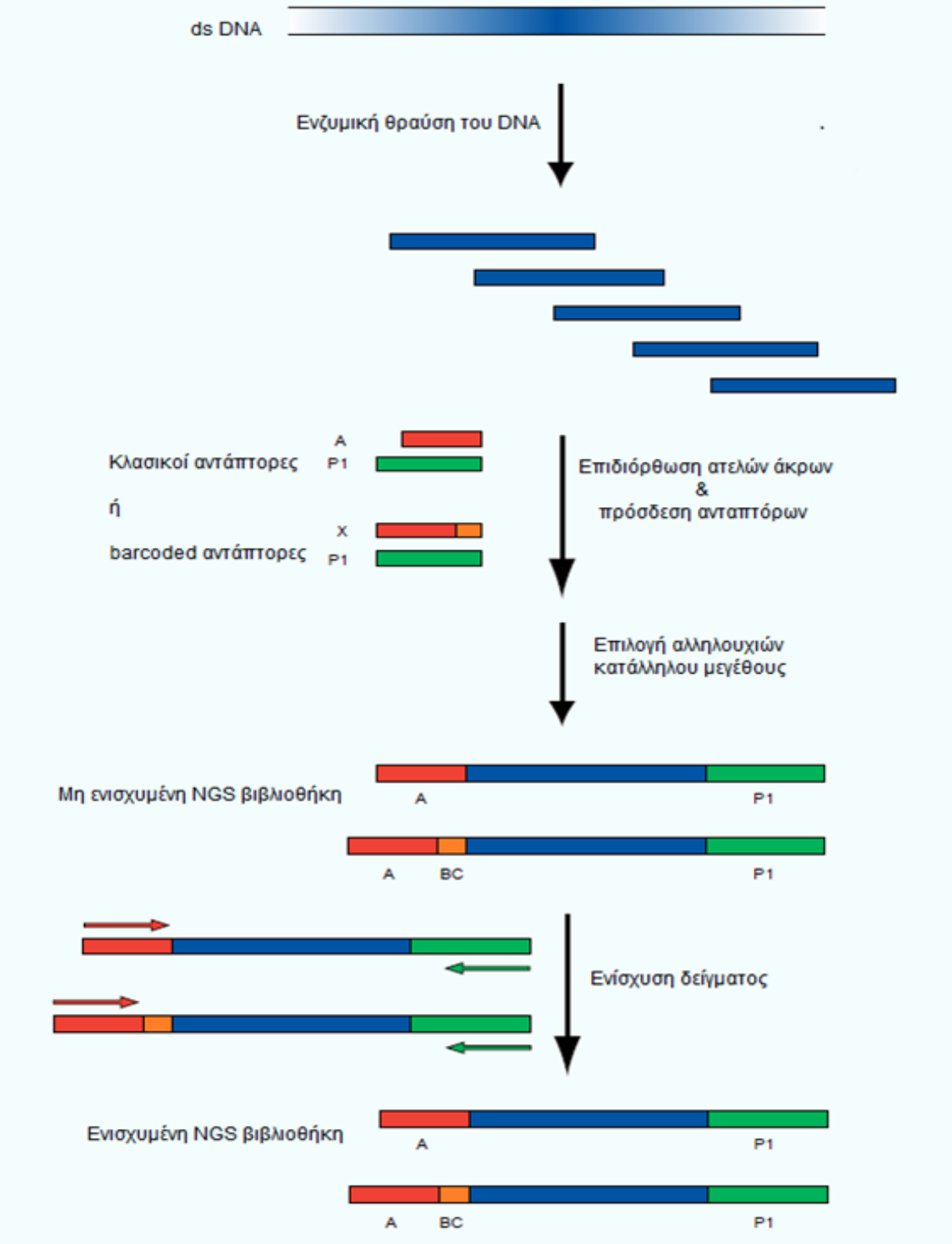
Μια βιβλιοθήκη αλληλούχησης επόμενης γενιάς αποτελείται από το επιθυμητό DNA, το οποίο έχει επεξεργαστεί κατάλληλα, ώστε να είναι αναγνωρίσιμο κατά τη διάρκεια των αντιδράσεων, οι οποίες θα επιτρέψουν τον προσδιορισμό του. Τα στάδια προετοιμασίας για την κατασκευή μιας NGS βιβλιοθήκης διαφέρουν ανάλογα με το είδος του νουκλεϊκού οξέος και την εφαρμογή που χρησιμοποιείται σε κάθε αντίδραση αλληλούχησης. Η μελέτη των εναλλακτικών μεταγράφων του *CDK4* γονιδίου στηρίχθηκε στην αλληλούχηση δίκλωνου μορίου DNA, το οποίο προήλθε από την ενίσχυση των cDNA μορίων, που δημιουργήθηκαν από την αντίδραση της αντίστροφης μεταγραφής μέσω PCR, και επομένως, η προετοιμασία του δείγματος βασίστηκε στην κατασκευή NGS βιβλιοθήκης μορίων DNA.

Αρχικά, το πρώτο βήμα για την κατασκευή της NGS βιβλιοθήκης είναι η διάσπαση της αλληλουχίας του *CDK4* σε μικρότερα τμήματα, το μήκος των οποίων

εξαρτάται από τον χρόνο αντίδρασης. Η θραύση του DNA μπορεί να επιτευχθεί με δύο τρόπους, είτε ενζυματικά είτε με ήχο. Στη παρούσα διπλωματική εργασία, η θραύση του DNA πραγματοποιήθηκε μέσω ενζυμικών αντιδράσεων χρησιμοποιώντας το Ion Xpress™ Plus gDNA Fragment Library Kit, το οποίο περιλαμβάνει τα απαραίτητα αντιδραστήρια, που είναι: α) ρυθμιστικό διάλυμα (Ion Shear™ Plus 10X Reaction Buffer), β) τα ένζυμα διάσπασης του DNA (Ion Shear™ Plus Enzyme Mix II), γ) ρυθμιστικό διάλυμα τερματισμού της αντίδρασης (Ion Shear™ Plus Stop Buffer), και δ) το διαλύτη της αντίδρασης (Low TE, -Tris-HCl + EDTA-). Συγκεκριμένα, ποσότητα 1 μg καθαρού PCR προϊόντος που αποτελεί το DNA της αλληλούχησης τεμαχίστηκε σε τυχαία θραύσματα DNA μήκους 300-400 bp και ακολούθησε καθαρισμός του προϊόντος από το σύνολο των αντιδραστηρίων με τη χρήση σφαιριδίων. Δεύτερο βήμα αποτελεί η επιδιόρθωση των ατελών άκρων των τμημάτων του DNA, ώστε να δημιουργηθούν κατάλληλες θέσεις για την πρόσδεση ανταπτόρων, οι οποίοι απαιτούνται για την πραγματοποίηση της αντίδρασης αλληλούχησης. Η διαδικασία ενσωμάτωσης των ανταπτόρων απαιτεί την παρουσία λιγάσης, η οποία καταλύει την αντίδραση πρόσδεσης στα άκρα κάθε αλυσίδας με την δημιουργία ενός φωσφοδιεστερικού δεσμού. Στη συνέχεια, έγινε καθαρισμός του DNA από το συνολικό μείγμα της αντίδρασης με τη χρήση των ειδικών σφαιριδίων KAPA™ Pure Beads (Kapa Biosystems Inc.).

Η ποιότητα των πειραματικών αλληλουχιών μειώνεται δραματικά στην περίπτωση αλληλούχησης μορίων που έχουν μήκος μεγαλύτερο των 400 bp, καθώς ο αλληλουχητής PGM της Ion Torrent™, που χρησιμοποιήθηκε, έχει την ικανότητα αλληλούχησης τμημάτων DNA μήκους έως και 400 bp. Επομένως, το τρίτο βήμα κατά τη διαδικασία κατασκευής της NGS βιβλιοθήκης είναι η επιλογή των αλυσίδων DNA που έχουν το κατάλληλο μήκος, η οποία μπορεί να γίνει με τους εξής τρεις τρόπους: α) ηλεκτροφορητική διαλογή, β) επιλογή με τη χρήση σφαιριδίων ή γ) αυτοματοποιημένα, με τη χρήση κατάλληλης συσκευής (Pippin Prep instrument). Στη παρούσα διπλωματική εργασία, η επιλογή των θραυσμάτων του DNA έγινε με τη χρήση των σφαιριδίων KAPA™ Pure Beads (Kapa Biosystems Inc.) στα οποία προσδένονται τα, επιθυμητού μήκους, θραύσματα DNA, στο στάδιο καθαρισμού του προϊόντος, το οποίο πραγματοποιήθηκε μετά την διαδικασία πρόσδεσης ανταπτόρων. Στο τέλος της προετοιμασίας της NGS βιβλιοθήκης, πραγματοποιήθηκε ποσοτικοποίηση της βιβλιοθήκης, μέσω PCR σε πραγματικό

χρόνο (real-time PCR, qPCR), με την χρήση των αντιδραστηρίων Ion Library TaqMan™ Quantitation Kit (Ion Torrent™), στο σύστημα ABI 7500 Fast Real-Time PCR (Applied Biosystems™).



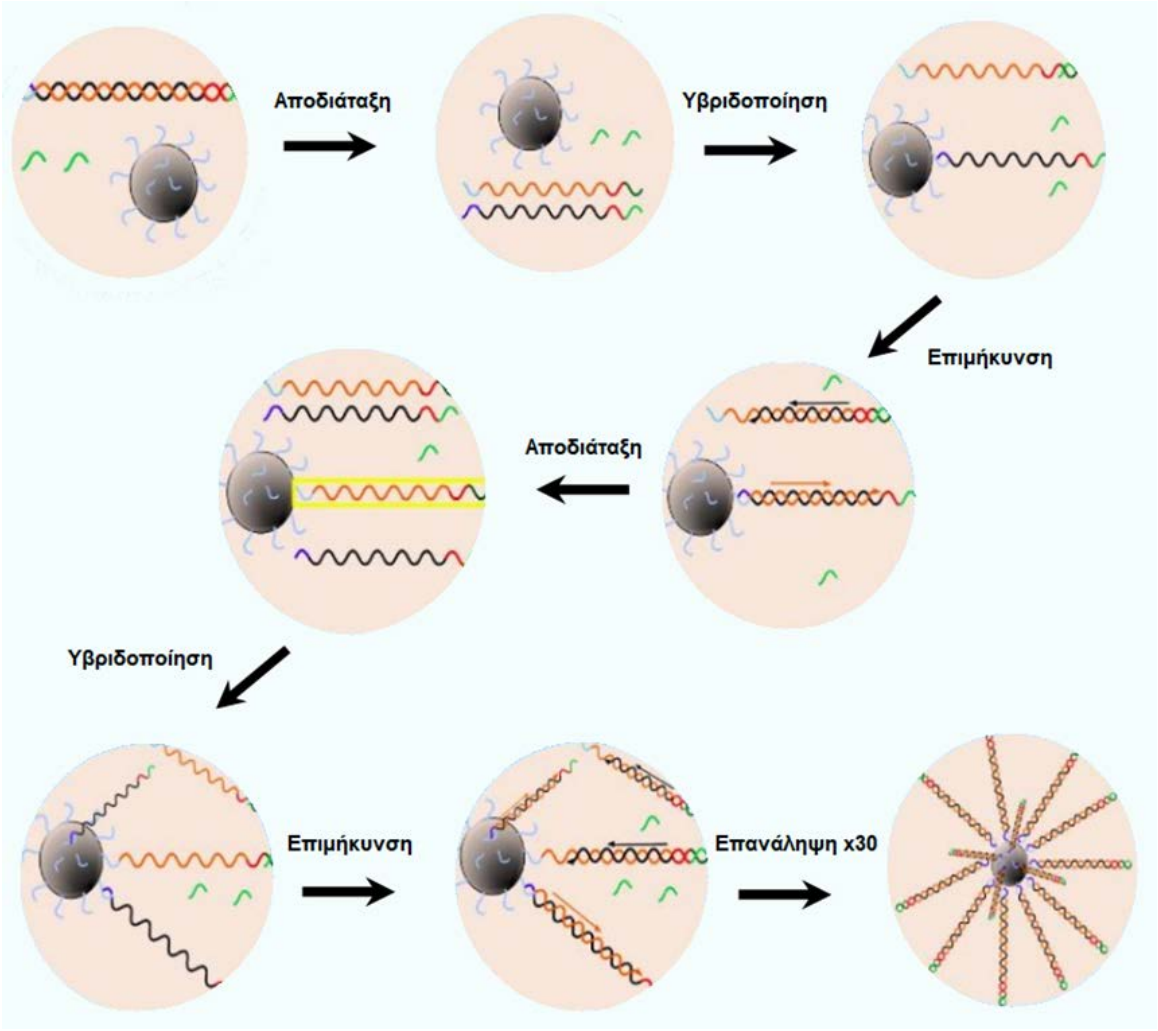
Εικόνα 13. Το διάγραμμα ροής παρουσιάζει τα βασικά βήματα, τα οποία ακολουθήθηκαν κατά τη κατασκευή της NGS βιβλιοθήκης, τα οποία πραγματοποιήθηκαν με τη χρήση του Ion Xpress™ Plus gDNA Fragment Library Kit.

2.9.2. Προετοιμασία και εμπλουτισμός του εκμαγείου

Η προετοιμασία του εκμαγείου της NGS βιβλιοθήκης, το οποίο πρόκειται να αλληλουχηθεί, περιλαμβάνει την ενίσχυση της βιβλιοθήκης, που κατασκευάστηκε μέσω της PCR και τον εμπλουτισμό του δείγματος, ώστε να επιλεχθούν οι κλώνοι, που φέρουν τα ενισχυμένα προϊόντα. Το πρώτο στάδιο της προετοιμασίας του δείγματος, δηλαδή η αντίδραση PCR, πραγματοποιήθηκε στο ειδικό σύστημα OneTouch™ 2 (Ion Torrent™) και χρησιμοποιήθηκε το kit Ion PGM™ Hi-Q™ View OT2 (Ion Torrent™), το οποίο περιέχει τα διαλύματα και τα αντιδραστήρια, που απαιτούνται για την πραγματοποίηση της αντίδρασης αλληλούχησης μέσω της πλατφόρμας Ion Torrent™.

Ακολουθώντας τις οδηγίες του πρωτοκόλλου, η ενίσχυση της NGS βιβλιοθήκης πραγματοποιήθηκε, μέσω της ειδικής PCR αντίδρασης σε γαλάκτωμα. Κατά την αντίδραση πολυμεριμού σε γαλάκτωμα, κάθε θραύσμα της βιβλιοθήκης απομονώνεται σε ελαιώδες μικροπεριβάλλον σε ειδικά σφαιρίδια, η επιφάνεια των οποίων είναι επικαλυμμένη με ολιγονουκλεοτιδικές αλληλουχίες εκκινητών. Κάθε τμήμα DNA της βιβλιοθήκης, που κατασκευάστηκε, δεσμεύεται σε μικροσφαιρίδια (Ion Sphere™ Particles, ISPs), μέσω υβριδοποίησης των ειδικών ανταπτόρων, που είναι προσδεδεμένοι στο DNA με τις συμπληρωματικές αλληλουχίες των εκκινητών, που φέρουν τα μικροσφαιρίδια. Η διαδικασία επώασης των μικροσφαιριδίων ISPs με την NGS βιβλιοθήκη μπορεί να οδηγήσει σε τρεις διαφορετικές περιπτώσεις, με βάση τα γεγονότα σύνδεσης, οι οποίες είναι: α) μονοκλωνικά μικροσφαιρίδια, που έχουν ενσωματώσει ένα θραύσμα DNA, β) πολυκλωνικά μικροσφαιρίδια στην επιφάνεια των οποίων βρίσκονται περισσότερα του ενός θραύσματα DNA, και γ) μικροσφαιρίδια, τα οποία δεν ενσωμάτωσαν καμία αλληλουχία. Τα ISPs, τα οποία θα ενσωματώσουν το DNA και θα αλληλουχηθούν επιτυχώς, είναι μόνο εκείνα τα μικροσφαιρίδια που φέρουν ένα μοναδικό κλώνο DNA. Αντίθετα, τα μικροσφαιρίδια, που έχουν ενσωματώσει περισσότερες από μία αλυσίδες DNA και είναι πολυκλωνικά, αν και θα εισαχθούν στην αντίδραση αλληλούχησης, δεν θα αλληλουχηθούν και, επιπλέον, η αυξημένη παρουσία τους μπορεί να οδηγήσει στη μειωμένη απόδοση της αντίδρασης αλληλούχησης.

Στη συνέχεια, τα μικροσφαιρίδια τοποθετούνται στο γαλάκτωμα της αντίδρασης σχηματίζοντας μικύλλια με την προσθήκη και ανάμειξη καθαρού νερού ελεύθερου από νουκλεάσες και του ελαίου της αντίδρασης (Ion OneTouch™ Reaction Oil). Αποτέλεσμα του σχηματισμού των μικυλλίων είναι η δημιουργία των μικροαντιδραστήρων της αντίδρασης σε καθένα από τους οποίους περιέχεται ένα σφαιρίδιο, που πιθανώς να φέρει ή όχι, έναν ή περισσότερους κλώνους DNA, ένζυμα για την διεξαγωγή της emPCR (Ion OneTouch™ Enzyme Mix), και αντιδραστήρια απαραίτητα για τον πολλαπλασιασμό του DNA (Ion OneTouch™ 2X Reagent Mix). Στο τέλος της αντίδρασης της PCR σε γαλάκτωμα, κάθε σφαιρίδιο φέρει στην επιφάνεια τους χιλιάδες αντίγραφα του κλώνου, που είχε δεσμευτεί εξαρχής, σε αυτό. Στην εικόνα 14 παρουσιάζονται τα τρία βασικά βήματα της αλυσιδωτής αντίδρασης πολυμερισμού, όπως αυτά πραγματοποιούνται, κατά τη διαδικασία της PCR σε γαλάκτωμα, στο ελαιώδες μικροπεριβάλλον που δημιουργείται.



Εικόνα 14. Η PCR σε γαλάκτωμα αποτελείται από τα τρία βασικά στάδια μιας PCR αντίδρασης με τη διαφορά ότι τα μόρια-στόχοι είναι ακινητοποιημένα σε μικροσφαιρίδια και οι αντιδράσεις πραγματοποιούνται σε κατάλληλους μικροαντιδραστήρες.

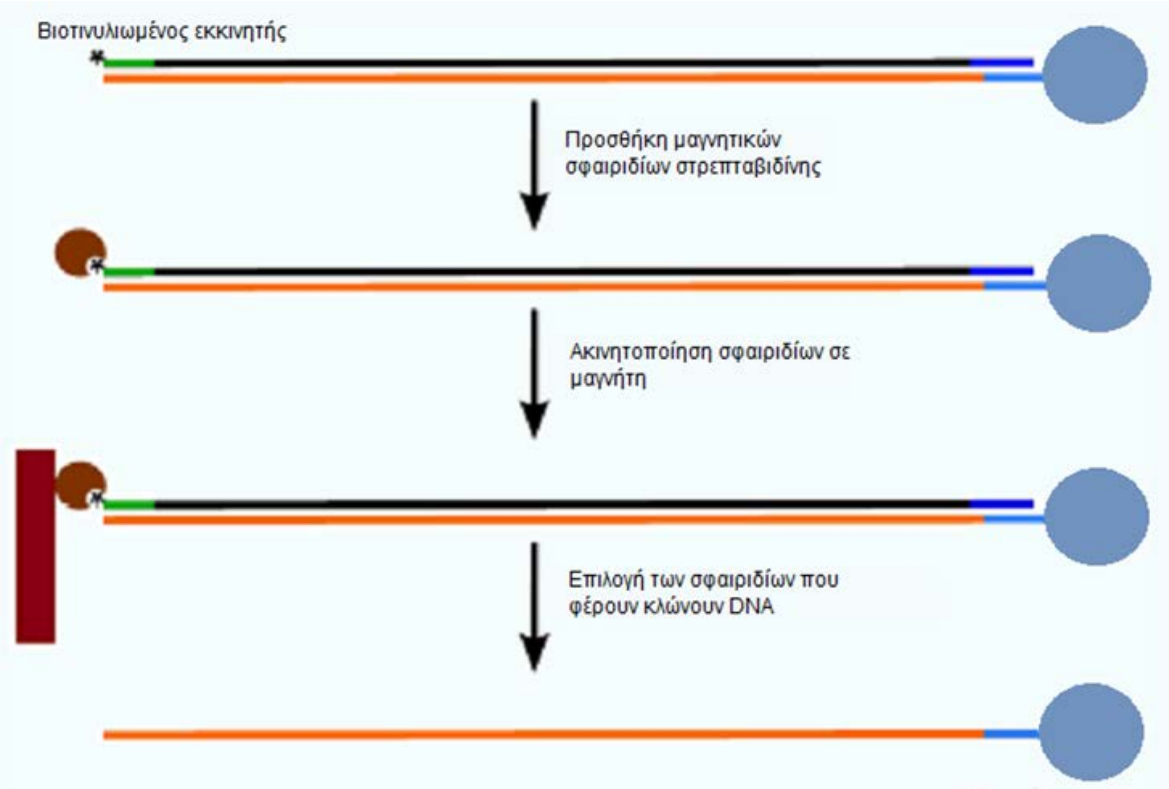
Ο ποιοτικός έλεγχος για την αξιολόγηση της emPCR πραγματοποιήθηκε, μέσω φθορισμομετρίας, σε κατάλληλο φθορισμόμετρο Qubit® 2.0 Fluorometer (Invitrogen™). Η μέθοδος βασίζεται στη μέτρηση του φθορισμού των ειδικών χρωστικών Alexa Fluor® 488 και Alexa Fluor® 647, που περιέχονται στο Ion Sphere™ Quality Control Kit. Η χρωστική Alexa Fluor® 488 προσδένεται στην επιφάνεια των μικροσφαιριδίων, καθώς είναι προσδεδεμένη σε ολιγονουκλεοτίδιο συμπληρωματικό ως προς το ολιγονουκλεοτίδιο της επιφάνειας των σφαιριδίων, ενώ η Alexa Fluor® 647 βρίσκεται συνδεδεμένη σε ολιγονουκλεοτίδιο συμπληρωματικό του αντάπτορα, που φέρουν οι αλυσίδες DNA.

Σύμφωνα με την αρχή μεθόδου, μετριέται η ένταση φθορισμού των δύο χρωστικών και συγκρίνεται ο λόγος τους, που αντιστοιχεί στον λόγο των μικροσφαιριδίων που φέρουν το εκμαγείο προς το σύνολο των σφαιριδίων, και

πρέπει να κυμαίνεται μεταξύ 0,15 - 0,30, ώστε η αντίδραση της emPCR, που προηγήθηκε, να θεωρηθεί επιτυχής. Τιμές μικρότερες από 0,15 υποδηλώνουν ότι μόνο ένα πολύ μικρό πλήθος μικροσφαιριδίων φέρει εκμαγείο, ενώ τιμές μεγαλύτερες του 0,30 υποδηλώνουν την παρουσία αυξημένου αριθμού μικροσφαιριδίων, τα οποία φέρουν πολλαπλές αλληλουχίες DNA. Τα σφαιρίδια, που δεν έχουν ενσωματώσει DNA, θα απομακρυνθούν από το σύνολο των σφαιριδίων, μέσω της διαδικασίας εμπλουτισμού, ενώ τα πολυκλωνικά μικροσφαιρίδια, τα οποία δεν είναι εφικτό να απομακρυνθούν, θα εισαχθούν στην αντίδραση αλληλούχησης, τελικά όμως, θα απορριφθούν ως κακής ποιότητας πειραματικές αλληλουχίες.

Μετά τον ποιοτικό έλεγχο του δείγματος, ακολουθεί το δεύτερο βασικό βήμα της προετοιμασίας του εκμαγείου, που είναι η απομάκρυνση των μικροσφαιριδίων, στα οποία δεν έχει προσδεθεί αλληλουχία DNA. Για το σκοπό αυτό, χρησιμοποιήθηκαν τα διαλύματα έκπλυσης Ion OneTouch™ Wash Solution και MyOne™ Beads Wash Solution, νερό ελεύθερο από νουκλεάσες, και μαγνητικά σφαιρίδια Dynabeads® MyOne™ Streptavidin C1 Beads, τα οποία περιέχονται στο Ion PGM™ Hi-Q™ View OT2 kit (Ion Torrent™). Η διαδικασία εμπλουτισμού πραγματοποιείται σε ξεχωριστή συσκευή εμπλουτισμού, η οποία ονομάζεται Ion OneTouch™ ES.

Η μέθοδος βασίζεται στην δημιουργία γέφυρας μεταξύ των μορίων βιοτίνης-στρεπταβιδίνης και στην χρήση μαγνητικού πεδίου, το οποίο συγκρατεί τα μαγνητικά σφαιρίδια, Dynabeads® MyOne™ Streptavidin C1 Beads, καθώς και τα μόρια που είναι συνδεδεμένα σε αυτά. Πιο αναλυτικά, σε κάθε μικροσφαιρίδιο, που φέρει εκμαγείο ο αντάπτορας, στο ελεύθερο άκρο του DNA, είναι βιοτινυλιωμένος. Πραγματοποιείται ανάμειξη των μικροσφαιριδίων με τα μαγνητικά σφαιρίδια, τα οποία έχουν στην επιφάνεια τους μόρια στρεπταβιδίνης, με αποτέλεσμα την αλληλεπίδραση μεταξύ των μορίων βιοτίνης- στρεπταβιδίνης, όπως παρουσιάζεται στην εικόνα 15. Τα μικροσφαιρίδια, που δεν έχουν ενσωματώσει στην επιφάνεια τους DNA αλληλουχίες, δεν συμπλοκοποιούνται μέσω της βιοτίνης στα μόρια στρεπταβιδίνης και απομακρύνονται από το διάλυμα της αντίδρασης. Στη συνέχεια, ακολουθεί ο τερματισμός της αλληλεπίδρασης μεταξύ των δύο μορίων και τα κλωνικά μικροσφαιρίδια επαναδιαλυτοποιούνται, ενώ τα μαγνητικά σφαιρίδια στρεπταβιδίνης απομακρύνονται με τη χρήση μαγνήτη.



Εικόνα 15. Σφαιρίδια στρεπταβιδίνης συμπλοκοποιούνται με τους βιοτινυλιωμένους εκκινητές, οι οποίοι είναι υβριδοποιημένοι στο ακινητοποιημένο DNA. Τελικό αποτέλεσμα είναι ο διαχωρισμός των μικροσφαιριδίων που φέρουν τους κλώνους του DNA από τα σφαιρίδια που δεν έχουν ενσωματώσει DNA.

2.9.3. Αντίδραση αλληλούχησης επόμενης γενιάς

Οι αντιδράσεις του τελευταίου σταδίου, για τον προσδιορισμό του μεταγράφου του *CDK4,* μέσω αλληλούχησης επόμενης γενιάς, πραγματοποιήθηκαν στο μηχάνημα αλληλούχησης Ion Torrent Personal Genome Machine™. Η διαδικασία περιλαμβάνει την προετοιμασία του αλληλουχητή, το φόρτωμα του δείγματος στον ημιαγωγό και την αντίδραση αλληλούχησης. Αρχικά, προσαρτώνται στον αλληλουχητή τα αντιδραστήρια που απαιτούνται για την διαδικασία της αλληλούχησης, τα διαλύματα τριφωσφορικών νουκλεοτιδίων και τα κατάλληλα ρυθμιστικά διαλύματα (Wash Buffers), τα οποία περιέχονται επίσης στο Ion PGM™ Hi-Q™ View Sequencing kit (Ion Torrent™). Το σύστημα έχει αφενός τη δυνατότητα ελέγχου των αντιδραστηρίων, που προστέθηκαν, και αφετέρου ειδοποιεί τον χειριστή για την ύπαρξη πιθανών σφαλμάτων. Το δείγμα DNA, το οποίο βρίσκεται στα ειδικά ISPs, εισάγεται στον ημιαγωγό της αντίδρασης και, τελικά, σε κάθε πηγάδι του ημιαγωγού ενσωματώνεται ένα μόνο μικροσφαιρίδιο. Στη συγκεκριμένη

αντίδραση αλληλούχησης χρησιμοποιήθηκε ο ημιαγωγός Ion 316™ Chip. Το σύστημα ειδοποιεί τον χειριστή, κατά την έναρξη του πειράματος, για την επιτυχία του φορτώματος του δείγματος στα πηγάδια (wells) του ημιαγωγού, υπολογίζοντας το ποσοστό των θέσεων που καλύφθηκαν από το δείγμα, καθώς είναι πιθανόν να μην καλυφθούν όλες οι θέσεις του chip με μικροσφαιρίδια. Η αντίδραση αλληλούχησης, μέσω της πλατφόρμας Ion Torrent™, βασίζεται στην μεθοδολογία αλληλούχησης μέσω σύνθεσης, η οποία περιγράφηκε στο κεφάλαιο της εισαγωγής. Η διαδικασία τερματίζεται όταν αλληλουχηθούν όλοι οι κλώνοι του DNA, που φορτώθηκαν στα πηγάδια του ημιαγωγού. Ακολουθεί η βιοπληροφορική ανάλυση των δεδομένων αλληλούχησης.

2.9.4. Βιοπληροφορική ανάλυση των NGS δεδομένων

Τα δεδομένα αλληλούχησης επόμενης γενιάς αποθηκεύονται σε συγκεκριμένο τύπο αρχείου .fastq. Το αρχείο FASTQ ανακτήθηκε από τον Server, μέσω του ειδικού περιηγητή Torrent Browser, ο οποίος επέτρεψε την λήψη του αρχείου σε Η/Υ. Τα αρχεία FASTQ, τα οποία παράγονται στο τέλος μιας NGS αντίδρασης, περιέχουν το σύνολο των αλληλουχούμενων μορίων- πειραματικών αλληλουχιών, το οποίο αποτελεί τα δεδομένα ενός πειράματος αλληλούχησης. Τα αρχεία FASTQ έχουν συγκεκριμένη δομή, η οποία αποτελείται από ένα μοτίβο τεσσάρων σειρών, οι οποίες αντιστοιχούν σε μια πειραματική αλληλουχία και, επομένως ο αριθμός των επαναλήψεων του μοτίβου αυτού σε ένα αρχείο FASTQ αντιστοιχεί στον αριθμό των αλυσίδων DNA, των οποίων προσδιορίστηκε η αλληλουχία τους. Κάθε μία από τις τέσσερεις γραμμές ενός μοτίβου δίνει τα δεδομένα για την κάθε πειραματική αλληλουχία (εικόνα 16) και, επομένως, κάθε σειρά αντιστοιχεί σε διαφορετική πληροφορία και είναι η εξής:

Η **πρώτη γραμμή** ξεκινάει πάντοτε με το σύμβολο «@» και περιέχει έναν μοναδικό αναγνωριστικό κωδικό της πειραματικής αλληλουχίας, ο οποίος αποτελείται από τον κωδικό του πειράματος και τις συντεταγμένες του πηγαδιού του ημιαγωγού, στο οποίο βρισκόταν η συγκεκριμένη αλληλουχία.

Η **δεύτερη γραμμή** αντιπροσωπεύει την νουκλεοτιδική αλληλουχία, η οποία προσδιορίστηκε στη συγκεκριμένη θέση του ημιαγωγού.

Η **τρίτη γραμμή** περιλαμβάνει το σύμβολο «+».

Η **τέταρτη γραμμή** περιλαμβάνει πάντοτε ίσο αριθμό χαρακτήρων με τους αντίστοιχους χαρακτήρες της δεύτερης γραμμής και αποτελείται από κωδικοποιημένα σύμβολα, τα οποία είναι: *!"#$%&'()*+,./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ\^_`abc defghijklmnopqrstuvwxyz*|~. Κάθε ένα από τους συγκεκριμένους χαρακτήρες αντιστοιχεί στο quality score της αντίστοιχης βάσης που βρίσκεται στη γραμμή 2, και, επομένως, κάθε σύμβολο αξιολογεί την αξιοπιστία με την οποία ο αλληλουχητής έχει προσθέσει το συγκεκριμένο νουκλεοτίδιο στην αντίστοιχη θέση.

Εικόνα 16. Σύμφωνα με τη δομή ενός FASTQ αρχείου τα δεδομένα μιας πειραματική αλληλουχίας παρουσιάζονται σε 4 γραμμές. Οι χαρακτήρες της δεύτερης σειράς φανερώνουν την αλληλουχία του τμήματος, όπως προσδιορίστηκε, ενώ τα σύμβολα της τελευταίας σειράς παρέχουν δεδομένα για τον έλεγχο ποιότητας της αλληλουχίας.

Ο έλεγχος ποιότητας των πρωτογενών δεδομένων, που συλλέχθηκαν από τις πειραματικές αλληλουχίες στο αρχείο FASTQ, πραγματοποιήθηκε με το εργαλείο FastQC. Το εργαλείο FastQC μπορεί να εγκατασταθεί και να χρησιμοποιηθεί ως εφαρμογή στον Η/Υ και επιτρέπει τον υπολογισμό μιας σειράς στατιστικών αναλύσεων, που αφορούν στην ποιότητα των πρωτογενών δεδομένων του πειράματος. Ο αλγόριθμος παράγει 7 βασικά γραφήματα ανάμεσα στα οποία είναι το γράφημα, το οποίο παρουσιάζει τη ποιότητα ανάγνωσης ανά νουκλεοτιδική βάση.

Στη συνέχεια, οι πειραματικές αλληλουχίες, που λήφθηκαν για το *CDK4*, στοιχίζονται στο γονιδίωμα αναφοράς (GRCh38) χρησιμοποιώντας τον αλγόριθμο HISAT2. Ο αλγόριθμος HISAT2 αποτελεί ένα γρήγορο και ακριβές πρόγραμμα στοίχισης και χαρτογράφησης των πειραματικών αλληλουχιών, που λαμβάνονται από πειράματα αλληλούχησης επόμενης γενιάς, και, επιπλέον, δίνει αξιόπιστα αποτελέσματα για ενδείξεις γεγονότων εναλλακτικού ματίσματος με βάση ένα γονιδίωμα αναφοράς [118]. Η ανάλυση των δεδομένων με την χρήση του

αλγόριθμου HISAT2 απαιτεί δύο αρχεία εισόδου, τα οποία είναι το αρχείο με τα πρωτογενή δεδομένα του πειράματος αλληλούχησης και ένα αρχείο, που περιέχει το γονιδίωμα αναφοράς πάνω στο οποίο θα στοιχηθούν οι πειραματικές αλληλουχίες. Τα αποτελέσματα στοίχισης λαμβάνονται από τον αλγόριθμο σε νέο SAM αρχείο εξόδου, το οποίο περιλαμβάνει το σύνολο των πειραματικών αλληλουχιών, οι οποίες έχουν στοιχηθεί στο γονιδίωμα αναφοράς, και απορρίπτει τις αλληλουχίες, οι οποίες δεν κατάφεραν να στοιχηθούν. Προκειμένου να οπτικοποιηθούν τα αποτελέσματα της στοίχισης, απαιτείται η μετατροπή του αρχείου SAM σε δύο τύπους αρχείων: ένα αρχείο BAM, το οποίο περιέχει εκείνες τις πειραματικές αλληλουχίες, η στοίχιση των οποίων έγινε με επιτυχία στο γονιδίωμα αναφοράς και ένα αρχείο BED, το οποίο περιέχει το σύνολο των γεγονότων συρραφής μεταξύ των εξωνίων, που εντοπίστηκαν κατά τη στοίχιση. Η μετατροπή του αρχείου SAM σε BAM πραγματοποιήθηκε με την χρήση του προγράμματος samtools και η οπτικοποίησή του με την χρήση του προγράμματος Integrative Genomics Viewer (IGV).

Επιπλέον, εκτός από την ανάλυση των αποτελεσμάτων με την χρήση του αλγόριθμου HISAT2 και το πρόγραμμα οπτικοποίησης, IGV, η διερεύνηση των γεγονότων εναλλακτικού ματίσματος του γονιδίου *CDK4* πραγματοποιήθηκε με τον αλγόριθμο “ASDT”, ο οποίος δημιουργήθηκε από μέλη του εργαστηρίου [119]. Ο αλγόριθμος ASDT χρησιμοποιεί ως αρχείο εισόδου τα πρωτογενή δεδομένα του FASTQ και ανιχνεύει τις θέσεις συρραφής μεταξύ των εξωνίων δίνοντας αποτελέσματα για νέα εναλλακτικά γεγονότα ματίσματος, όπως νέες θέσεις συρραφής, διατήρηση εσωνίων ή προεκτάσεις εξωνίων, και παράγει αντίστοιχα .txt αρχεία εξόδου.

2.10. Αλληλούχηση Τρίτης Γενιάς (Third-generation sequencing, TGS)

Η αντίδραση Αλληλούχησης Τρίτης Γενιάς για την μελέτη γεγονότων εναλλακτικού ματίσματος του γονίδιο *CDK4* πραγματοποιήθηκε με τη χρήση του συστήματος MinION® Mk1C, της πλατφόρμας αλληλούχησης ONT, χρησιμοποιώντας το flow cell FLO-MIN106 χημείας R9.4.1. Η διαδικασία αλληλούχησης τρίτης γενιάς με τη χρήση βιολογικών νανοπόρων βασίζεται σε συγκεκριμένη μεθοδολογία δύο σταδίων, τα οποία είναι η προετοιμασία της TGS βιβλιοθήκης, δηλαδή του δείγματος προς αλληλούχηση, και η αντίδραση αλληλούχησης.

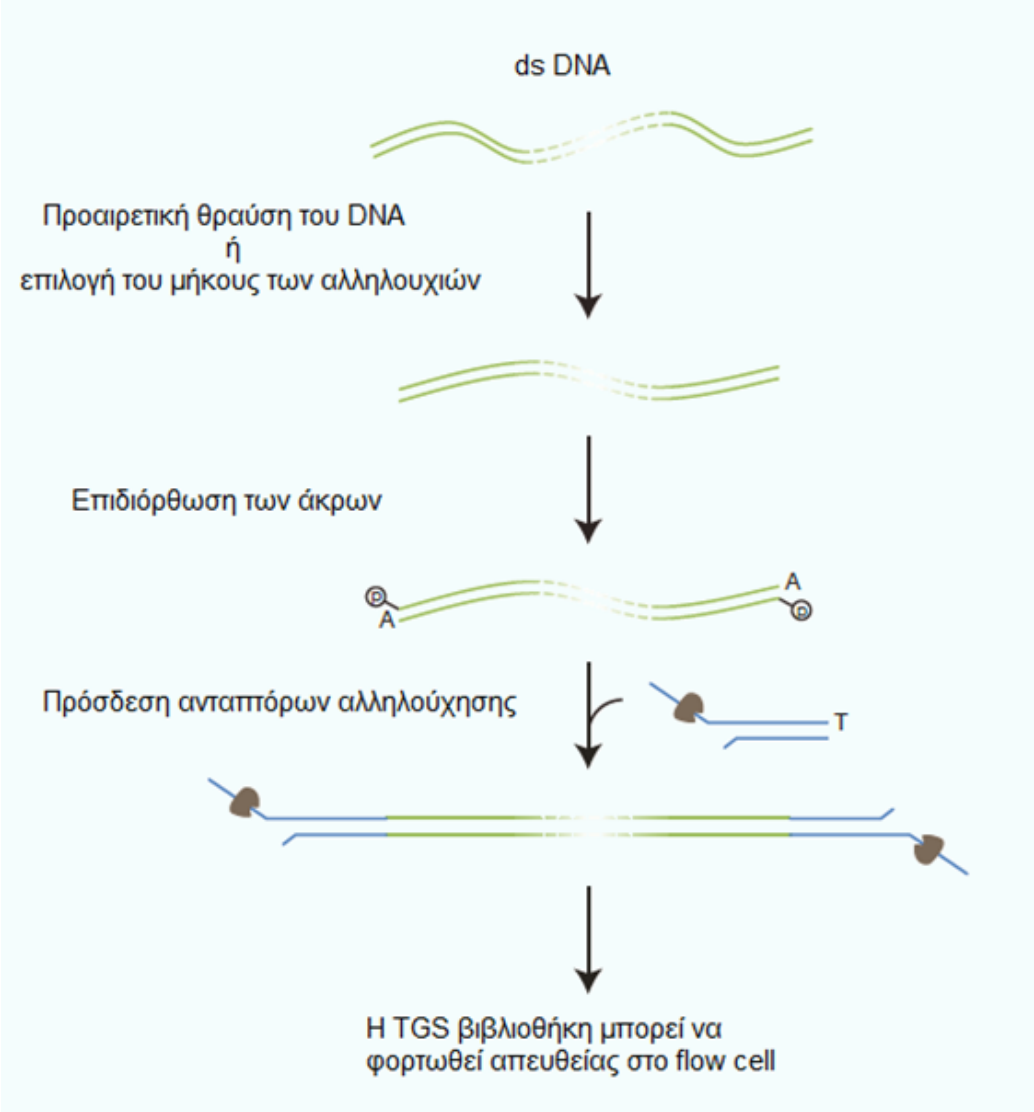
2.10.1. Κατασκευή TGS βιβλιοθήκης

Για την κατασκευή της TGS βιβλιοθήκης χρησιμοποιήθηκε το kit αλληλούχησης Ligation Sequencing Kit (SQK-LSK109, ONT), σύμφωνα με τις οδηγίες του αντίστοιχου πρωτοκόλλου. Πιο αναλυτικά, σε σωλήνα Eppendorf 1,5 mL, προστέθηκε αρχική ποσότητα 1 μg δείγματος του ενισχυμένου PCR προϊόντος του γονιδίου *CDK4,* το οποίο αραιώθηκε σε 49 μL νερό ελεύθερο από νουκλεάσες. Επόμενο βήμα είναι η ενζυμική επεξεργασία του DNA, ώστε να δημιουργηθούν κατάλληλα άκρα για την πρόσδεση ανταπτόρων με την χρήση του kit NEBNext® Ultra™ II End Repair / dA-Tailing Module (New England Biolabs, Inc). Σε σωλήνα 0,2 mL προστέθηκαν: 1 μl DNA CS, το οποίο χρησιμοποιείται για τον ποιοτικό έλεγχο της αντίδρασης, 47 μl του δείγματος DNA, τα ρυθμιστικά διαλύματα NEBNext FFPE DNA Repair Buffer και Ultra II End-prep reaction buffer σε ίσο όγκο 3,5 μL, 2 μl NEBNext FFPE DNA Repair Mix, το οποίο περιέχει ένζυμα για την επιδιόρθωση των άκρων του δείγματος, και 3 μl Ultra II End-prep enzyme, mix που αποτελείται από τα ένζυμα τερματισμού της αντίδρασης. Πραγματοποιήθηκε επώαση στο θερμικό κυκλοποιητή για 5 λεπτά στους 20º C και μετά για 5 λεπτά στους 65º C.

Στη συνέχεια, πραγματοποιήθηκε καθαρισμός του δείγματος, από τα αντιδραστήρια, με την χρήση των σφαιριδίων καθαρισμού AMPure XP beads (Beckman Coulter, USA). Σύμφωνα με τις οδηγίες του πρωτοκόλλου, το δείγμα DNA μεταφέρθηκε σε σωλήνα Eppendorf 1,5 mL, στον οποίο προστέθηκαν 60 μL AMPure XP® σφαιριδίων, ακολούθησε επώαση, με συνεχή ανάδευση για 5 λεπτά σε θερμοκρασία δωματίου. Το δείγμα τοποθετήθηκε σε μαγνήτη και με την χρήση πιπέτας απορρίφθηκε το διάλυμα, το οποίο περιέχει τα αντιδραστήρια του προηγούμενου δείγματος, ενώ το DNA βρίσκεται προσδεδεμένο στα μαγνητικά σφαιρίδια. Κατόπιν, προστέθηκαν 200 μL 70% αιθανόλης και, στη συνέχεια, το υγρό απορρίφθηκε με τη χρήση πιπέτας. Η διαδικασία έκπλυσης με αιθανόλη πραγματοποιήθηκε και δεύτερη φορά. Το δείγμα απομακρύνθηκε από τον μαγνήτη και το DNA επαναδιαλυτοποιήθηκε σε 61 μL νερού ελεύθερο από νουκλεάσες. Ακολούθησε επώαση 2 λεπτών σε θερμοκρασία δωματίου και, στη συνέχεια, ο σωλήνας τοποθετήθηκε και πάλι σε μαγνήτη. Τέλος, συλλέχθηκε το DNA με τη χρήση πιπέτας σε σωλήνα Eppendorf 1,5 mL. Ο ποσοτικός προσδιορισμός του

δείγματος έγινε με τη χρήση του φθορισμόμετρου Qubit® 2.0 Fluorometer (Invitrogen™).

Επόμενο βήμα είναι η πρόσδεση ανταπτόρων στα άκρα του DNA, η οποία πραγματοποιήθηκε σε σωλήνα Eppendorf 1,5 mL, στον οποίο προστέθηκαν 60 mL DNA, 25 μL ρυθμιστικού διαλύματος πρόσδεσης, Ligation Buffer (LNB), 10 μL ενζύμου λιγάσης, NEBNext Quick T4 Ligase (New England Biolabs, Inc), και 5 μL διαλύματος Adapter Mix (AMX), το οποίο περιέχει τους αντάπτορες. Η επώαση της αντίδρασης διήρκησε 10 λεπτά σε θερμοκρασία δωματίου και πραγματοποιήθηκε επόμενος καθαρισμός του δείγματος με την χρήση των σφαιριδίων AMPure XP®. Στο διάλυμα προστέθηκα 40 μL σφαιριδίων, το οποίο επωάστηκε με συνεχή ανάδευση για 5 λεπτά σε θερμοκρασία δωματίου. Το δείγμα τοποθετήθηκε σε μαγνήτη και απορρίφθηκε το υγρό, ενώ η έκπλυση των σφαιριδίων έγινε με τη χρήση 250 μl του ρυθμιστικού διαλύματος Short Fragment Buffer (SFB), καθώς το DNA έχει μήκος <3 kb. Απορρίφθηκε, ξανά, το υγρό και ακολούθησε και δεύτερη έκπλυση του δείγματος. Το DNA επαναδιαλυτοποιήθηκε σε 15 μL ρυθμιστικού διαλύματος, Elution Buffer (EB) και, ύστερα από επώαση 10 λεπτών σε θερμοκρασία δωματίου, συλλέχθηκαν 15 μL καθαρού DNA. Τέλος, μετρήθηκε η ποσότητα του καθαρού DNA με τη χρήση του φθορισμόμετρου Qubit® 2.0 Fluorometer (Invitrogen™).



Εικόνα 17. Στο διάγραμμα ροής παρουσιάζονται τα βασικά στάδια προετοιμασίας του DNA, τα οποία ακολουθήθηκαν για τη διεξαγωγή της TGS αλληλούχησης. Κατά την αλληλούχηση τρίτης γενιάς δεν απαιτείται η θραύση του DNA και η ενίσχυσή του κατά την προετοιμασία της TGS βιβλιοθήκης.

2.10.2. Αντίδραση αλληλούχησης τρίτης γενιάς

Η αλληλούχηση των μεταγράφων του *CDK4*, μέσω των πρωτεϊνικών νανοπόρων, πραγματοποιήθηκε με τη χρήση του, τρίτης γενιάς, αλληλουχητή MinION™ Mk1C της ONT. Η διαδικασία περιλαμβάνει την προετοιμασία του flow cell, στο οποίο θα γίνει η αλληλούχηση, το φόρτωμα του δείγματος στο flow cell και την αντίδραση αλληλούχησης με ανίχνευση του ηλεκτρικού ρεύματος στη μεμβράνη των νανοπόρων.

Το στάδιο προετοιμασίας του flow cell, στο οποίο θα φορτωθεί η TGS βιβλιοθήκη που κατασκευάστηκε, απαιτεί τη προσθήκη ρυθμιστικών διαλυμάτων στη συσκευή, στο σημείο που φέρει τους νανοπόρους- κανάλια-, ώστε να διαμορφωθούν οι κατάλληλες συνθήκες pH για την αντίδραση αλληλούχησης. 30 μL του διαλύματος Flush Tether (FLT) αναμειγνύονται με το διάλυμα Flush Buffer (FB), τα οποία περιέχονται στο Flow cell Priming Kit (EXP-FLP002) και 800 μL του τελικού διαλύματος φορτώνεται σε κατάλληλη θύρα (priming port) στο flow cell της αντίδρασης. Μετά το πέρας 5 λεπτών, φορτώνονται εκ νέου 200 μL του ίδιου διαλύματος στην ίδια θύρα του flow cell. Στη συνέχεια, 12 μL της DNA βιβλιοθήκης αναμειγνύονται με 37,5 μL ρυθμιστικού διαλύματος αλληλούχησης (Sequencing Buffer, SQB) και 25,5μL σφαιριδίων φόρτωσης, Loading Beads (LB). Τελικά, στη συσκευή αλληλούχησης φορτώνονται 75 μL μέσω την θύρας φόρτωσης του δείγματος. Απαιτείται ιδιαίτερη προσοχή κατά τη διαδικασία της φόρτωσης του δείγματος, καθώς η παρουσία φυσαλίδων, δηλαδή αέρα, στην επιφάνεια του flow cell θα οδηγήσει σε μειωμένη απόδοση της αντίδρασης. Στο συγκεκριμένο πείραμα, ο χρόνος αλληλούχησης της DNA βιβλιοθήκης, που κατασκευάστηκε, ήταν 3 ώρες και οδήγησε στον προσδιορισμό 1,78 εκατομμυρίων πειραματικών αλληλουχιών μέσου μήκους 0,8 kb.

2.10.3. Βιοπληροφορική ανάλυση TGS δεδομένων

Ο αλγόριθμος Guppy, με την χρήση του οποίου έγινε η αρχική επεξεργασία των αποτελεσμάτων του πειράματος αλληλούχησης τρίτης γενιάς, είναι ένα εργαλείο επεξεργασίας πρωτογενών δεδομένων, το οποίο παρέχει τη δυνατότητα βιοπληροφορικής επεξεργασίας, όπως η υπολογιστική διαδικασία της αντιστοίχισης του ηλεκτρικού σήματος σε νουκλεοτιδική αλληλουχία (basecalling), η αφαίρεση των αλληλουχιών, που αποτελούν τους αντάπτορες, και η αξιολόγηση της ποιότητας των παραγόμενων πειραματικών δεδομένων [120]. Οι πειραματικές αλληλουχίες, οι οποίες προκύπτουν ύστερα από τη διαδικασία του basecalling, αξιολογούνται με βάση τη βαθμολογία της ποιότητάς τους και διαχωρίζονται σε δύο ξεχωριστούς φακέλους, από τους οποίους ο ένας περιέχει τις πειραματικές αλληλουχίες, οι οποίες αξιολογήθηκαν ως υψηλής ποιότητας (pass) και ο δεύτερος περιέχει τις νουκλεοτιδικές αλληλουχίες των οποίων ο έλεγχος ποιότητας υπέδειξε χαμηλή αξιοπιστία για τα δεδομένα αλληλούχησής τους.

Κατά τη βιοπληροφορική ανάλυση, για την μελέτη του εναλλακτικού ματίσματος στο γονίδιο *CDK4,* χρησιμοποιήθηκαν μόνο οι πειραματικές αλληλουχίες, οι οποίες χαρακτηρίστηκαν από υψηλή αξιοπιστία. Επόμενο βήμα ήταν η στοίχιση των πρωτογενών δεδομένων για το γονίδιο *CDK4* του FASTQ αρχείου στο ανθρώπινο γονιδίωμα αναφοράς (GRCh38). Η στοίχιση πραγματοποιήθηκε με τη χρήση του αλγόριθμου Minimap2 [121], που αποτελεί ένα γρήγορο πρόγραμμα στοίχισης και ευθυγράμμιση νουκλεοτιδικών αλληλουχιών, το οποίο μπορεί, επιπλέον, να αναγνωρίσει αλληλοεπικαλυπτόμενα άκρα σε μεγάλου μήκους πειραματικές αλληλουχίες, να εντοπίσει γεγονότα συρραφής μεταξύ των περιοχών του γονιδιώματος και να συναρμολογήσει τις αλληλουχίες αυτές με βάση την ακολουθία στο γονιδίωμα αναφοράς. Ο αλγόριθμος απαιτεί FASTQ αρχεία εισόδου και παράγει δύο τύπους αρχείων εξόδου, PAF ή SAM. Στη συνέχεια χρησιμοποιήθηκε το πρόγραμμα samtools για την μετατροπή του αρχείου SAM σε BAM, ώστε να είναι δυνατή η οπτικοποίηση των αποτελεσμάτων με τη χρήση του προγράμματος οπτικοποίησης Integrative Genomics Viewer (IGV) για την ανίχνευση των γεγονότων εναλλακτικού ματίσματος. Επιπλέον, πραγματοποιήθηκε βιοπληροφορική επεξεργασία των δεδομένων, που περιέχονται στο FASTQ αρχείο, με τη χρήση του αλγόριθμου ASDT για την ανίχνευση νέων εναλλακτικών θέσεων συρραφής μεταξύ των εξωνίων.

2.11. Μελέτη του προφίλ έκφρασης των εναλλακτικών μεταγράφων του *CDK4* με ποσοτική real-time PCR (qPCR)

Η μελέτη του προφίλ έκφρασης των νέων εναλλακτικών μεταγράφων του γονιδίου *CDK4* πραγματοποιήθηκε με την μέθοδο της αλυσιδωτής αντίδρασης πολυμεράσης σε πραγματικό χρόνο (real-time PCR, RT-PCR), με σκοπό τον ποσοτικό προσδιορισμό κάθε μεταγράφου σε διαφορετικούς τύπους ιστών.

2.11.1. Αρχή της μεθόδου

Η αλυσιδωτή αντίδραση πολυμεράσης πραγματικού χρόνου, real-time PCR, απαιτεί την παρουσία ενός δίκλωνου DNA, το οποίο λειτουργεί ως υπόστρωμα για τον ενζυμικό πολλαπλασιασμό ενός μικρότερου τμήματος DNA (50-400 bp), με ταυτόχρονο προσδιορισμό της ποσότητας του παραγόμενου προϊόντος. Για το σκοπό αυτό, η μέθοδος της real-time PCR χρησιμοποιεί ένα σετ εκκινητών

(primers), καθένας από τους οποίους προσδένεται εκλεκτικά στην μία εκ των δύο αλυσίδων του DNA με βάση τον κανόνα της συμπληρωματικότητας των βάσεων. Η ενίσχυση του DNA πραγματοποιείται με την χρήση του ενζύμου DNA πολυμεράση, η οποία επιμηκύνει την αλληλουχία των εκκινητών με κατεύθυνση 5'-3' χρησιμοποιώντας ως εκμαγείο την μονόκλωνη αλυσίδα του DNA στόχου. Η αντίδραση πραγματοποιείται σε 3 βασικά στάδια, όπως και στην απλή PCR, τα οποία αποτελούν ένα κύκλο και είναι η αποδιάταξη του δίκλωνου DNA, η υβριδοποίηση των εκκινητών στο τμήμα του DNA που είναι συμπληρωματικό και η επιμήκυνση της νεοσυντιθέμενης αλυσίδας.

Η ποσοτική real-time PCR επιτρέπει την ποσοτικοποίηση του DNA-στόχου που ενισχύεται σε πραγματικό χρόνο καθόλη τη διάρκεια της αντίδρασης ενίσχυσης. Η μέθοδος βασίζεται στην παρακολούθηση του ρυθμού αύξησης του φθορισμού κάποιας φθορίζουσας χρωστικής. Πιο συγκεκριμένα, σε κάθε κύκλο της PCR, καταγράφεται ο φθορισμός, που παράγεται κατά την ενίσχυση του DNA, και τελικά προκύπτει καμπύλη ενίσχυσης (amplification plot), η οποία παρουσιάζει τις μεταβολές του φθορισμού. Η καμπύλη ενίσχυσης αποτελείται από τρεις φάσεις: την εκθετική, τη γραμμική και τη φάση κορεσμού. Στην εκθετική φάση όλα τα αντιδραστήρια βρίσκονται σε περίσσεια και επομένως, σε κάθε κύκλο της αντίδρασης πραγματοποιείται ακριβής διπλασιασμός των αντιγράφων του DNA-στόχου. Κατά τη γραμμική φάση, αρχίζει η εξάντληση των αντιδραστηρίων και η επιβράδυνση της αντίδρασης, καθώς μειώνεται η αποδοτικότητά της, και τελικά φτάνει σε σημείο κορεμού (plateau).

Ο ποσοτικός προσδιορισμός του PCR προϊόντος επιτυγχάνεται κατά την εκθετική φάση της αντίδρασης, με τον προσδιορισμό της τιμής Ct (threshold cycle). Η τιμή Ct είναι αντιστρόφως ανάλογη της ποσότητας του υποστρώματος που χρησιμοποιήθηκε αρχικά, δηλαδή, όσο μικρότερη είναι η τιμή Ct τόσο υψηλότερη είναι η αρχική ποσότητα του DNA-στόχου, και αντίστροφα.

Η real-time PCR επιτρέπει, εκτός από τον ποσοτικό προσδιορισμό του δείγματος και τον ποιοτικό προσδιορισμό του προϊόντος μέσω της καμπύλης τήξης (melt curve) που κατασκευάζεται μετά το πέρας της αντίδρασης. Κάθε ενισχυμένος στόχος DNA χαρακτηρίζεται από ένα μοναδικό σημείο τήξης (temperature melting, T_m) και επομένως η παρουσία περισσότερων της μίας κορυφής στην καμπύλη τήξης ενός DNA-στόχου υποδηλώνει την παρουσία παραπροϊόντων.

Τα ειδικά και τα μη ειδικά συστήματα ανίχνευσης είναι τα δύο είδη χημείας που χρησιμοποιούνται για τον προσδιορισμό του PCR προϊόντος που παράγεται κατά την real-time PCR. Στα μη ειδικά συστήματα χρησιμοποιούνται φθορίζουσες χρωστικές οι οποίες ενσωματώνονται στο δίκλωνο DNA. Η SYBR-Green αποτελεί μία ευρέως διαδεδομένη φθορίζουσα χρωστική, η οποία, κατά τη σύνθεση του δίκλωνου DNA, ενσωματώνεται σε αυτό με αποτέλεσμα την εκπομπή φθορισμού. Στην παρούσα διπλωματική πραγματοποιήθηκε real-time PCR με την μέθοδο SYBR-Green. Η μέθοδος βασίζεται στην ανίχνευση του φθορισμού που παράγεται κατά την δεύσμεση των φθορίζουσων χρωστικών στο δίκλωνο DNA, το οποίο συντίθεται σε κάθε κύκλο της αντίδρασης. Αντίθετα, στα ειδικά συστήματα η ανίχνευση του προϊόντος επιτυγχάνεται με τη χρήση ειδικού, ως προς την αλληλουχία-στόχο, ανιχνευτή, όπως στην περίπτωση της Taqman qPCR.

2.11.2. Μελέτη της έκφρασης των νέων εναλλακτικών μεταγράφων του *CDK4*

Στην παρούσα διπλωματική εργασία, η μελέτη των επιπέδων έκφρασης των μεταγράφων του γονιδίου *CDK4* πραγματοποιήθηκε σε ένα ευρύ φάσμα ανθρώπινων κυτταρικών σειρών με τη χρήση της μεθόδου SYBR-Green qPCR. Για το σκοπό αυτό σχεδιάστηκαν εκκινητές που στοχεύουν συγκεκριμένα κάθε μετάγραφο (πίνακας 6). Κάθε ζεύγος εκκινητών σχεδιάστηκε κατάλληλα ώστε να στοχεύει ειδικά κάθε μετάγραφο του γονιδίου και επομένως, σε κάθε αντίδραση να ενισχύεται μόνο ένα εναλλακτικό *CDK4* μετάγραφο. Τα ζεύγη των εκκινητών που χρησιμοποιήθηκαν για τον προσδιορισμό του κάθε μεταγράφου παρουσιάζονται στον πίνακα 7. Επιπροσθέτως, η ανάλυση του προφίλ έκφρασης, με την χρήση της qPCR, πραγματοποιήθηκε στα 17 cDNAs, τα οποία χρησιμοποιήθηκαν και κατά την αλληλούχηση. Τελικά, ο ποσοτικός προσδιορισμός των νέων *CDK4* μεταγράφων πραγματοποιήθηκε σε κυτταρικές σειρές από αδενοκαρκίνωμα μαστού, πορογενές αδενοκαρκίνωμα μαστού, καρκίνο των ωοθηκών, αδενοκαρκίνωμα του ενδομητρίου, καρκίνωμα του τραχήλου της μήτρας, καρκίνο του προστάτη, καρκίνο της ουροδόχου κύστης, καρκίνωμα του νεφρού, καρκίνο του παχέος εντέρου, γαστρικό αδενοκαρκίνωμα, ηπατοκυτταρικό αδενοκαρκίνωμα, καρκίνο του εγκεφάλου, αδενοκαρκίνωμα του πνεύμονα, μελάνωμα, λέμφωμα, λευχαιμικά κύτταρα και σε φυσιολογικό εμβρυονικό νεφρό. Για σκοπούς κανονικοποίησης των αποτελεσμάτων χρησιμοποιήθηκε το γονίδιο *GAPDH* ως γονίδιο αναφοράς.

Πίνακας 6. Οι εκκινητές που χρησιμοποιήθηκαν κατά την μελέτη της έκφρασης των εναλλακτικών μεταγράφων του *CDK4*. Το όνομα κάθε εκκινητή βασίζεται στον αριθμό του εξωνίου που στοχεύει και το σύμβολο "/" δηλώνει τη θέση συρραφής στην οποία στοχεύει κάθε εκκινητής. Η θερμοκρασία τήξης (T_m) κάθε εκκινητή υπολογίστηκε με τη χρήση του Primer-BLAST.

Κατεύθυνση	**Ονομασία**	**Αλληλουχία (5′→3′)**	**Μήκος (nt)**	**T_m (ºC)**
Πρόσθιος	2/3F	CAATGTTGTCCGGCTGATGG	20	59.55
	2/4F	CCAATGTTGTCCGGATCTGATG	22	59.38
	2/5F	CCCAATGTTGTCCGGTTGTTAC	22	60.03
	2/6F	CAATGTTGTCCGGCCTCTCTT	21	60.61
	2F	GCCCTCAAGAGTGTGAGAGT	20	59.03
	1/3F	CTGGCGTGAGGCTGATGGA	19	62.02
	1/4F	GCTGGCGTGAGGATCTGAT	19	59.55
	1/5F	CTGGCGTGAGGTTGTTACAC	20	59.13
	1/6F	CTGGCGTGAGGCCTCTCTT	19	61.65
	1/7F	GCTGGCGTGAGCCTGATT	18	60.44
	1/8F	GGCTGGCGTGAGGAAATG	18	58.81
Ανάστροφος	2/8R	GCATTTCCGGACAACATTGGG	21	60.40
	3/6R	CACAGAAGAGAGGCCTTGATCG	22	60.74
	3/8R	CAGCATTTCCTTGATCGTTTCG	22	58.58
	4/6R	CAGAAGAGAGGCCACGGGT	19	61.29
	4/8R	GTCAGCATTTCCACGGGTGTA	21	60.61
	5/6R	GAAGAGAGGCTTTCGACGAAAC	22	59.59
	5/7R	CAGCCCAATCAGGTTTCGAC	20	59.20
	5/8R	AGTCAGCATTTCTTTCGACGAAAC	24	60.32
	6/7R	CAGCCCAATCAGGTCAAAGAT	21	58.27
	6/8R	GTCAGCATTTCTCAAAGATTTTGCC	25	59.88
	7R	TCGAGGCCAGTCATCCTCTG	20	61.04
	7outR	GCTCCCGACTCCTCCATC	18	58.87
	8R	AGCCACTCCATTGCTCACTC	20	60.04

Οι αντιδράσεις qPCR πραγματοποίθηκαν στον κυκλοποιητή 7500 Fast Real-Time PCR system (Applied Biosystems) με τελικό όγκο 10 μl, που περιείχε: 5 μl από 2X Kapa SYBR® Fast qPCR Master Mix (Kapa Biosystems, Inc., Woburn, MA, USA), 1 μl κάθε εκκινητή, σε συγκέντρωση 2 μM, και 1 μl από το cDNA δείγμα.

Πίνακας 7. Τα ειδικά, για κάθε μετάγραφο, ζεύγη των εκκινητών που χρησιμοποιήθηκαν κατά την RT-qPCR για τη μελέτη του προφιλ έκφρασης των νέων εναλλακτικών *CDK4* μορίων.

***CDK4* μετάγραφο**	**Ονομασία εκκινητή**		**Μέγεθος προϊόντος (bp)**
	Πρόσθιος	**Ανάστροφος**	
v.2	2/3F	6/8R	488
v.3	2/3F	5/7R	439
v.4	2/3F	5/8R	438
v.5	2/3F	4/6R	328
v.6	2/3F	4/8R	327
v.7	2/3F	5/6R	268
v.8	2/3F	5/8R	270
v.9	2/3F	3/6R	162
v.10	2/3F	3/8R	157
v.11	2/4F	5/8R	303
v.12	2/5F	6/7R	188
v.13	2/5F	6/8R	186
v.14	2/5F	5/8R	136
v.15	2/6F	7R	100
v.16	2/6F	6/8R	74
v.17	2F	2/8R	129
v.18	1/3F	6/7R	488
v.19	1/3F	6/8R	486
v.20	1/3F	5/7R	437
v.21	1/3F	5/8R	436
v.22	1/3F	6/7R	378
v.23	1/3F	6/8R	376
v.24	1/3F	4/8R	325
v.25	1/3F	6/7R	320
v.26	1/3F	6/8R	318
v.27	1/3F	5/8R	268
v.28	1/3F	6/7R	210
v.29	1/3F	6/8R	208
v.30	1/3F	3/8R	155
v.31	1/4F	6/7R	353
v.32	1/4F	5/8R	301
v.33	1/4F	4/6R	191
v.34	1/4F	4/8R	190
v.35	1/5F	6/7R	184
v.36	1/5F	6/8R	182
v.37	1/5F	5/7R	133
v.38	1/5F	5/8R	132
v.39	1/6F	6/7R	74
v.40	1/6F	6/8R	72
v.41	1/7F	7outR	134
v.42	1/8F	8R	119

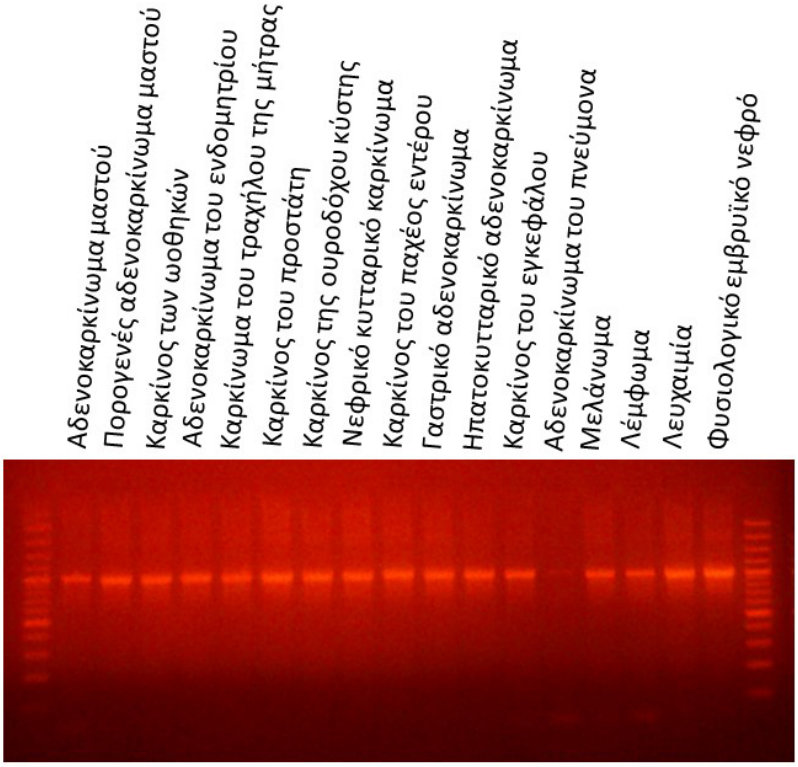
2.12. Παραγωγή και οπτικοποίηση των 3D μοντέλων των νέων CDK4 ισομορφών

Τα εναλλακτικά μετάγραφα του *CDK4*, τα οποία ταυτοποιήθηκαν, στη συνέχεια, ελέγχθηκαν ως προς την παρουσία αλληλουχιών, οι οποίες περιέχουν ανοιχτό πλαίσιο ανάγνωσης (Open Reading Frame, ORF), και, επομένως, έχουν την ικανότητα να κωδικοποιούν πρωτεϊνικές ισομορφές της CDK4 κινάσης. Τα μετάγραφα, που παρουσιάζουν ανοιχτό πλαίσιο ανάγνωσης, χαρακτηρίστηκαν ως κωδικά, ενώ όσα δεν φέρουν ORFs, αποτελούν μη κωδικά μόρια. Για τον προσδιορισμό της πρωτεϊνικής αλληλουχίας κάθε νέου κωδικού mRNA χρησιμοποιήθηκε το διαδικτυακό πρόγραμμα ExPASy, το οποίο επιτρέπει τη μετάφραση μιας νουκλεοτιδικής αλληλουχίας (DNA / RNA) σε αλληλουχία αμινοξέων [122]. Τελικά, για τα mRNA μετάγραφα του *CDK4,* τα οποία προβλέπεται να κωδικοποιούν πρωτεϊνικές ισομορφές, υψηλής ομολογίας με την κύρια CDK4, δημιουργήθηκαν τρισδιάστατα μοντέλα, τα οποία αναπαριστούν την δομή των πρωτεϊνικών ισομορφών της κινάσης. Η παραγωγή των 3D μοντέλων έγινε με τη χρήση του διαδικτυακού εργαλείου I-TASSER [123], ενώ η οπτικοποίηση των μοντέλων πραγματοποιήθηκε με το πρόγραμμα PyMOL.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1. Προφίλ έκφρασης του γονιδίου *CDK4* σε διαφορετικούς τύπους ιστών

Το γονίδιο *CDK4* ενισχύθηκε σε 17 δείγματα cDNA κυτταρικών σειρών, οι οποίες προέρχονται από διαφορετικούς τύπους ιστών, μέσω απλής PCR. Η ηλεκτροφόρηση των προϊόντων της απλής PCR οδήγησε στην επιβεβαίωση της έκφρασης του γονιδίου *CDK4* στη συντριπτική πλειοψηφία των δειγμάτων που μελετήθηκαν, όπως παρουσιάζεται στην εικόνα 18. Επιπλέον, το αποτέλεσμα της ηλεκτροφόρησης φανερώνει ότι η έκφραση του *CDK4* μεταγράφου σε κύτταρα που προέρχονται από ιστό του πνεύμονα είναι μειωμένη.

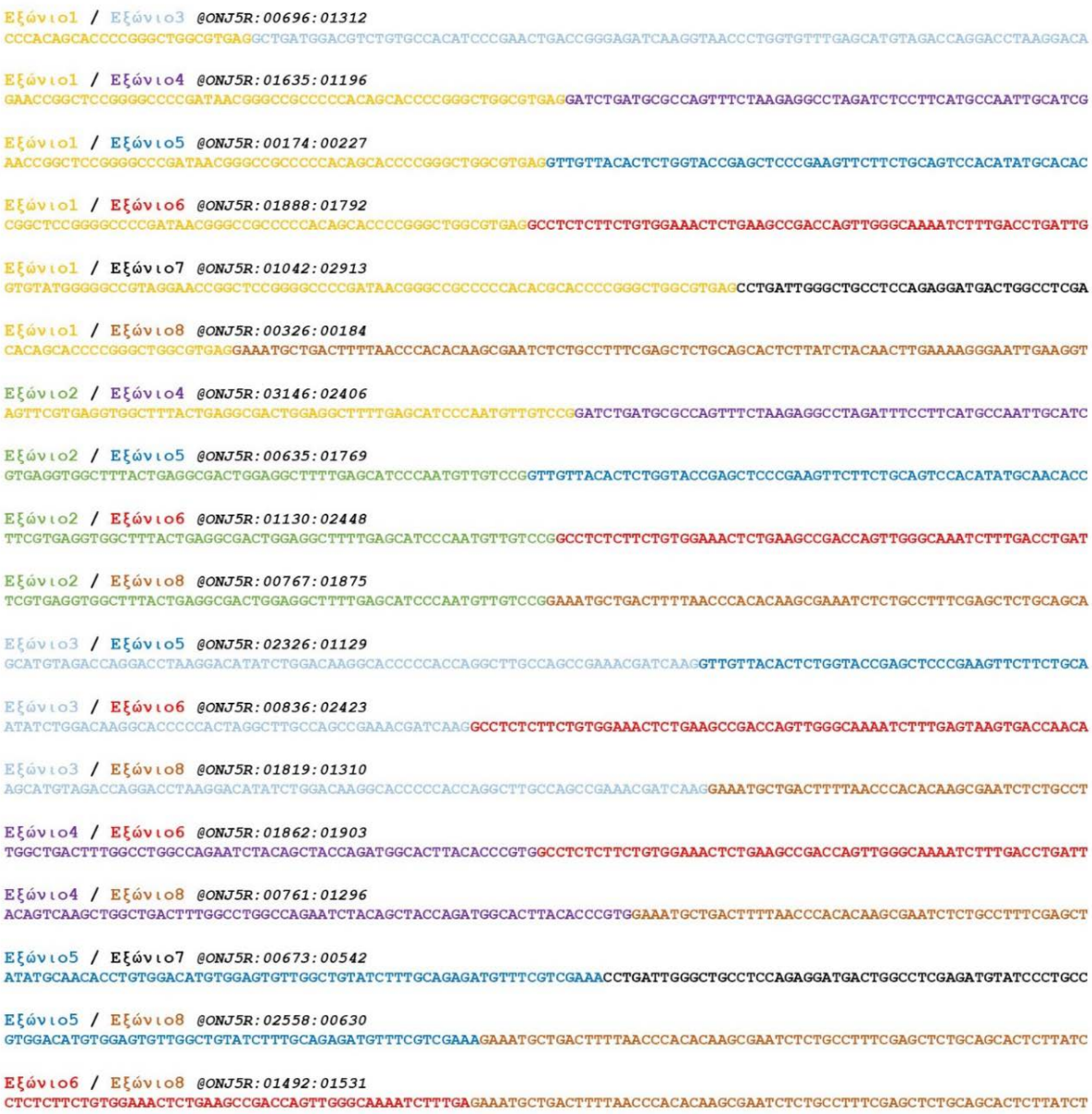


Εικόνα 18. Το αποτέλεσμα της ηλεκτροφόρησης των προϊόντων PCR 17 διαφορετικών δειγμάτων, που προέρχονται από 17 διαφορετικές κυτταρικές σειρές, αποκαλύπτει το προφίλ έκφρασης του γονιδίου *CDK4* στους διάφορους ιστούς.

3.2. Νέες εναλλακτικές θέσεις συρραφής του γονιδίου *CDK4*

Τα αποτελέσματα, τα οποία λήφθηκαν από την αλληλούχηση επόμενης γενιάς, μέσω της πλατφόρμας PGM της Ion Torrent™, αναλύθηκαν μέσω του προγράμματος οπτικοποίησης IGV και μέσω του αλγορίθμου “ASDT” και

επιβεβαίωσαν την ύπαρξη των 8 γνωστών εξωνίων του γονιδίου, καθώς και τα γεγονότα συρραφής μεταξύ των εξωνίων, τα οποία συγκροτούν το πλήρες λειτουργικό γνωστό μετάγραφο του *CDK4* (*CDK4* v.1, NM_000075.4). Επιπλέον, τα αποτελέσματα της βιοπληροφορικής ανάλυσης, μέσω του IGV, καθώς και με τη χρήση του αλγορίθμου ASDT, αποκάλυψαν την ύπαρξη 18 νέων θέσεων εναλλακτικού ματίσματος, οι οποίες αφορούν στα 8 γνωστά εξώνια του γονιδίου (εικόνα 19).



Εικόνα 19. Οι πειραματικές αλληλουχίες των NGS αποτελεσμάτων αποκάλυψαν την ύπαρξη 18 νέων εναλλακτικών θέσεων συρραφής μεταξύ των γνωστών εξωνίων του γονιδίου *CDK4*.

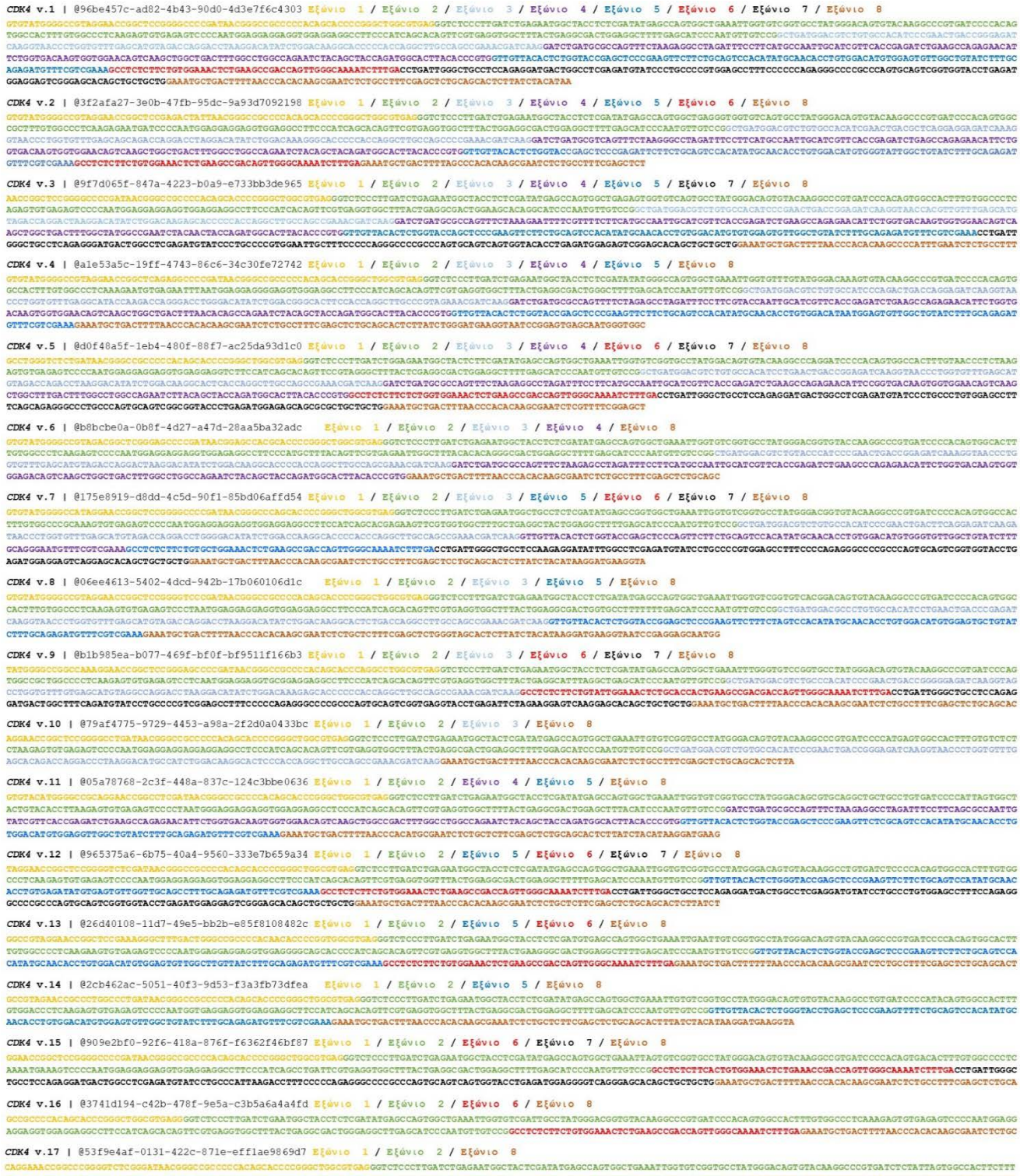
3.3. Νέα εναλλακτικά μετάγραφα του γονιδίου *CDK4*

Η στοίχιση των πειραματικών αλληλουχιών της αλληλούχησης τρίτης γενιάς, με την χρήση της πλατφόρμας αλληλούχησης ONT, πραγματοποιήθηκε με τη βοήθεια του προγράμματος IGV, καθώς και με τον αλγόριθμο “ASDT” και επικυρώνει όλα τα αποτελέσματα, που λήφθηκαν ύστερα από τη βιοπληροφορική ανάλυση των NGS δεδομένων. Στον πίνακα 6 παρουσιάζονται οι γνωστές θέσεις συρραφής μεταξύ των εξωνίων, καθώς και οι νέες θέσεις εναλλακτικής συρραφής, όπως αποκαλύφθηκαν από την βιοπληροφορική ανάλυση των δεδομένων, που συλλέχθηκαν και με τις δύο μεθοδολογίες μαζικής παράλληλης αλληλούχησης.

Πίνακας 8. Λίστα των γνωστών και των νέων θέσεων συρραφής, οι οποίες αποκαλύφθηκαν από τη βιοπληροφορική ανάλυση των αποτελεσμάτων, και ο αριθμός των πειραματικών αλληλουχιών, που επιβεβαιώνουν κάθε γεγονός συρραφής, μέσω μεθοδολογιών TGS και NGS αλληλούχησης.

	Θέσεις συρραφής μεταξύ γνωστών εξωνίων	**Αριθμός πειραματικών αλληλουχιών για κάθε γεγονός συρραφής μέσω TGS**	**Αριθμός πειραματικών αλληλουχιών για κάθε γεγονός συρραφής μέσω NGS**
Γνωστές θέσεις συρραφής	1 - 2	62381	94800
	2- 3	171786	158610
	3 - 4	172207	211454
	4 - 5	128893	207149
	5 - 6	102348	156111
	6 - 7	175501	150030
	7 - 8	199642	168203
Εναλλακτικές θέσεις συρραφής	1 - 3	10908	10670
	1 - 4	516	655
	1 - 5	4467	6520
	1 - 6	2025	3488
	1 - 7	54	61
	1 - 8	197	895
	2 - 4	115	207
	2 - 5	151	308
	2 - 6	42	73
	2 - 8	15	23
	3 - 5	1329	2142
	3 - 6	84	157
	3 - 8	41	62
	4 - 6	281	567
	4 - 8	114	199
	5 - 7	1343	885
	5 - 8	1062	1642
	6 - 8	2875	3718

Το χαρακτηριστικό πλεονέκτημα της αλληλούχησης τρίτης γενιάς, μέσω της πλατφόρμας MinION® Mk1C (ONT), επέτρεψε, χάρη στη χρήση της εφαρμογής στοχευμένης DNA αλληλούχησης (targeted DNA-seq), την παραγωγή πειραματικών αλληλουχιών μεγάλου μήκους, με αποτέλεσμα οι πειραματικές αλληλουχίες, που περιέχουν τις εναλλακτικές θέσεις συρραφής, να αποτελούνται από το συνολικό cDNA μόριο, δηλαδή να περιλαμβάνουν όλη την περιοχή του μορίου από το πρώτο ως το τελευταίο εξώνιο και επομένως, δεν υπάρχει η ανάγκη της συναρμολόγησης του μεταγράφου, καθώς κάθε πειραματική αλληλουχία αναπαριστά ένα πλήρους μήκους *CDK4* μετάγραφο. Οι εικόνες 20 και 21 παρουσιάζουν πειραματικές αλληλουχίες, που λήφθηκαν από τα αποτελέσματα της TGS αλληλούχησης και οι οποίες οδήγησαν στην ανακάλυψη των νέων μεταγράφων του γονιδίου *CDK4*. Συγκεκριμένα, ταυτοποιήθηκαν 41 νέα *CDK4* μετάγραφα, τα οποία προέρχονται από ένα ή περισσότερα γεγονότα εναλλακτικής συρραφής μεταξύ των οκτώ εξωνίων.



Εικόνα 20. Πειραματικές αλληλουχίες της μεθοδολογίας TGS, οι οποίες επιβεβαιώνουν την ύπαρξη νέων εναλλακτικών μεταγράφων του ανθρώπινιου γονιδίου *CDK4* (*CDK4 v.2 – v.17*).

Εικόνα 21. Πειραματικές αλληλουχίες οι οποίες λήφθηκαν από την αλληλούχηση τρίτης γενιάς, οι οποίες επιβεβαιώνουν την ύπαρξη νέων εναλλακτικών μεταγράφων του ανθρώπινιου γονιδίου *CDK4* (*CDK4 v.18 – v.42*).

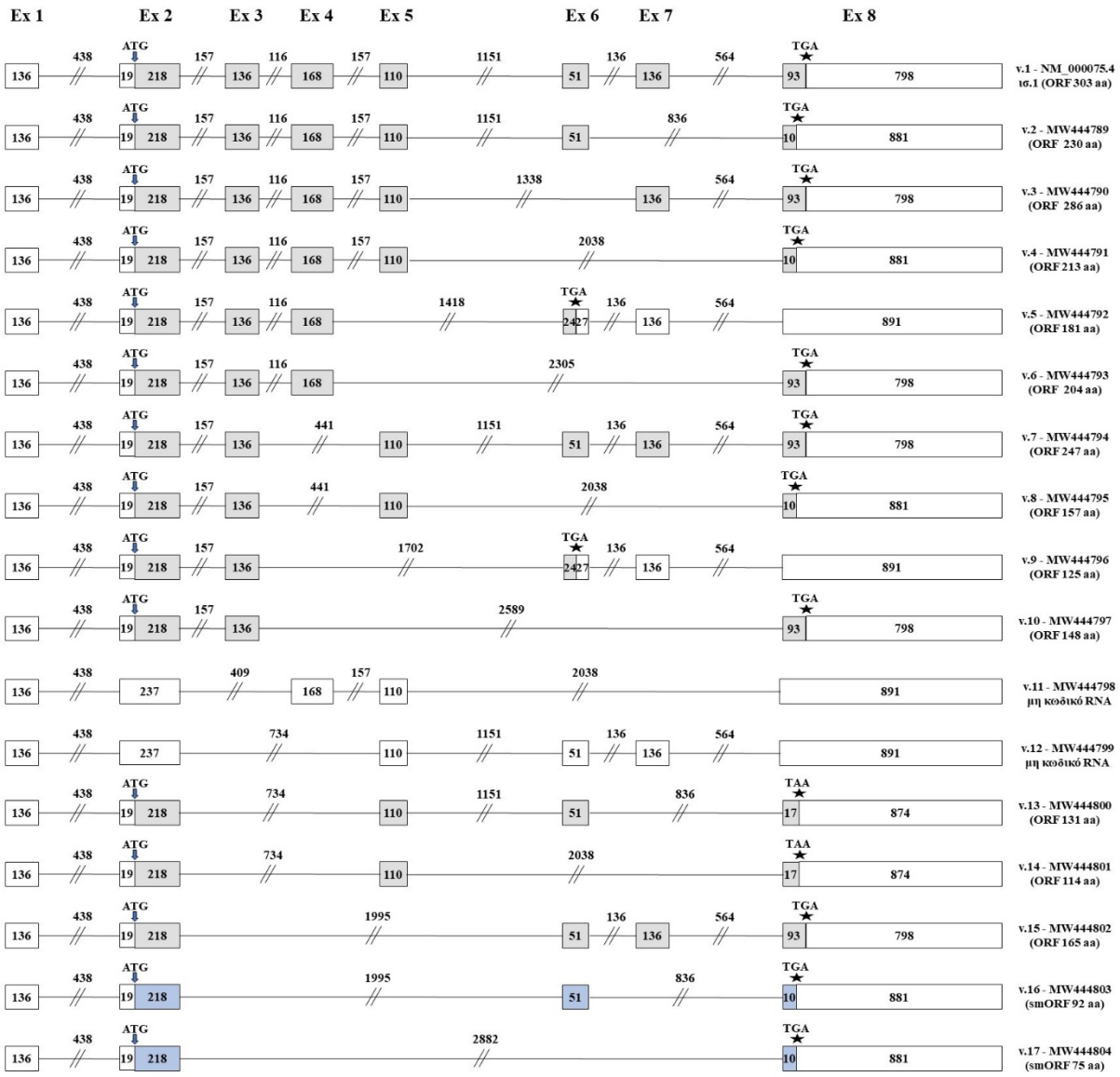
Με δεδομένο ότι, το γνωστό κωδικόνιο έναρξης “ATG” βρίσκεται στο δεύτερο εξώνιο του γονιδίου και με βάση τα αποτελέσματα της οπτικοποίησης των πειραματικών αλληλουχιών, μέσω του προγράμματος IGV (εικόνα 22), μπορούμε να κατατάξουμε τα νέα μετάγραφα σε δύο ομάδες ανάλογα με την ύπαρξη ή όχι του δεύτερου εξωνίου στην νουκλεοτιδική ακολουθία του κάθε μεταγράφου και, επομένως, η πρώτη ομάδα αποτελείται από τα μετάγραφα *CDK4* v.2 - v.17, ενώ η δεύτερη από τα υπόλοιπα μετάγραφα (*CDK4* v.18 - v.42), τα οποία πιθανότατα να φέρουν εναλλακτικά κωδικόνια έναρξης.

Εικόνα 22. Η οπτικοποίηση των πειραματικών αλληλουχιών, που λήφθηκαν από την TGS αλληλούχηση, πραγματοποιήθηκε με την χρήση του προγράμματος IGV και προσδιορίστηκαν 41 νέα εναλλακτικά μετάγραφα, τα οποία συγκροτούνται ύστερα από διαφορετικά γεγονότα εναλλακτικής συρραφής των 8 γνωστών εξωνιών του *CDK4* γονιδίου. Τα 42 συνολικά μετάγραφα του γονιδίου μπορούν να κατηγοριοποιηθούν σε δύο ομάδες, με βάση την παρουσία ή απουσία του εξωνίου 2 στην νουκλεοτιδική τους αλληλουχία.

3.4. Μελέτη της δομής των νέων *CDK4* μεταγράφων που μοιράζονται το γνωστό κωδικόνιο έναρξης

Η ανάλυση των αποτελεσμάτων της TGS αλληλούχησης οδήγησε στην ανακάλυψη 16 νέων εναλλακτικών μεταγράφων του γονιδίου *CDK4* (*CDK4* v.2 - v.17), τα οποία περιέχουν το γνωστό κωδικόνιο έναρξης, το οποίο είναι τμήμα της αλληλουχίας του δεύτερου εξωνίου του γονιδίου (εικόνα 23). Οι πειραματικές αλληλουχίες, οι οποίες λήφθηκαν από την αλληλούχηση μέσω της Oxford Nanopore Technologies® και αντιστοιχούν στην ομάδα των μεταγράφων που αναφέρθηκαν, προέρχονται από γεγονότα εναλλακτικής συρραφής και αφορούν την παράλειψη ενός ή περισσοτέρων εξωνίων ή / και τον αμοιβαίο αποκλεισμό εξωνίων. Ενδεικτικά, στην εικόνα 20 παρουσιάζεται για καθένα από τα μετάγραφα *CDK4* v.2 - v.17 μία πειραματική αλληλουχία, όπως αυτή προσδιορίστηκε κατά την αλληλούχηση τρίτης γενιάς με τη χρήση της πλατφόρμας αλληλούχησης MinION™ Mk1C της ONT.

Με βάση την μελέτη, που πραγματοποιήθηκε για την πρόβλεψη των μεταγράφων που έχουν ανοιχτό πλαίσιο ανάγνωσης (ORF), αποδεικνύεται ότι εκτός από δύο μετάγραφα *CDK4* (v.11 και v.12) τα υπόλοιπα 14 έχουν ORFs και, ως εκ τούτου, προβλέπεται να κωδικοποιούν νέες πρωτεϊνικές ισομορφές (εικόνα 23). Επιπλέον, οπτικοποιημένα δεδομένα για κάθε μετάγραφο, τα οποία παρουσιάζονται αναλυτικά στην εικόνα 23, φανερώνουν ότι τα μετάγραφα v.2 - v.10 της *CDK4* κινάσης περιέχουν τα γνωστά εξώνια 2 και 3 και, επομένως, έχουν την ικανότητα να κωδικοποιούν πρωτεϊνικές ισομορφές, οι οποίες έχουν πανομοιότυπο αμινο-τελικό τμήμα (αμινοξέα 1-96) τόσο μεταξύ τους όσο και με την γνωστή πρωτεϊνική κινάση, η οποία κωδικοποιείται από το κύριο *CDK4* μετάγραφο. Αντίθετα, οι αλληλουχίες των υπόλοιπων μεταγράφων του γονιδίου, που έχουν ORFs, (*CDK4* v.13 - v.17) στερούνται του εξωνίου 3 και για το λόγο αυτό, οι πρωτεϊνικές ισομορφές, που δύναται να κωδικοποιούνται, διαθέτουν αμινοτελικές περιοχές με εμφανώς διαφοροποιημένη δομή σε σχέση με την κύρια πρωτεΐνη. Επιπλέον, κάθε εν δυνάμει πρωτεϊνική ισομορφή, που κωδικοποιείται από τα νέα μετάγραφα του *CDK4* v.2 - v.17 (εκτός των v.11 και v.12), διαθέτει διαφορετικές δομικές παραλλαγές στο καρβοξυτελικό άκρο, καθώς τα γεγονότα εναλλακτικού ματίσματος αφορούν στα εξώνια που κωδικοποιούν για την καρβοξυτελική περιοχή της πρωτεΐνης.



Εικόνα 23. Σχηματική αναπαράσταση των 17 μεταγράφων του γονιδίου *CDK4* που φέρουν και μοιράζονται το γνωστό κωδικόνιο έναρξης, το οποίο εδράζεται στο εξώνιο 2. Στην εικόνα παρουσιάζονται με ορθογώνια κουτιά τα εξώνια που συγκροτούν κάθε μετάγραφο, με βέλος (↓) η θέση του κωδικονίου έναρξης και με αστερίσκο (*) το πιθανό κωδικόνιο λήξης. Το γκρι χρώμα αντιπροσωπεύει τις κωδικές αλληλουχίες, ενώ το λευκό χρώμα τα μη κωδικά μετάγραφα. Το μπλε χρώμα αντιπροσωπεύει τις κωδικές αλληλουχίες που χαρακτηρίζονται από smORFs (<100 aa). Ο αριθμός κάθε μεταγράφου, το μήκος του ORF (μόνο για τα κωδικά mRNAs), καθώς και ο κωδικός πρόσβασης στη GenBank® παρουσιάζονται στα δεξιά κάθε μεταγράφου.

Τα εναλλακτικά μετάγραφα του *CDK4* v.2, v.3 και v.4 μοιράζονται ακριβώς την ίδια νουκλεοτιδική αλληλουχία ως το 3΄ άκρο του 5ου εξωνίου και με το γνωστό

CDK4 μετάγραφο v.1, καθώς τα γεγονότα εναλλακτικής συρραφής συμβαίνουν καθοδικά του 5ου εξωνίου (εικόνα 23). Επομένως, με βάση τη δομή τους, τα μετάγραφα αυτά έχουν την μεγαλύτερη πιθανότητα να διαθέτουν την ικανότητα κωδικοποίησης νέων λειτουργικών πρωτεϊνικών ισομορφών. Με βάση τα ORFs, κάθε πιθανή νέα πρωτεϊνική ισομορφή αποτελείται από 230, 286 και 213 αμινοξέα, αντίστοιχα, για κάθε ένα από τα μετάγραφα που αναφέρθηκαν, και κάθε μία από τις πρωτεΐνες αυτές περιέχει όλες τις δομικές περιοχές που συγκροτούν μια λειτουργική πρωτεϊνική κινάση αντίστοιχη της γνωστής CDK4. Πιο αναλυτικά, οι προβλεπόμενες πρωτεϊνικές ισομορφές, που μπορούν να κωδικοποιούνται, περιλαμβάνουν την περιοχή του βρόχου που είναι πλούσια σε γλυκίνη (GVGAYG), τη θέση δέσμευσης της κυκλίνης D1 (PISTVRE), το λειτουργικό μοτίβο που απαιτείται για την φωσφορυλίωση της κινάσης DFG-APE, καθώς και το βρόχο ενεργοποίησης της κινάσης που περιέχει τη θέση της φωσφορυλίωσης (T-loop) (εικόνα 24).

Το ανοιχτό πλαίσιο ανάγνωσης για καθένα από τα μετάγραφα v.5 και v.6 έχει μήκος 181 και 204 αμινοξέων, αντίστοιχα. Και τα δύο μετάγραφα στερούνται το εξώνιο 5 και επομένως, οι προβλεπόμενες ισομορφές των πρωτεϊνών αυτών, αν και περιλαμβάνουν τις περισσότερες από τις λειτουργικές περιοχές της κινάσης, περιέχουν μια τροποποιημένη περιοχή φωσφορυλίωσης και ένα ελλιπές μοτίβο DFG-APE, καθώς απουσιάζουν τα αμινοξέα APE. Τα μετάγραφα *CDK4* v.7 - v.10, που ταυτοποιήθηκαν, χαρακτηρίζονται από την κοινή απουσία του 4ου εξωνίου. Καθένα από αυτά τα μετάγραφα συγκροτείται από το εναλλακτικό μάτισμα του εξωνίου 3 με κάποιο απομακρυσμένο εξώνιο και επομένως, διαφέρουν οι προβλεπόμενες πρωτεϊνικές ισομορφές που δύναται να παράγονται. Η απουσία του 4ου εξωνίου υποδηλώνει αλλαγές στην περιοχή της φωσφορυλίωσης και την παρουσία ελαττωματικού DFG-APE μοτίβου, καθώς απουσιάζουν τα τρία αμινοξέρα DFG (εικόνες 24 & 26).

Στη συνέχεια, το νέο εναλλακτικό μετάγραφο v.11 είναι το μοναδικό μετάγραφο που φέρει την εναλλακτική θέση συρραφής του εξωνίου 2 με το εξώνιο 4. Το μετάγραφο αυτό αποτελεί ένα μη κωδικό μόριο RNA καθώς περιέχει ένα πρώιμο κωδικόνιο λήξης (premature termination codon, PTC). Επιπλέον, τα μετάγραφα *CDK4* v.12 - v.14 χαρακτηρίζονται από την κοινή απουσία των εξωνίων 3 και 4 (εικόνα 23). Το μετάγραφο v.12 περιέχει ένα πρώιμο κωδικόνιο λήξης και, για το λόγο αυτό, χαρακτηρίζεται ως μη κωδικό μόριο. Οι νουκλεοτιδικές αλληλουχίες των

μεταγράφων v.13 και v.14 είναι πιθανό να κωδικοποιούν για διαφορετικές πρωτεϊνικές ισομορφές η δομή των οποίων όμως χαρακτηρίζεται από σημαντικές ελλείψεις, καθώς και οι δύο στερούνται εντελώς την περιοχή ενεργοποίησης της φωσφορυλίωσης και το μοτίβο DFG-APE. Επιπλέον, η πρωτεϊνική ισομορφή, που προβλέπεται ότι προέρχεται από το μετάγραφο v.15, αναμένεται να χαρακτηρίζεται επίσης, από την απουσία των δύο αυτών βασικών περιοχών της πρωτεϊνικής κινάσης και, επιπλέον, φέρει περισσότερες δομικές αλλοιώσεις (εικόνες 24 & 26).

Τέλος, σύμφωνα με τα αποτελέσματα της ανάλυσης για τα μετάγραφα *CDK4* v.16 και v.17, οι προβλεπόμενες πιθανές πρωτεϊνικές ισομορφές, που μπορούν να κωδικοποιηθούν, χαρακτηρίζονται από άκρως σημαντικές δομικές ελλείψεις, καθώς απουσιάζουν σημαντικές νουκλεοτιδικές αλληλουχίες στο mRNA, λόγω των εναλλακτικών θέσεων συρραφής του εξωνίου 2 με απομακρυσμένα εξώνια του γονιδίου (εικόνα 23). Τα μετάγραφα αυτά περιέχουν μικρά σε μήκος πλαίσια ανάγνωσης (small ORFs, smORFs), μικρότερα από 100 κωδικόνια, (92aa και 75aa, αντίστοιχα) και τα πιθανά παραγόμενα πεπτίδια, που μπορεί να κωδικοποιούνται, μπορεί να έχουν διαφορετικό λειτουργικό ή ρυθμιστικό ρόλο.

Αμινοξική ακολουθία

v.1	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQS TYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHK RISAFRALQHSYLHKDEGNPE
v.2	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQS TYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFEKC
v.3	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQS TYATPVDMWSVGCIFAEMFRRNLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDE GNPE
v.4	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQS TYATPVDMWSVGCIFAEMFRRKKC
v.5	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVASLLWKL
v.6	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVEMLTFNPHKRISAFR ALQHSYLHKDEGNPE
v.7	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSL PRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.8	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKKC
v.9	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFE HVDQDLRTYLDKAPPPGLPAETIKASLLWKL

Αμινοξική ακολουθία	
v.10	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGLPISTVREVALLRRLEAFEHPNVVRLMDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.13	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGLPISTVREVALLRRLEAFEHPNVVRLLHSGTELPKFFCSPHMQHLWTCGVLAVSLQRCFVESLSSVETLKPTSWAKSLRNADF
v.14	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGLPISTVREVALLRRLEAFEHPNVVRLLHSGTELPKFFCSPHMQHLWTCGVLAVSLQRCFVERNADF
v.15	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGLPISTVREVALLRRLEAFEHPNVVRPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.16	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGLPISTVREVALLRRLEAFEHPNVVRPLFCGNSEADQLGKIFEKC
v.17	MATSRYEPVAEIGVGAYGTVYKARDPHSGHFVALKSVRVPNGGGGGGLPISTVREVALLRRLEAFEHPNVVRKC
v.18	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.19	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFEKC
v.20	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRNLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.21	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKKC
v.22	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVASLLWKL
v.23	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVASLLWKL
v.24	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKDLMRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.25	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.26	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFEKC
v.27	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKKC
v.30	MDVCATSRTDREIKVTLVFEHVDQDLRTYLDKAPPPGLPAETIKEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.31	MRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.32	MRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVVVTLWYRAPEVLLQSTYATPVDMWSVGCIFAEMFRRKKC
v.34	MRQFLRGLDFLHANCIVHRDLKPENILVTSGGTVKLADFGLARIYSYQMALTPVEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.35	MWSVGCIFAEMFRRKPLFCGNSEADQLGKIFDLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE
v.36	MWSVGCIFAEMFRRKPLFCGNSEADQLGKIFEKC
v.37	MWSVGCIFAEMFRRNLIGLPPEDDWPRDVSLPRGAFPPRGPRPVQSVVPEMEESGAQLLLEMLTFNPHKRISAFRALQHSYLHKDEGNPE

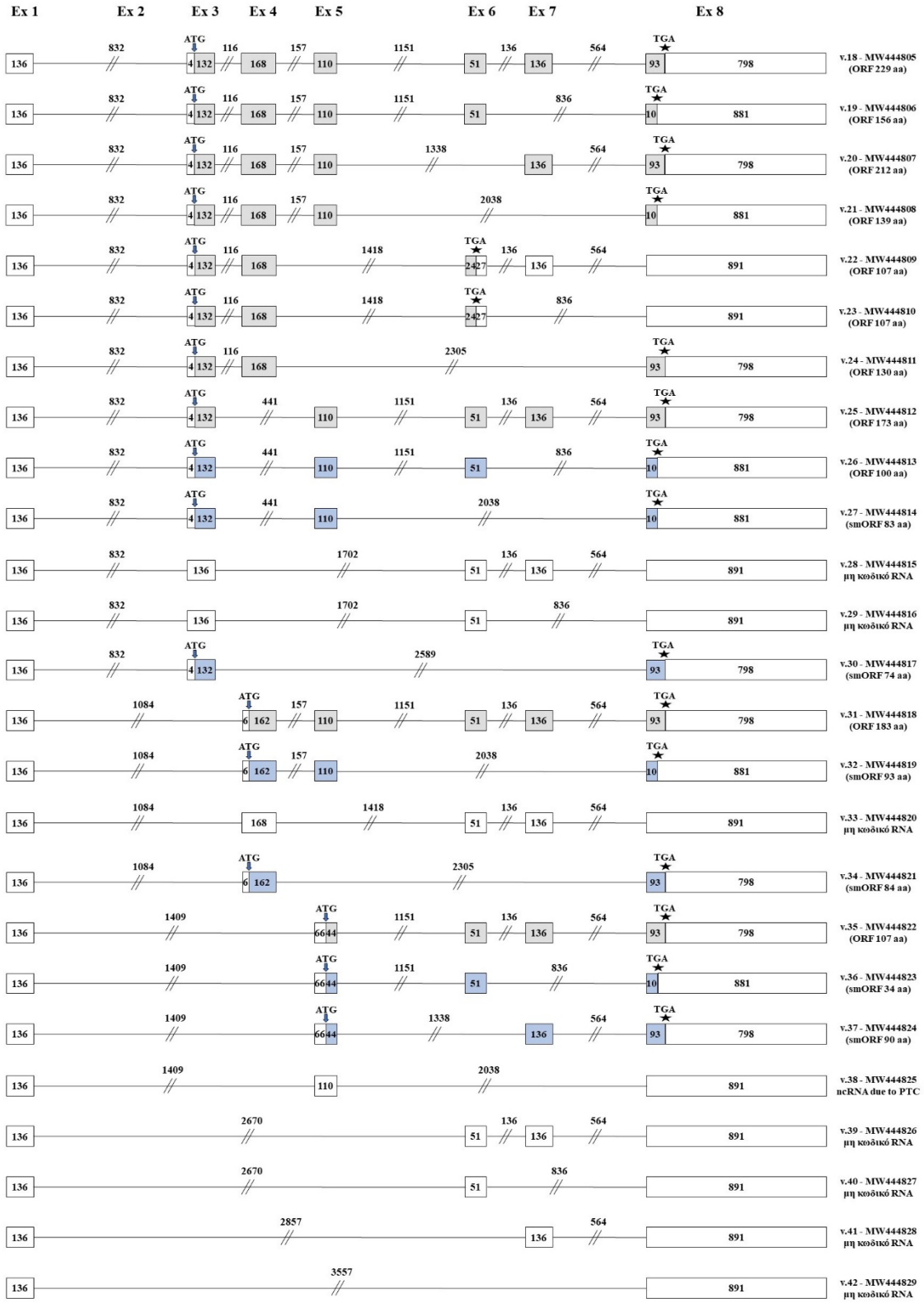
Εικόνα 24. Η αμινοξική ακολουθία για κάθε μία από τις προβλεπόμενες πρωτεϊνικές ισομορφές, σύμφωνα με τα ORFs που προκύπτουν από κάθε μετάγραφο. Η παρουσία ή απουσία κάθε εξωνίου επηρεάζει άμεσα την σειρά των αμινοξέων στην κάθε πρωτεΐνη. Τα αμινοξέα, που εμφανίζονται με διαφορετικό χρώμα, είναι υπεύθυνα για το σχηματισμό χαρακτηριστικών δομικών περιοχών της πρωτεΐνης. Συγκεκριμένα, με μπλε χρώμα παρουσιάζεται η περιοχή πλούσια σε γλυκίνη (GVGAYG), με γαλάζιο το αμινοξύ λυσίνη (K35) στη θέση 35, που είναι υπεύθυνο για τη σύνδεση της πρωτεΐνης με άλλα μόρια, με πορτοκαλί χρώμα η περιοχή σύνδεσης της κινάσης με την κυκλίνη (PISTVRE). Το μπορντό

χρώμα αφορά στα τελευταία αμινοξέα της αμινοτελικής περιοχής (FEHV), το μωβ χρώμα προσδιορίζει το ασπαραγινικό οξύ (D140) στη θέση 140, που είναι υπεύθυνο για την ενεργοποίηση της κινάσης, ενώ με κόκκινο χρώμα εμφανίζεται το μοτίβο DFG-APE, που δημιουργεί τη θηλιά στην οποία βρίσκεται η θέση φωσφορυλίωση (T172) του ενζύμου, η οποία παρουσιάζεται με πράσινο χρώμα (QMALTPVVVTLW).

3.5. Ανάλυση της δομής των νέων *CDK4* μεταγράφων που διαθέτουν εναλλακτικά κωδικόνια έναρξης

Η βιοπληροφορική ανάλυση και η ανίχνευση νέων εναλλακτικών θέσεων συρραφής μεταξύ των εξωνίων αποκάλυψαν την ύπαρξη νέων μεταγράφων, τα οποία χαρακτηρίζονται από πλήρη απουσία του δεύτερου εξωνίου, και, επομένως, δεν διαθέτουν το γνωστό κωδικόνιο έναρξης, το οποίο είναι απαραίτητο για την έναρξη της πρωτεϊνοσύνθεσης (εικόνα 25). Συγκεκριμένα, 13 μετάγραφα του γονιδίου *CDK4,* και συγκεκριμένα τα μετάγραφα v.18 - v.30, περιλαμβάνουν την νέα θέση εναλλακτικού ματίσματος μεταξύ των εξωνίων 1 και 3. Το μετάγραφο v.18 περιέχει όλα τα άλλα εξώνια του γνωστού μεταγράφου, ενώ τα υπόλοιπα 12 μετάγραφα προέρχονται από περισσότερα του ενός γεγονότα εναλλακτικού ματίσματος μεταξύ των εξωνίων γεγονός που οδηγεί σε αρκετά διαφοροποιημένες νουκλεοτιδικές αλληλουχίες σε σχέση με το κύριο μετάγραφο *CDK4* v.1.

Σε επίπεδο πρωτεΐνης, το ανοιχτό πλαίσιο ανάγνωσης κάθε μεταγράφου οδήγησε στην αναγνώριση μιας τριπλέτας “ATG”, η οποία βρίσκεται στο εξώνιο 3 και μπορεί να λειτουργεί ως εναρκτήριο κωδικόνιο για την σύνθεση πρωτεϊνικών ισομορφών, καθώς απουσιάζει το γνωστό κωδικόνιο έναρξης, που κωδικοποιείται από την αλληλουχία που βρίσκεται στο εξώνιο 2. Αναλυτικότερα, το νέο πιθανό κωδικόνιο έναρξης βρίσκεται στο 5΄ άκρο του εξωνίου 3 και συγκεκριμένα ξεκινάει από το πέμπτο νουκλεοτίδιο (εικόνα 25). Με την προϋπόθεση ότι το νέο αυτό κωδικόνιο έναρξης είναι λειτουργικό και χρησιμοποιείται ως εναρκτήρια θέση για πρωτεϊνοσύνθεση, τα εναλλακτικά μετάγραφα v.18 - v.30, εκτός mRNA των *CDK4* v.28 και v.29, διαθέτουν ORFs και, ως εκ τούτου, πιθανότατα αποτελούν κωδικά μόρια ικανά να συνθέσουν πρωτεΐνες. Επιπλέον, το μετάγραφο *CDK4* v.22 διαθέτει ένα πρώιμο κωδικόνιο λήξης και, για το λόγο αυτό, πιθανότατα, λειτουργεί ως μη κωδικό μόριο.

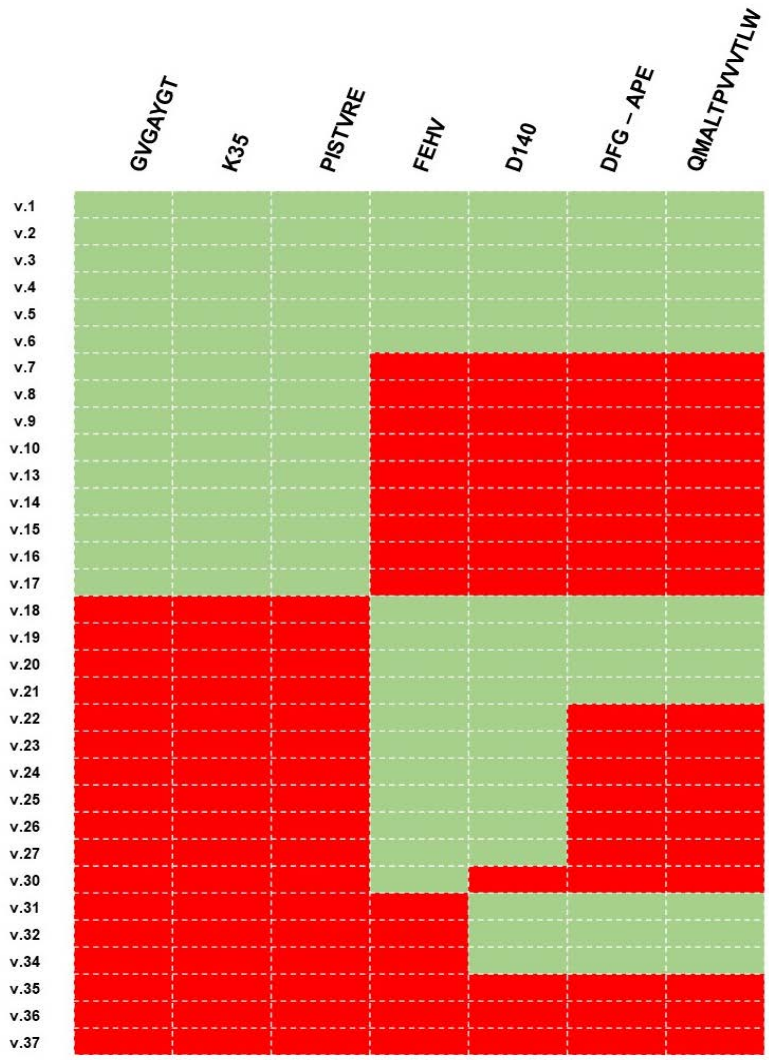


Εικόνα 25. Σχηματική αναπαράσταση των *CDK4* μεταγράφων που φέρουν νέες θέσεις συρραφής μεταξύ του εξωνίου 1 και των εξωνίων 3, 4, 5, 6, 7 και 8. Παρουσιάζονται τα εξώνια κάθε μεταγράφου, η θέση του κωδικονίου έναρξης, το πιθανό κωδικόνιο λήξης και ο αριθμός των αμινοξέων που συγκροτούν την πιθανή προβλεπόμενη πρωτεΐνη που κωδικοποιείται από κάθε μετάγραφο. Με γκρι εμφανίζονται οι κωδικές αλληλουχίες, ενώ με

λευκό τα μη κωδικά μετάγραφα. Τα μπλε χρώμα αντιπροσωπεύει τις κωδικές αλληλουχίες που χαρακτηρίζονται από smORFs (<100 aa).

Σύμφωνα με τη μελέτη της δομής των συγκεκριμένων προβλεπόμενων πρωτεϊνών, που προέρχονται από τα μετάγραφα που περιέχουν την εναλλακτική θέση συρραφής μεταξύ των εξωνίων 1 και 3, η εκάστοτε πρωτεϊνική ισομορφή αναμένεται να φέρει σημαντικές παραλλαγές σε σύγκριση με την λειτουργική κινάση, που κωδικοποιείται από το κύριο μετάγραφο v.1, λόγω της πλήρης απουσίας του εξωνίου 2, το οποίο κωδικοποιεί για τα 74 από τα συνολικά 96 αμινοξικά κατάλοιπα που συγκροτούν το αμινοτελικό τμήμα της CDK4 πρωτεΐνης. Κατά συνέπεια, οι προβλεπόμενες πρωτεϊνικές ισομορφές διαθέτουν μια διαφορετική αμινοτελική περιοχή η οποία διαθέτει 22 μόνο αμινοξέα που κωδικοποιούνται από τη νουκλεοτιδική αλληλουχία του εξωνίου 3 και επομένως, δεν διαθέτουν τον πλούσιο σε γλυκίνη βρόχο (G-rich loop), καθώς ούτε και την θέση σύνδεσης της κυκλίνης (PISTVRE) (εικόνες 24 & 26).

Τέλος, οι πειραματικές αλληλουχίες σύμφωνα με την αλληλούχηση τρίτης γενιάς αποκάλυψαν την ύπαρξη 12 ακόμα μεταγράφων. Τα 12 αυτά μετάγραφα (*CDK4* v.31 - v.42) στερούνται των εξωνίων 2 και 3 και χαρακτηρίζονται από την εναλλακτική συρραφή του πρώτου εξωνίου με τα πιο απομακρυσμένα εξώνια 4, 5, 6, 7 και 8 (εικόνα 25). Συγκεκριμένα, τέσσερα από τα μετάγραφα αυτά (*CDK4* v.31 - v.34) διαθέτουν τη νέα θέση συρραφής μεταξύ των εξωνίων 1 και 4. Ομοίως, τέσσερα επιπλέον μετάγραφα (*CDK4* v.35 - v. 38) χαρακτηρίζονται από την εναλλακτική συρραφή μεταξύ των εξωνίων 1 και 5, η οποία περιλαμβάνει την ταυτόχρονη παράλειψη τριών διαδοχικών εξωνίων, που αφορούν στα εξώνια 2, 3 και 4. Τα τελευταία τέσσερα μετάγραφα του *CDK4,* που ταυτοποιήθηκαν, περιλαμβάνουν σχεδόν αποκλειστικά νέες θέσεις συρραφής και χαρακτηρίζονται από σημαντική αλλαγή όσο αφορά στη νουκλεοτιδική τους αλληλουχία σε σχέση με το γνωστό *CDK4* μόριο. Εν συντομία, τα μετάγραφα *CDK4* v.39 και v.40 μοιράζονται το νέο εναλλακτικό γεγονός ματίσματος μεταξύ των εξωνίων 1 και 6, ενώ το *CDK4* v.41 περιλαμβάνει τη συρραφή του εξωνίου 1 με το εξώνιο 7 και το *CDK4* v.42 παράγεται από την άμεση σύνδεση του εξωνίου 1 με το τελευταίο εξώνιο του γονιδίου, το εξώνιο 8.



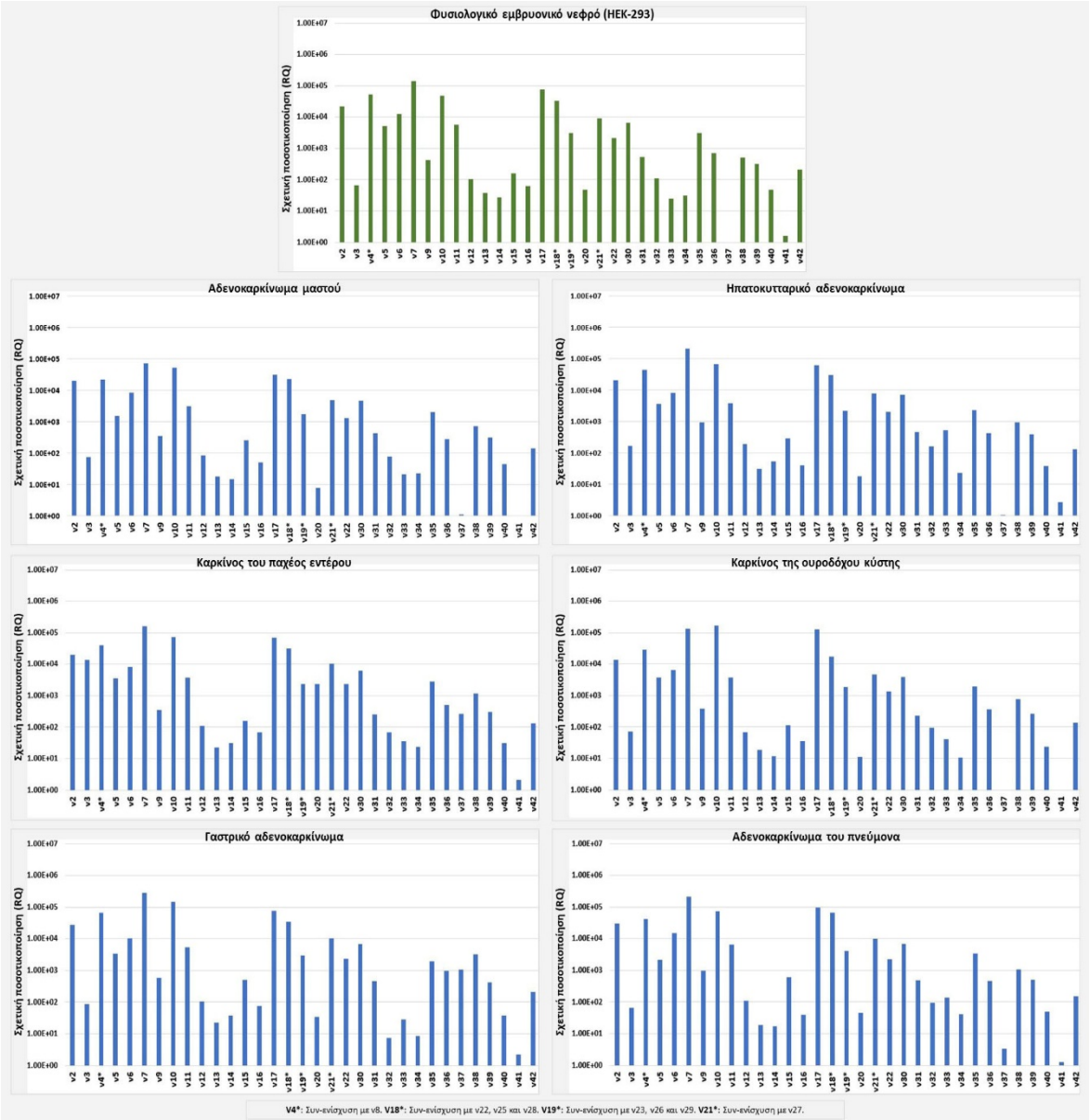
Εικόνα 26. Σχηματική αναπαράσταση των δομικών περιοχών που συγκροτούν κάθε προβλεπόμενη πρωτεινική ισομορφή που προκύπτει από τα κωδικά μετάγραφα του *CDK4,* ύστερα από τη διερεύνηση της ύπαρξης ORF σε καθένα από αυτά. Το πράσινο χρώμα υποδηλώνει την παρουσία του μοτίβου, ενώ το κόκκινο χρώμα την απουσία της συγκεκριμένης περιοχής.

Η διερεύνηση της ύπαρξης ORFs αποκάλυψε την παρουσία εναλλακτικών κωδικονίων έναρξης στα εξώνια 4 και 5, για τα μετάγραφα που διαθέτουν τις θέσεις συρραφής μεταξύ των εξωνίων 1 και 4 και 1 και 5 αντίστοιχα (εικόνα 25). Ωστόσο, εκτός του μεταγράφου v.31, του οποίου το ORF οδηγεί στην παραγωγή μιας πρωτεΐνης 183 αμινοξέων, τα προβλεπόμενα πεπτίδια, που πιθανότατα κωδικοποιούνται από τα συγκεκριμένα μετάγραφα, έχουν μήκος μικρότερο των 100 αμινοξέων και διαθέτουν διαφορετικό ρόλο από αυτόν που φέρει μια λειτουργική κινάση. Το μετάγραφο *CDK4* v.31 έχει την ικανότητα κωδικοποίησης μιας

πρωτεϊνικής ισομορφής από την οποία όμως απουσιάζουν βασικά μοτίβα που βρίσκονται στην αμινοτελική περιοχή της CDK4 κινάσης και κωδικοποιούνται από τα εξώνια 2 και 3. Τέλος, τα μετάγραφα που συγκροτούνται από την συρραφή του 1ου εξωνίου με ένα από τα εξώνια 6, 7 και 8 δεν διαθέτουν ORFs και επομένως, χαρακτηρίζονται ως μη κωδικά μόρια.

3.6. Μελέτη του προφίλ έκφρασης των νέων εναλλακτικών *CDK4* μεταγράφων

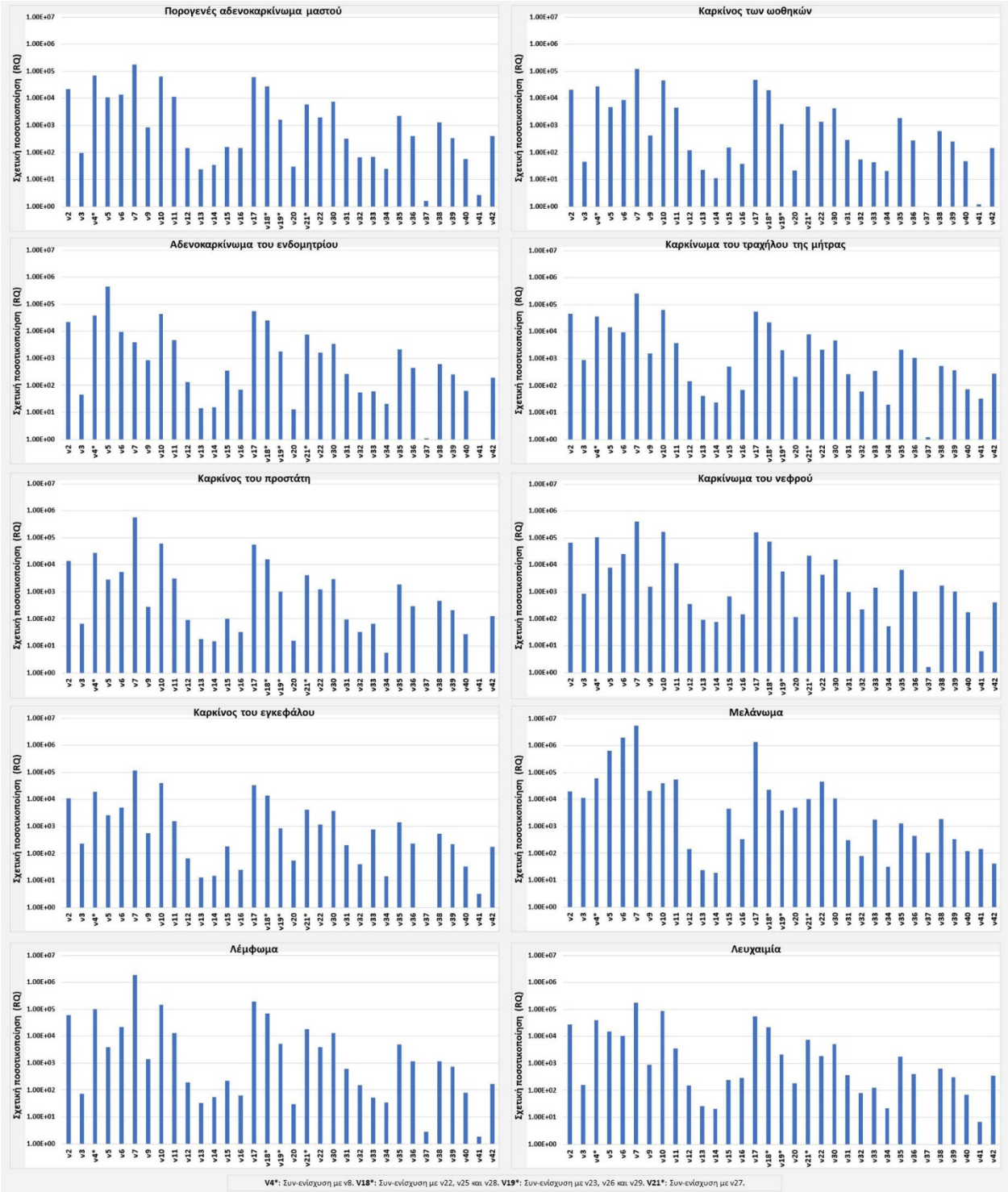
Με βάση τα αποτελέσματα που λήφθηκαν από τις αντιδράσεις qPCR, που πραγματοποιήθηκαν, τα νέα εναλλακτικά μετάγραφα του γονιδίου *CDK4,* τα οποία περιγράφονται στην παρούσα διπλωματική εργασία, παρουσιάζουν ένα ευρύ προφίλ έκφρασης, καθώς προσδιορίζονται στους περισσότερους από τους ανθρώπινους ιστούς στους οποίους έγινε η μελέτη. Αξιοσημείωτο είναι το γεγονός ότι τα συγκεκριμένα εναλλακτικά μετάγραφα δεν εκφράζονται μόνο στους καρκινικούς ιστούς, στους οποίους η έκφραση του κύριου *CDK4* μεταγράφου (*CDK4* v.1) είναι υψηλή, αλλά ανιχνεύθηκαν και στην κυτταρική σειρά HEK-293, η οποία προέρχεται από φυσιολογικό εμβρυονικό νεφρό (εικόνα 27).



Εικόνα 27. Τα διαγράμματα παρουσιάζουν τα σχετικά επίπεδα έκφρασης κάθε νέου εναλλακτικού *CDK4* μεταγράφου, τα οποία προσδιορίστηκαν με ποσοτική PCR στη φυσιολογική κυτταρική σειρά HEK-293 καθώς και σε 6 ανθρώπινες κακοήθειας, οι οποίες, σύμφωνα με την βιβλιογραφία, παρουσιάζουν υψηλά ποσοστά έκφρασης του *CDK4* γονιδίου. Τα επίπεδα έκφρασης κάθε μεταγράφου (ή ομάδας μεταγράφων) υπολογίστηκαν σε συνάρτηση με την αντίστοιχη έκφραση του mRNA του γονιδίου αναφοράς *GAPDH*. Η σχετική ποσότητα κάθε μεταγράφου παρουσιάζεται ως ο λόγος του αριθμού των αντιγράφων του συγκεκριμένου μεταγράφου προς 10^6 αντίγραφα του μεταγράφου *GAPDH* (άξονας Υ). Το σύμβολο * αντιπροσωπεύει τα μετάγραφα, στα οποία δεν ήταν δυνατός ο ειδικός ποσοτικός τους προσδιορισμός, μέσω της qPCR, λόγω των πολύ μικρών διαφορών που φέρουν οι αλληλουχίες τους και επομένως, για τα συγκεκριμένα μετάγραφα πραγματοποιήθηκε ποσοτικός προσδιορισμός της ομάδας που ενισχύθηκε.

Τα ευρήματα, που παρουσιάζονται ενισχύουν την υπόθεση ότι τα περισσότερα από τα νέα μετάγραφα, τα οποία προσδιορίστηκαν, διαθέτουν άγνωστες, αλλά, άκρως σημαντικές λειτουργίες, που, πιθανότατα, σχετίζονται με την κυτταρική ομοιόσταση. Επιπλέον, όλα τα εναλλακτικά μετάγραφα ανιχνεύθηκαν στους περισσότερους ιστούς, ωστόσο, τα επίπεδα έκφρασής τους είναι αρκετά διαφοροποιημένα. Συγκεκριμένα, σύμφωνα με την ποσοτική μελέτη που εφαρμόστηκε, το νέο *CDK4* μετάγραφο v.7 φαίνεται να υπερεκφράζεται στο σύνολο των 16 ανθρώπινων κακοηθειών, που μελετήθηκαν, ενώ τα μετάγραφα v.37 και v.41 ανιχνεύονται αρκετά δύσκολα στους περισσότερους ιστούς (εικόνες 27 & 28). Επιπροσθέτως, τα μετάγραφα *CDK4* v.2, v.10 και v.17 ακολουθούν ένα αρκετά υψηλό προφίλ έκφρασης στους περισσότερους τύπους καρκίνου, καθώς και στη φυσιολογική κυτταρική σειρά εμβρυονικού νεφρού.

Ωστόσο, εξαιτίας του μεγάλου πλήθους των νέων εναλλακτικών μεταγράφων, τα οποία ταυτοποιήθηκαν, και λόγω των μικροδιαφορών που φέρουν μεταξύ τους ορισμένα από αυτά, εξαιτίας των γεγονότων εναλλακτικής συρραφής που συμβαίνουν, σε ορισμένες περιπτώσεις, είναι αδύνατη η ειδική ενίσχυση ενός *CDK4* μεταγράφου με τη μέθοδο της qPCR. Κατά συνέπεια, δύο ζεύγη μεταγράφων: v.4 / v.8 και v.21 / v.27, ενισχύονται ταυτόχρονα με το ίδιο ζεύγος εκκινητών και συνεπώς τα επίπεδα έκφρασης τους δεν μπορούν να διακριθούν μέσω qPCR. Ο ίδιος περιορισμός ισχύει και για δύο σύνολα τεσσάρων μεταγράφων, τα οποία αφορούν στα *CDK4* v.18, v.22, v.25 και v.28 και τα *CDK4* v.19, v.23, v.26, v.29 (εικόνες 27 & 28). Τα τέσσερα μετάγραφα κάθε ομάδας ενισχύονται ταυτόχρονα και επομένως, το επίπεδο έκφρασης καθενός εκ των τεσσάρων μεταγράφων παραμένει ασαφές. Ο διαχωρισμός των συγκεκριμένων μεταγράφων θα μπορούσε να επιτευχθεί με την χρήση nested PCR. Ωστόσο η προσέγγιση αυτή δεν πραγματοποιήθηκε καθώς η μέθοδος εισάγει σφάλμα στον ποσοτικό προσδιορισμό του κάθε μεταγράφου.

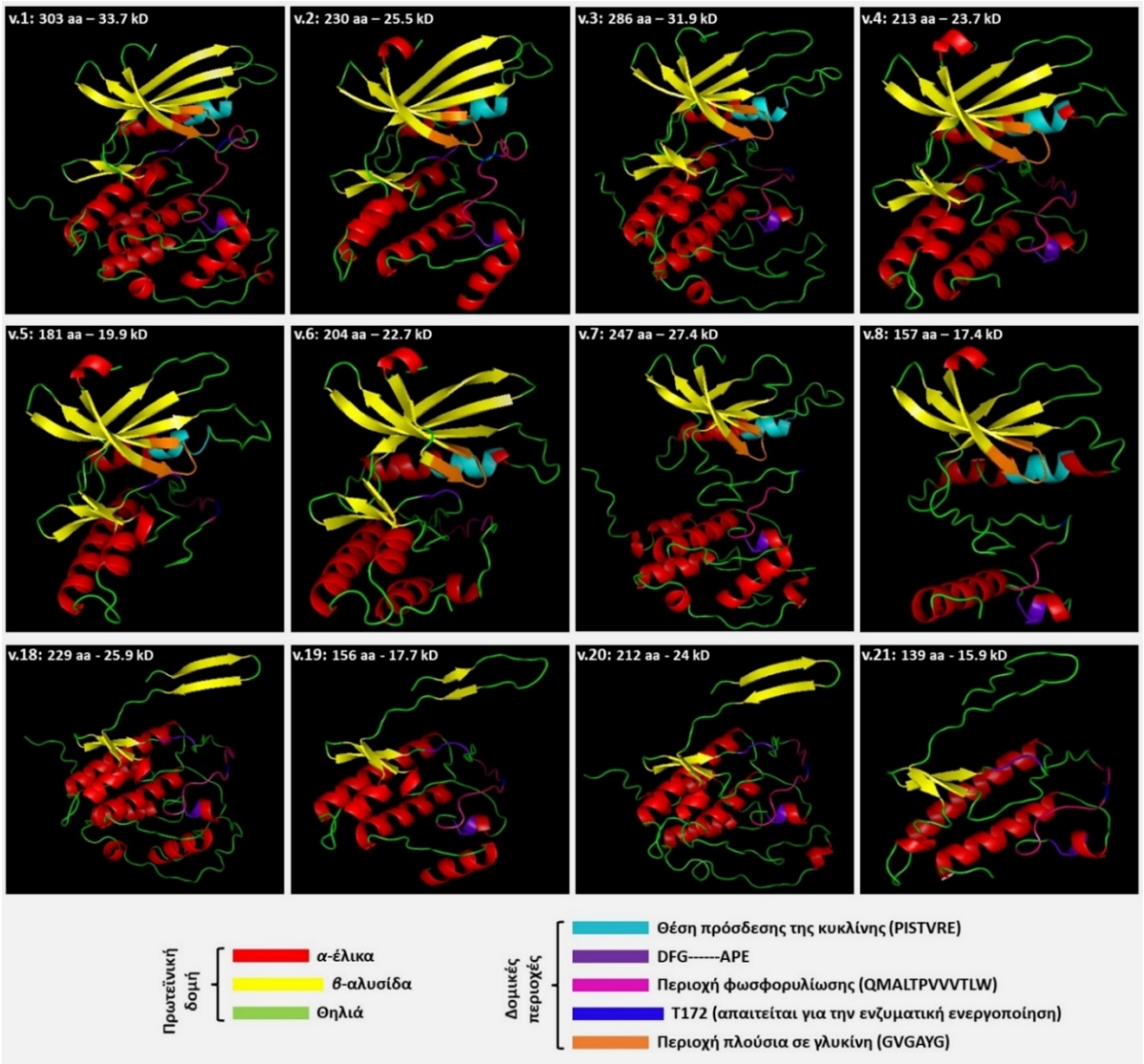


Εικόνα 28. Τα διαγράμματα παρουσιάζουν τα σχετικά επίπεδα έκφρασης κάθε νέου εναλλακτικού CDK4 μεταγράφου, τα οποία προσδιορίστηκαν με ποσοτική PCR σε 10 ανθρώπινες κακοήθειας. Τα επίπεδα έκφρασης κάθε μεταγράφου (ή ομάδας μεταγράφων) υπολογίστηκαν συναρτήσει της έκφρασης του *GAPDH* ως γονίδιο αναφοράς. Η σχετική ποσότητα κάθε μεταγράφου παρουσιάζεται ως ο λόγος του αριθμού των αντιγράφων του συγκεκριμένου μεταγράφου προς 106 αντίγραφα του μεταγράφου GAPDH (άξονας Y). Το σύμβολο * αντιπροσωπεύει τα μετάγραφα, στα οποία δεν ήταν δυνατός ο ειδικός ποσοτικός

προσδιορισμός, μέσω qPCR και επομένως, πραγματοποιήθηκε ποσοτικός προσδιορισμός της ομάδας μεταγράφων που ενισχύθηκε.

3.7. Προβλεπόμενα πρωτεϊνικά μοντέλα

Το εργαλείο I-TASSER επέτρεψε την παραγωγή τρισδιάστατων πρωτεϊνικών μοντέλων τα οποία μελετήθηκαν ως προς την παρουσία δομικών περιοχών που συγκροτούν μία ενεργή κινάση.



Εικόνα 29. Προβλεπόμενα μοντέλα τρισδιάστατης δομής των πρωτεινικών ισομορφών, τα οποία κωδικοποιούνται από εναλλακτικά μετάγραφα του *CDK4* γονιδίου. Κάθε συντηρημένη λειτουργική περιοχή παρουσιάζεται με διαφορετικό χρώμα. Για κάθε πρωτεϊνική ισομορφή απεικονίζεται μόνο η 3D δομή με την υψηλότερη βαθμολογία αξιοπιστίας σύμφωνα με το εργαλείο i-Tasser.

Οι προβλεπόμενες πρωτεϊνικές δομές των 7 μεταγράφων του *CDK4* (*CDK4* v.2 – v.8) έχουν δομή όμοια με την κλασική δομή του διπλού λοβού που παρατηρείται στην κύρια CDK4 πρωτεΐνη (εικόνα 29). Αντίθετα, τα πρωτεϊνικά μόρια, που προβλέπεται να παράγονται από την μεταφραστική ικανότητα των εναλλακτικών μεταγράφων του *CDK4*, στα οποία απουσιάζει το εξώνιο 2, (*CDK4* v.18-v. 21), παρουσιάζουν εμφανώς διαφοροποιημένη δομή με σχεδόν πλήρη απουσία της αμινοτελικής περιοχής (εικόνα 29).

4. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΣΥΖΗΤΗΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

Η φυσιολογική λειτουργία των κυττάρων, που απαρτίζουν τους ιστούς και επιτελούν άκρως σημαντικές διεργασίες για την ομαλή λειτουργία ενός οργανισμού, καθώς και η μελέτη των μηχανισμών, που διαταράσσουν την ομοιόσταση αυτή, αποτελούν ένα πολλά υποσχόμενο πεδίο έρευνας στον τομέα της Βιολογίας. Ο τρόπος με τον οποίο ένα κύτταρο επικοινωνεί και αλληλεπιδρά με το περιβάλλον του, καθώς και η ικανότητα ανταπόκρισής του σε αλλαγές που συμβαίνουν καθ' όλη τη διάρκεια της ζωής του καθορίζονται σε μεγάλο βαθμό από τη γενετική πληροφορία που διαθέτει και επομένως, το DNA αποτελεί τη βάση για κάθε επιστημονική μελέτη, που στόχο έχει να απαντήσει τέτοιου είδους βιολογικά ερωτήματα.

Ως εκ τούτου, το εναλλακτικό μάτισμα είναι ένας σημαντικός μηχανισμός που οδηγεί στην παραγωγή πολλαπλών μεταγράφων mRNA από ένα μόνο γονίδιο και χρησιμοποιείται για τη διατήρηση της κυτταρικής ομοιόστασης και ανάπτυξης, ρυθμίζοντας τη γονιδιακή έκφραση [61]. Ωστόσο, η απορρύθμιση των μηχανισμών, που αφορούν στην διαδικασία του εναλλακτικού ματίσματος, σχετίζεται με την εμφάνιση κακοηθειών στον άνθρωπο. Εναλλακτικά μετάγραφα, τα οποία ανιχνεύονται σε καρκινικούς ιστούς, έχει αποδειχθεί ότι συμμετέχουν στην αιτιοπαθογένεια και στην εξέλιξη ασθενειών. Επομένως, ο προσδιορισμός νέων εναλλακτικών μεταγράφων, που προέρχονται από ένα γονίδιο, μελετώντας τα γεγονότα εναλλακτικού ματίσματος και τους μηχανισμούς της μεταγραφής, είναι ένα ουσιαστικό βήμα για την κατανόηση της πολυπλοκότητας του μεταγραφώματος και, κατά συνέπεια, και του πρωτεώματος στα ευκαρυωτικά κύτταρα σε φυσιολογικές και παθολογικές καταστάσεις. Στη παρούσα διπλωματική εργασία, στόχος ήταν η μελέτη του γονιδίου *CDK4* σε καρκινικά κύτταρα και η διερεύνηση της ύπαρξης νέων εναλλακτικών μεταγράφων του γονιδίου με την χρήση των μεθοδολογιών αλληλούχησης επόμενης και τρίτης γενιάς. Τελικός σκοπός ήταν η διεξοδική μελέτη του γονιδίου *CDK4,* ώστε να επιτευχθεί η πλήρης κατανόηση του ρόλου του γονιδίου αυτού κατά τη διάρκεια του κυτταρικού κύκλου και ο τρόπος με τον οποίο εμπλέκεται στην καρκινογένεση σε διάφορους τύπους ιστών.

Οι μέθοδοι αλληλούχησης του γονιδιώματος, που έχουν αναπτυχθεί, αποτελούν αξιόπιστα εργαλεία για την μελέτη των βιολογικών μηχανισμών που διέπουν τα κύτταρα και κατά συνέπεια, τους οργανισμούς. Ιδιαίτερα, τις τελευταίες δεκαετίες, οι μεθοδολογίες προσδιορισμού νουκλεοτιδικών αλληλουχιών, όπως η NGS και TGS αλληλούχηση, έχουν αποδειχθεί πολύτιμα εργαλεία για τον χαρακτηρισμό του

μεταγραφώματος, τον προσδιορισμό γεγονότων εναλλακτικού ματίσματος και τον ποσοτικό προσδιορισμό των επιπέδων της μεταγραφής και συμβάλλουν στην κατανόηση των διαδικασιών ρύθμισης της γονιδιακής έκφρασης [124, 125]. Επιπλέον, τα, πολύ υψηλής απόδοσης, αποτελέσματα, που προσφέρουν οι, τελευταίας γενιάς, τεχνολογίες αλληλούχησης, έχουν οδηγήσει στην ενίσχυση της έρευνας, που αφορά στη μελέτη των μηχανισμών που προκαλούν ανθρώπινες ασθένειες και, ειδικότερα, καρκίνο. Η μαζική παράλληλη αλληλούχηση, τόσο με μεθόδους επόμενης γενιάς όσο και με την χρήση των νέων μεθοδολογιών αλληλούχησης τρίτης γενιάς, χρησιμοποιείται για με σκοπό τον προσδιορισμό ολόκληρων γονιδιωμάτων ή / και την μελέτη συγκεκριμένων τμημάτων του γονιδιώματος και μπορεί να ανιχνευθούν αλλαγές που συμβαίνουν κατά μήκος ολόκληρου του γονιδιώματος στις οποίες συμπεριλαμβάνονται αντικαταστάσεις απαλοιφές ή προσθήκες βάσεων.

Ειδικότερα, οι μέθοδοι αλληλούχησης NGS και TGS αποτελούν ισχυρά εργαλεία με τα οποία μπορεί να μελετηθεί διεξοδικά όλο το γονιδίωμα και συμβάλλουν στην αποκάλυψη νέων πληροφοριών για την αρχιτεκτονική με την οποία συγκροτούνται τα γονίδια, καθώς δίνουν τη δυνατότητα ανίχνευσης και ταυτοποίησης, με μεγάλη ακρίβεια, πολλαπλών μεταγράφων, τα οποία προέρχονται από ένα γονίδιο [74, 126]. Στην παρούσα εργασία, χρησιμοποιήθηκε η πλατφόρμα νέας γενιάς αλληλούχησης PGM της Ion Torrent™ καθώς και ο αλληλουχητής τρίτης γενιάς MinION™ Mk1C (Oxford Nanopore Technologies Ltd, ONT) για την ανίχνευση γεγονότων εναλλακτικής συρραφής μεταξύ των εξωνίων του γονιδίου *CDK4* και την ανακάλυψη νέων μεταγράφων. Η μελέτη του γονιδίου βασίζεται στο mRNA που προκύπτει κατά τη μεταγραφή και στα γεγονότα ματίσματος που συμβαίνουν κατά τη διάρκεια της επεξεργασίας RNA. Ωστόσο, η ταυτοποίηση των νέων μεταγραφών επιτεύχθηκε με την εφαρμογή πρωτοκόλλων, που βασίζονται στο προσδιορισμό του DNA, στοχεύοντας στη λήψη όσο το δυνατόν περισσότερων δεδομένων για ανάλυση, καθώς το πλήθος των νουκλεοτιδικών αλυσίδων, που πρόκειται να προσδιοριστούν, είναι διπλάσιο όταν ο προς αλληλούχηση στόχος είναι δίκλωνο DNA.

Στην παρούσα διπλωματική εργασία, αναπτύχθηκε μια τρίτης γενιάς μεθοδολογία αλληλούχησης, βασισμένη στον προσδιορισμό του DNA, χρησιμοποιώντας την συσκευή MinION™ Mk1C, με στόχο την ανίχνευση νέων πλήρους μήκους εναλλακτικών μεταγράφων του ανθρώπινου γονιδίου *CDK4*. Η βιοπληροφορική ανάλυση των δεδομένων οδήγησε στον εντοπισμό 41 νέων εναλλακτικών *CDK4* μεταγράφων, τα

οποία προέρχονται από πολλαπλούς συνδυασμούς εναλλακτικής συρραφής μεταξύ των 8 γνωστών εξωνίων του γονιδίου. Συνολικά προσδιορίστηκαν 18 νέα γεγονότα συρραφής μεταξύ των εξωνίων, τα οποία επιβεβαιώθηκαν και από την ανάλυση των αποτελεσμάτων της NGS αλληλούχησης μέσω Ion Torrent™. Οι νέες θέσεις συρραφής προκύπτουν από τον κυρίαρχο μηχανισμό εναλλακτικού ματίσματος, ο οποίος αφορά στην παράλειψη ενός ή / και περισσοτέρων εξωνίων [93], ενώ δεν πραγματοποιήθηκε μελέτη για τον εντοπισμό άλλων μεταγράφων που προέρχονται από άλλους μηχανσιμούς εναλλακτικής συρραφής, όπως διατήρηση εσωνίου.

Επιπλέον, μελετήθηκαν γνωστές μικρές «υπο-αλληλουχίες» της cDNA αλληλουχίας, οι οποίες ονομάζονται Expressed Sequence Tags (ESTs) και είναι διαθέσιμες στη GenBank®, καθώς και δημόσια διαθέσιμα στη βάση δεδομένων Sequence Read Archive (SRA) αποτελέσματα TGS, που προέκυψαν από πειράματα RNA-seq, προκειμένου να διερευνηθεί η πιθανή ύπαρξη κάθε νέου *CDK4* μεταγράφου, που περιγράφεται στην παρούσα μελέτη. Τα TGS αποτελέσματα αφορούν στην αλληλούχηση του μεταγραφώματος που εκφράζεται στις κυτταρικές σειρές HepG2 and U87 (SRA πειραματικές αλληλουχίες SRR11861906 και SRR11262667). Η *in-silico* ανάλυση επιβεβαίωσε ότι κανένα από τα νέα mRNA μόρια δεν ανιχνεύτηκε σε ESTs ή στις πειραματικές αλληλουχίες της RNA-seq αλληλούχησης, γεγονός που ενισχύει την εφαρμογή πρωτοκόλλων που αφορούν αλληλούχηση DNA για την μελέτη των γεγονότων εναλλακτικής συρραφής.

Η εφαρμογή πρωτοκόλλων που βασίζονται στο DNA σε σύγκριση με τον άμεσο προσδιορισμό της αλληλουχίας του RNA επιτρέπει, κατά την TGS αλληλούχηση, την ανίχνευση σπάνιων και πλήρους μήκους μεταγραφών mRNA. Αντίθετα, η απευθείας αλληλούχηση RNA μορίων χαρακτηρίζεται από μειωμένη παρουσία σπάνιων μεταγραφών, που συνήθως προκύπτουν από το εναλλακτικό μάτισμα των πρώιμων mRNAs, τα οποία ενδέχεται να μην προσδιοριστούν κατά την αντίδραση αλληλούχησης αν το βάθος ανάγνωσης είναι μικρό. Ως αποτέλεσμα, η μέθοδος μελέτης του δίκλωνου DNA για τον προσδιορισμό εναλλακτικών μεταγράφων επιτρέπει αυξημένα, ακριβή και αξιόπιστα αποτελέσματα. Επιπλέον, η κατασκευή DNA βιβλιοθήκης στο στάδιο προετοιμασίας της TGS βιβλιοθήκης, που προηγείται της αντίδρασης αλληλούχησης, καθώς και η ανάλυση του DNA, αποτελείται από καθορισμένα βήματα, τα οποία πραγματοποιούνται σε μειωμένο χρόνο, ενώ από την άλλη πλευρά, η προετοιμασία RNA βιβλιοθήκης είναι μια χρονοβόρα διαδικασία λόγω των πολλαπλών βημάτων που

απαιτούνται. Κατά συνέπεια, ο προσδιορισμός του δίκλωνου DNA προσφέρει ιδιαίτερα αυξημένο βάθος ανάγνωσης, υψηλή κάλυψη και πολύ υψηλή εξειδίκευση για τη μελέτη των μεταγράφων του *CDK4* στο συντομότερο δυνατό χρόνο.

Τη δεκαετία του 2000, η ανάπτυξη της μαζικής παράλληλης αλληλούχησης με την ανάπτυξης των μεθοδολογιών αλληλούχησης επόμενης γενιάς αποτέλεσε καινοτομία στον τομέα της γονιδιωματικής [82]. Η NGS αλληλούχηση είναι ένα χρήσιμο εργαλείο για την μελέτη του γονιδιώματος και, τα τελευταία χρόνια, αποκτά μεγάλο αντίκτυπο στη διάγνωση, την πρόγνωση, καθώς και τη θεραπεία πολλών ασθενειών συμπεριλαμβανομένου του καρκίνου [127]. Πλέον, η νέα καινοτόμος προσέγγιση προσδιορισμού του γονιδιώματος μέσω της TGS αλληλούχησης θεωρείται ότι θα αντικαταστήσει σε μεγάλο βαθμό τις παλαιότερες μεθοδολογίες, καθώς προσφέρει πλήθος πλεονεκτημάτων τα οποία ελαχιστοποιούν το χρόνο της αντίδρασης και παράλληλα αυξάνουν την απόδοση. Σύμφωνα με τα αποτελέσματα της παρούσας μελέτης και συγκρίνοντας τις δύο τεχνικές αλληλούχησης, η τεχνολογία αλληλούχησης τρίτης γενιάς με την χρήση της συσκευής MinION™ Mk1C (ONT) είναι δυνατόν να αντιμετωπίσει σημαντικές προκλήσεις τις οποίες αδυνατεί η αλληλούχηση επόμενης γενιάς.

Τα αποτελέσματα της συγκεκριμένης μελέτης, που λήφθηκαν από τις δύο μεθόδους αλληλούχησης, επέτρεψαν την περαιτέρω σύγκριση μεταξύ των μεθοδολογιών NGS και TGS. Ειδικότερα, η TGS αλληλούχηση παρέχει πειραματικές αλληλουχίες μεγάλου μήκους (> 500 bp) αντίθετα με τις νουκλεοτιδικές αλληλουχίες, οι οποίες προσδιορίζονται μέσω των τεχνολογιών, που προσφέρουν οι πλατφόρμες Ion Torrent™ και Illumina®, που δεν ξεπερνούν τις 600 bp σε μήκος. Το πλεονέκτημα αυτό οδηγεί στο συμπέρασμα ότι η αλληλούχηση τρίτης γενιάς αποτελεί την πιο κατάλληλη μέθοδο για την μελέτη των εναλλακτικών μεταγράφων ενός γονιδίου, καθώς είναι δυνατόν να προσδιοριστεί ακέραιη η πλήρης αλληλουχία ενός μεταγράφου χωρίς να έχει προηγηθεί θραύση του στόχου σε μικρότερα τμήματα και ως επακόλουθο, παραλείπεται, στη συνέχεια, το στάδιο της συναρμολόγησης των πειραματικών αλληλουχιών κατά την ανάλυση των αποτελεσμάτων. Συγκεκριμένα, η παρούσα μελέτη αποδεικνύει την ικανότητα ανίχνευσης και ταυτοποίησης, με μεγάλη αξιοπιστία, νέων εναλλακτικών μεταγράφων μέσω της αλληλούχησης τρίτης γενιάς.

Ένα ακόμα πλεονέκτημα των αντιδράσεων προσδιορισμού του DNA, μέσω των πλατφόρμων αλληλούχησης της ONT αποτελεί η ταχύτητα με την οποία ολοκληρώνεται

ο προσδιορισμός κάθε νουκλεοτιδικής αλυσίδας. Η αλληλούχηση τρίτης γενιάς προσφέρει ανάλυση σε πραγματικό χρόνο και τα αποτελέσματα μπορούν να ληφθούν σε λίγα μόλις λεπτά ή ώρες. Από την άλλη πλευρά, το πρωτόκολλο προετοιμασίας μιας NGS βιβλιοθήκης, αλλά και το στάδιο της αντίδρασης αλληλούχησης μπορεί να διαρκέσει αρκετές ώρες έως και ημέρες. Η αλληλούχηση επόμενης γενιάς απαιτεί την προετοιμασία και τον εμπλουτισμό του δείγματος, το οποίο θα προσδιοριστεί στη συνέχεια. Αντίθετα, το βήμα αυτό παραλείπεται στην αλληλούχηση τρίτης γενιάς με άμεσο αποτέλεσμα τη μείωση του συνολικού χρόνου που απαιτείται για την διεξαγωγή ενός πειράματος αλληλούχησης και παράλληλα, μειώνεται ο αριθμός, και συνεπώς το κόστος, των αντιδραστηρίων που χρησιμοποιούνται. Επιπρόσθετα, το στάδιο της προετοιμασίας του εκμαγείου απαιτεί επιπλέον αντιδράσεις, κατά τις οποίες ενδέχεται να προκύψουν πειραματικά σφάλματα, όπως, για παράδειγμα, κατά την emPCR μπορεί να ενισχυθούν συγκεκριμένα τμήματα DNA, ενώ άλλα τμήματα να μην πολλαπλασιαστούν με αποτέλεσμα την απώλεια πληροφορίας κατά την αλληλούχηση.

Οι πειραματικές αλληλουχίες, που λαμβάνονται από πειράματα αλληλούχησης τρίτης γενιάς, είναι μεγαλύτερου μήκους και δεν απαιτείται η συναρμολόγηση των θραυσμάτων του DNA, γεγονός που προσφέρει πολλές ευκαιρίες για διερεύνηση μεγάλων περιοχών DNA των οποίων η ανίχνευση ήταν δύσκολη με τη χρήση προηγούμενων προσεγγίσεων, όπως το NGS. Ωστόσο, ο αριθμός των διαθέσιμων εργαλείων για την βιοπληροφορική ανάλυση των δεδομένων της αλληλούχησης τρίτης γενιάς είναι ακόμη περιορισμένος. Εν κατακλείδι, χάρη στη χρήση των δύο τεχνολογιών αλληλούχησης επιτεύχθηκε ο τελικός στόχος της παρούσας διπλωματικής εργασίας, ο οποίος ήταν η κατανόηση των γεγονότων εναλλακτικού ματίσματος στο γονίδιο *CDK4,* τα οποία ενδέχεται να αποκαλύψουν τους τρόπους εμπλοκής του γονιδίου σε διαδικασίες που οδηγούν στην ογκογένεση.

Η παρουσία των εναλλακτικών μεταγράφων που ανιχνεύθηκαν για το γονίδιο *CDK4* δεν αποτελεί ξεχωριστό και μοναδικό χαρακτηριστικό του γονιδίου αλλά γενικό χαρακτηριστικό της οικογένειας των *CDK* μορίων. Τα περισσότερα μέλη της οικογένειας των *CDKs* γονιδίων υπόκεινται σε εναλλακτικό μάτισμα και, επομένως, είναι ικανά παράγουν πολλαπλά αντίγραφα από ένα μόνο γονίδιο. Χαρακτηριστικό είναι το παράδειγμα του γονιδίου *CDK7,* το οποίο σύμφωνα με τα διαθέσιμα δεδομένα της GenBank®, παρουσιάζει 10 εναλλακτικά μετάγραφα. Ωστόσο, ο αριθμός των μεταγράφων του γονιδίου *CDK4,* τα οποία ταυτοποιήθηκαν, είναι τετραπλάσιος παρόλο

που το γονίδιο *CDK7* διαθέτει 13 γνωστά εξώνια ενώ το *CDK4* μόλις 8. Ως εκ τούτου, μπορούν να δημιουργηθούν τα εξής ερωτήματα:

- Υπάρχει πιθανότητα να παράγονται επιπλέον εναλλακτικά μετάγραφα του γονιδίου *CDK7*; Μπορεί τα γεγονότα εναλλακτικού ματίσματος να συμβαίνουν σε όλα τα μέλη της οικογένειας των *CDKs*;
- Υπάρχει συσχετισμός του αριθμού γεγονότων εναλλακτικού ματίσματος με τον αριθμό των μεταγράφων που μπορεί να παράγεται από κάθε γονίδιο;
- Πως το κύτταρο επιλέγει ποιο γεγονός συρραφής θα συμβεί κατά την ωρίμανση κάθε παραγόμενου πρώιμου mRNA; Για ποιο λόγο ορισμένα γεγονότα εναλλακτικού ματίσματος εμφανίζονται περισσότερες φορές από άλλα;

Για την απάντηση των ερωτημάτων, που μόλις αναφέρθηκαν, απαιτείται η οργάνωση και διενέργεια περαιτέρω πειραματικών διαδικασιών, καθώς τα αποτελέσματα, που λήφθηκαν στα πλαίσια της παρούσας διπλωματικής εργασίας, αδυνατούν να δώσουν σαφή απάντηση. Ωστόσο, με βάση την παρούσα μελέτη και βιβλιογραφικές αναφορές που μελετούν τους μηχανισμούς εναλλακτικού ματίσματος, είναι δυνατόν να δοθούν ορισμένες εξηγήσεις, οι οποίες πιθανώς να αφορούν στα συγκεκριμένα ερωτήματα. Αρχικά, η παρουσία 41 εναλλακτικών *CDK4* μεταγράφων και η ύπαρξη εναλλακτικών μεταγράφων στο γονίδιο *CDK7* υποδηλώνουν πως γεγονότα εναλλακτικού ματίσματος είναι δυνατόν να συμβαίνουν σε όλους τους αντιπροσώπους της οικογένειας των *CDKs* και, ιδιαίτερα, στα γονίδια που έχουν άμεση σχέση με λειτουργίες ρύθμισης του κυτταρικού κύκλου, όπως τα *CDK1, CDK2* και *CDK6*.

Ακολούθως, το μέγεθος κάθε γονιδίου χαρακτηρίζει το πλήθος των γεγονότων εναλλακτικής συρραφής που μπορεί να συμβούν. Όπως γίνεται αντιληπτό, ένα γονίδιο, που φέρει, για παράδειγμα, μόλις 4 εξώνια, έχει τη δυνατότητα σχηματισμού λιγότερων εναλλακτικών θέσεων συρραφής από ένα γονίδιο με 10 εξώνια. Επομένως, μπορεί να θεωρηθεί ότι ένα γονίδιο με περισσότερα εξώνια διαθέτει περισσότερες θέσεις εναλλακτικής συρραφής, αν και αυτό δεν είναι απόλυτο, καθώς οι μηχανισμοί εναλλακτικού ματίσματος ενδέχεται να μην λαμβάνουν χώρα σε όλα τα γονίδια. Κατά την διαδικασία ωρίμανσης ενός πρώιμου mRNA μπορούν να συμβούν διαφορετικά γεγονότα ματίσματος που επηρεάζουν το παραγόμενο mRNA που θα προκύψει. Δεδομένου ότι τα φυσιολογικά κύτταρα των ανώτερων ευκαρυωτικών οργανισμών διαθέτουν συστήματα πλήρους ελέγχου των διαδικασιών που λαμβάνουν χώρα κάθε χρονική στιγμή, ώστε να διατηρείται η ομοιόστασή τους, συμπεραίνουμε ότι το κύτταρο

διαθέτει ειδικούς μηχανισμούς που ρυθμίζουν τον τρόπο συρραφής των εξωνίων και επομένως, δεν εμφανίζονται όλα τα γεγονότα συρραφής που συμβαίνουν σε ένα mRNA με την ίδια συχνότητα. Εάν για παράδειγμα, σε μία συγκεκριμένη χρονική στιγμή υπάρχει ανάγκη ρύθμισης του κυτταρικού κύκλου, η διαδικασία ωρίμανσης των πρόδρομων mRNA μορίων του *CDK4* γονιδίου θα ρυθμιστεί κατάλληλα ώστε να δράσουν μηχανισμοί, που θα οδηγήσουν στην παραγωγή του κύριου *CDK4* v.1 μεταγράφου, το οποίο θα επάγει την σύνθεση της CDK4. Αντίθετα, αν δεν υπάρχει απαίτηση ρύθμισης του κυτταρικού πολλαπλασιασμού, τα πρόδρομα *CDK4* μετάγραφα, που έχουν παραχθεί, πιθανότατα θα οδηγούνται σε μηχανισμούς εναλλακτικής συρραφής, ώστε να παραχθούν εναλλακτικά κωδικά ή μη κωδικά ώριμα μετάγραφα, τα οποία διαθέτουν διαφορετικό λειτουργικό ρόλο ή τα οποία είναι πολύ ασταθή μόρια και οδηγούνται προς αποικοδόμηση.

Στη περίπτωση των γεγονότων εναλλακτικού ματίσματος του *CDK4* μεταγράφου, η μελέτη για τον προσδιορισμό των θέσεων εναλλακτικής συρραφής, που ανιχνεύθηκαν, χρησιμοποιήθηκε δείγμα από καρκινικές κυτταρικές σειρές που καλλιεργήθηκαν στο εργαστήριο. Επομένως, σημαντική κρίνεται η προσπάθεια εντοπισμού των γεγονότων αυτών σε φυσιολογικές κυτταρικές σειρές. Ο εντοπισμός τυχών διαφορών μεταξύ των γεγονότων συρραφής, που συμβαίνουν σε κύτταρα, που προέρχονται από φυσιολογικούς και καρκινικούς ιστούς, μπορεί να υποδείξει τη συσχέτιση τους με την εκδήλωση κακοηθειών ή την προδιάθεση για εμφάνιση καρκίνου; Επιπλέον, με δεδομένο ότι το δείγμα που χρησιμοποιήθηκε προέρχεται από κυτταρικές σειρές προκύπτει το ερώτημα αν τα συμβάντα εναλλακτικού ματίσματος, που εντοπίστηκαν στις καρκινικές αυτές κυτταρικές σειρές που αναφέρθηκαν, μπορούν να εντοπιστούν και σε ιστούς ασθενών και αν ναι, σε ποιο στάδιο ανάπτυξης καρκίνου. Υπάρχει, επιπλέον, συσχέτιση των μεταγράφων αυτών με την κλινική εικόνα των ασθενών; Θα μπορούσε κάποιο ή / και κάποια από τα *CDK4* μετάγραφα να χαρακτηριστεί ως βιοδείκτης; Η παρουσία ή η απουσία των εναλλακτικών αυτών μεταγράφων σχετίζεται με την πρόγνωση;

Επιπλέον, κρίνεται απαραίτητο να τονιστεί ότι στην παρούσα μελέτη προσδιορίστηκαν μόνο τα εναλλακτικά μετάγραφα τα οποία προκύπτουν ύστερα από την παράλειψη ενός ή περισσοτέρων εξωνίων και όχι το σύνολο των μηχανισμών εναλλακτικού ματίσματος. Επομένως, ένα επόμενο ερώτημα είναι η πιθανότητα ύπαρξης ακόμη περισσότερων μεταγράφων που παράγονται από το *CDK4.* Όπως

γίνεται αντιληπτό, τα ερωτήματα αυτά χρήζουν περαιτέρω διερεύνησης και δεν μπορούν να απαντηθούν με βάση τα αποτελέσματα της παρούσας διπλωματικής εργασίας.

Στην παρούσα μελέτη, η ανίχνευση των εναλλακτικών θέσεων συρραφής μεταξύ των εξωνίων του *CDK4* γονιδίου και ο προσδιορισμός των νέων εναλλακτικών μεταγράφων οδηγεί στη διερεύνηση του ρόλου και της τύχης των *CDK4* μεταγράφων στα κύτταρα. Τα νέα προσδιοριζόμενα εναλλακτικά μετάγραφα του *CDK4* είτε μπορεί να είναι κωδικά mRNA μόρια και επομένως, να διαθέτουν την ικανότητα να κωδικοποιούν νέες πρωτεϊνικές ισομορφές με διαφορετικά βιολογικά και λειτουργικά χαρακτηριστικά, είτε αντιπροσωπεύουν, λόγω της ύπαρξης PTCs, μη κωδικά RNAs, τα οποία πιθανόν έχουν ρυθμιστικό ρόλο. Για να επιτευχθεί ο διαχωρισμός των *CDK4* μεταγράφων σε υποομάδες, με σκοπό την διευκόλυνση στη μελέτη τους, κάθε ακολουθία εξετάστηκε ξεχωριστά ως προς το μήκος του πλαισίου ανάγνωσης και την ύπαρξη πρώιμων κωδικονίων λήξης.

Σύμφωνα με βιβλιογραφικά δεδομένα, το κωδικόνιο λήξης σε ένα κωδικό mRNA βρίσκεται περίπου 50 νουκλεοτίδια ανάντη του τελευταίου γεγονότος συρραφής και σε οποιαδήποτε θέση κατάντη αυτού του ορίου, χαρακτηριστικό το οποίο προσδίδει σταθερότητα στο μόριο [128, 129]. Αντίθετα, όταν το κωδικόνιο λήξης βρίσκεται σε περισσότερο απομακρυσμένες θέσεις χαρακτηρίζεται ως πρώιμο (premature termination codon) και το παραγόμενο RNA μόριο είναι πιο ασταθές και αποσυντίθεται εύκολα [130]. Σύμφωνα με αυτά τα κριτήρια, τα *CDK4* μετάγραφα κατηγοριοποιούνται σε 3 ομάδες: τα mRNAs που έχουν ORFs και, επομένως, έχουν την ικανότητα να κωδικοποιούν πρωτεΐνες, τα mRNAs που διαθέτουν smORFs, και η ικανότητα κωδικοποίησης πρωτεϊνών κρίνεται αμφίβολη, και τα μη κωδικά μόρια RNAs με PTCs (εικόνες 23 & 25). Ωστόσο, το σημαντικότερο ερώτημα που τίθεται σχετίζεται με την μεταφραστική ικανότητα των νέων κωδικών mRNAs. Πως επηρεάζει τα κύτταρα η παρουσία των νέων προβλεπόμενων CDK4 πρωτεϊνικών ισομορφών; Μπορεί ένα τέτοιο μόριο να συνδεθεί με κυκλίνη και να δράσει ως ενεργή κινάση αντικαθιστώντας ή ενισχύοντας την λειτουργία της κύριας πρωτεΐνης ή μήπως, τα μόρια αυτά αποτελούν μη λειτουργικά πεπτίδια, τα οποία διαθέτουν διαφορετικούς ρόλους πέρα από την ρύθμιση του κυτταρικού κύκλου.

Το κύριο *CDK4* μετάγραφο (*CDK4* v.1) συγκροτείται από τη σύνδεση των 8 γνωστών εξωνίων και κωδικοποιεί την εξαρτώμενη από κυκλίνη κινάση 4, μία πρωτεΐνη 303 αμινοξέων, η οποία χαρακτηρίζεται από την παρουσία όλων των βασικών δομών

και των διατηρημένων περιοχών μιας τυπικής ευκαρυωτικής κινάσης [131]. Το κωδικόνιο “ATG” που σηματοδοτεί την έναρξη της πρωτεϊνοσύνθεσης βρίσκεται στο 2ο εξώνιο και επομένως, αναμφίβολα, όλα τα εναλλακτικά *CDK4* μετάγραφα, που περιέχουν το εξώνιο 2 (*CDK4* v.2-v.17), μοιράζονται το ίδιο κωδικόνιο έναρξης και παρουσιάζουν υψηλή πιθανότητα να κωδικοποιούν πρωτεϊνικές ισομορφές παρόμοιες με το γνωστό πρωτεινικό μόριο. Ωστόσο, τα νέα *CDK4* v.2, v.3 και v.4 μετάγραφα μπορούν να χαρακτηριστούν ως τα περισσότερο υποσχόμενα για την παραγωγή πρωτεϊνών παρόμοιων της κύριας CDK4. Οι νουκλεοτιδικές αλληλουχίες των τριών αυτών μεταγράφων έχουν ORFs, τα οποία οδηγούν στην πρόβλεψη πρωτεϊνικών ισομορφών, οι οποίες διατηρούν όλες τις κρίσιμες περιοχές της κινάσης και επομένως, πληρούν όλα τα απαιτούμενα κριτήρια για την κωδικοποίηση πλήρως λειτουργικών μορίων (εικόνα 24).

Με βάση το ανοιχτό πλαίσιο ανάγνωσης, που αντιστοιχεί σε καθένα από τα μετάγραφα 2, 3 και 4, οι προβλεπόμενες CDK4 πρωτεϊνικές ισομορφές μοιράζονται την ίδια αμινοτελική περιοχή με την κύρια κινάση 4, που αποτελείται από τα αμινοξέα 1-96 (εικόνα 24). Οι συγκεκριμένες ισομορφές παρουσιάζουν την καταλυτική δράση της κινάσης, καθώς διαθέτουν τον πλούσιο σε γλυκίνη βρόγχο, ο οποίος είναι υπεύθυνος για την καθοδήγηση του ATP στη θέση της φωσφορυλίωσης. Επιπλέον, η παρουσία του συντηρημένου μοτίβου “PISTVRE”, επιτρέπει τη δέσμευση της κυκλίνης- D1 σε αυτές τις ισομορφές και επομένως, στο σχηματισμό του συμπλόκου της κινάσης με την κυκλίνη που αποτελεί το πρώτο βήμα για την ενεργοποίηση του ενζύμου [102, 108]. Η υπόθεση της ύπαρξης πρωτεϊνικών ισομορφών που προκύπτουν από την μετάφραση των κωδικών μεταγράφων 2, 3 και 4 του γονιδίου *CDK4* ενισχύεται περαιτέρω από τα παραγόμενα τρισδιάστατα μοντέλα πρόβλεψης της δομής της κάθε ισομορφής, τα οποία παρουσιάζονται στην εικόνα 29. Τα μοντέλα, που δημιουργήθηκαν με τη χρήση του εργαλείου i-Tasser και μελετήθηκαν με το πρόγραμμα PyMOL, φανερώνουν πως κάθε μία από τις τρεις ισομορφές της CDK4 πρωτεΐνης εμφανίζει σαφώς σημαντική δομική ομοιότητα με την κύρια κινάση 4 που παράγεται φυσιολογικά. Στην εικόνα 29, είναι εμφανής η διάκριση των δομικών περιοχών κάθε προβλεπόμενης πρωτεΐνης και επιπλέον, παρατηρείται η τυπική δομή του διπλού λοβού που χαρακτηρίζει τόσο την CDK4 αλλά και άλλες κινάσες.

Συνολικά, σύμφωνα με όσα αναφέρθηκαν προηγουμένως, συμπεραίνουμε ότι οι συγκεκριμένες CDK4 ισομορφές διαθέτουν τον ίδιο βαθμό συγγένειας δέσμευσης της

κυκλίνη με την κύρια παραγόμενη πρωτεΐνη. Υπό αυτές τις συνθήκες, οι προβλεπόμενες πρωτεΐνες δημιουργούν ισχυρούς δεσμούς με το μόριο της κυκλίνης με αποτέλεσμα την ενεργοποίηση του συμπλόκου κινάσης-κυκλίνης. Παράλληλα, η παρουσία των σηματοδοτικών αλληλουχιών, που βρίσκονται στην αμινοτελική περιοχή, προετοιμάζουν το μόριο για την ενζυματική ενεργοποίηση του καρβοξυτελικού τμήματος. Επιπρόσθετα, θα πρέπει να αναφερθεί ότι, για κάθε μία από τις πρωτεϊνικές ισομορφές 2, 3 και 4, στο καρβοξυτελικό άκρο βρίσκονται όλες οι κρίσιμες περιοχές για την λειτουργία της κινάσης (εικόνα 24 & 26). Συγκεκριμένα, όπως παρουσιάζεται με μωβ χρώμα στην εικόνα 29, στην καρβοξυτελική περιοχή υπάρχει το δομικό μοτίβο DFG-APE το οποίο ρυθμίζει την δραστικότητα της κινάσης με τη διαμόρφωση τη θηλιά της φωσφορυλίωσης, που περιλαμβάνει το αμινοξύ T172, και λειτουργεί ως το ενεργό κέντρο του ενζύμου.

Τα δομικά στοιχεία, που χαρακτηρίζουν τις τρεις πρωτεΐνες, οι οποίες κωδικοποιούνται από τα μετάγραφα *CDK4* 2, 3 και 4, τα οποία περιγράφηκαν, οδηγούν στο συμπέρασμα ότι μικρές παραλλαγές διαφοροποιούν τα μόρια αυτά με την κύρια κινάση. Η ταυτόχρονη παρουσία των 4 αυτών, σχεδόν πανομοιότυπων πρωτεϊνών, στα κύτταρα και η ενεργοποίηση τους μπορεί να λειτουργήσει ως ένα σήμα για την υπερ-ενεργοποίηση και τον πολλαπλασιασμό των κυττάρων, γεγονός που ενδέχεται να οδηγήσει σε καρκινογένεση ή την συνεχή εξέλιξη των καρκινικών ιστών και την δημιουργία μεταστάσεων. Συγκεκριμένα, η αυξημένη δράση των πρωτεϊνικών κινασών που ρυθμίζουν τον κυτταρικό κύκλο έχει ως αποτέλεσμα την υπερφωσφορυλίωση των πρωτεϊνών του ρετινοβλαστώματος και την επαγωγή του κυτταρικού πολλαπλασιασμού λόγω της ενεργοποίησης του μεταγραφικού παράγοντα E2F (εικόνα 10). Όσο περισσότερα ενεργοποιημένα μόρια δίνουν το σήμα φωσφορυλίωσης των πρωτεϊνών, τόσο ενισχύεται η γονιδιακή έκφραση γεγονός που μπορεί να οδηγήσει σε ανεξέλεγκτο πολλαπλασιασμό των κυττάρων και ως εκ τούτου καρκινογένεση.

Από την άλλη πλευρά, η παρουσία των πρωτεϊνικών ισομορφών στα κύτταρα, ενδέχεται να αναπληρώνει τη δράση της κύριας CDK4 σε περιπτώσεις όπου, η δράση του συγκεκριμένου ενζύμου είναι μειωμένη ή μηδενική. Μια περίπτωση θα μπορούσε να είναι η παρουσία μετάλλαξης σε κάποια περιοχή του γονιδιώματος που δεν επιτρέπει την έκφραση του κύριου μεταγράφου αλλά μόνο εναλλακτικών mRNA μορίων. Επομένως, τα εναλλακτικά μετάγραφα, καθώς και η κωδικοποιητική τους ικανότητας, μπορεί να είναι κρίσιμα για την αποκατάσταση της φυσιολογικής λειτουργίας των

κυττάρων. Συμπερασματικά, ο λειτουργικός ρόλος των μεταγράφων αυτών καθώς και η ανίχνευση των προβλεπόμενων πρωτεϊνών αξίζει περαιτέρω διερεύνησης, καθώς δύναται να απαντηθούν καίρια ερωτήματα για τους μηχανισμούς που διέπουν τα κύτταρα.

Ένα επόμενο ζήτημα είναι οι λειτουργικές ιδιότητες των υπόλοιπων mRNAs του *CDK4,* που χαρακτηρίστηκαν ως κωδικά. Για παράδειγμα, τα μετάγραφα 5 και 6 έχουν ORFs, το μήκος των οποίων αντιστοιχεί σε 181 και 204 αμινοξέα, αντίστοιχα, για τις πιθανές παραγόμενες πρωτείνες. Οι προβλεπόμενες ισομορφές περιλαμβάνουν όλα τα κρίσιμα λειτουργικά μοτίβα, που είναι υπεύθυνα για την σύνδεση της κυκλίνης και την τυπική δράση κινάσης, καθώς διαθέτουν ακέραιη την αμινοτελική περιοχή της πρωτείνης (εικόνα 24 & 26). Ωστόσο, στον καρβοξυτελικό λοβό το τμήμα ενεργοποίησης του ενζύμου, που εκτίνεται από το μοτίβο DFG στο μοτίβο APE και περιέχει τη θηλιά της φωσφορυλίωσης, είναι διαταραγμένο, καθώς απουσιάζει η περιοχή που περιλαμβάνει την αμινοξική ακολουθία APE (εικόνα 24 & 26), το οποίο υποδηλώνει ότι τα μόρια αυτά μπορούν να αλληλεπιδράσουν με την κυκλίνη, αλλά χαρακτηρίζονται από απώλεια ή δραματική μείωση της ενζυματικής τους δράσης.

Παρόμοια ευρήματα παρατηρούνται, επίσης, και στα *CDK4* μετάγραφα 7 και 8, τα οποία εμφανίζουν τη δομή διπλού λοβού που παρατηρείται στην κύρια πρωτεΐνη CDK4. Το αμινοτελικό τμήμα των προβλεπόμενων αυτών πρωτεϊνικών ισομορφών είναι πανομοιότυπο με το αντίστοιχο της κύριας πρωτεΐνης, αλλά η θηλιά της φωσφορυλίωσης στην καρβοξυτελική περιοχή είναι σημαντικά παραλλαγμένη (εικόνα 27). Τα υπόλοιπα μετάγραφα (*CDK4* v.9 – v.17), εκτός των μη κωδικών μεταγράφων 11 και 12, που μοιράζονται το γνωστό κωδικόνιο έναρξης, περιέχουν πλήθος δομικών διαφορών σε σχέση με το κύριο *CDK4* μετάγραφο και ακολούθως, οι προβλεπόμενες πρωτεϊνικές ισομορφές που κωδικοποιούν φέρουν σημαντικές δομικές αλλαγές που διαταράσσουν πιθανώς εξ' ολοκλήρου το λειτουργικό ρόλο της κινάσης (εικόνες 24 & 26). Το βασικό ερώτημα που δημιουργείται σχετικά με την δράση των πρωτεϊνών αυτών αφορά στην ικανότητά τους να δεσμεύονται ισχυρά με την κυκλίνη. Ο πιθανός σχηματισμός του συμπλόκου της ισομορφής της κινάσης με την κυκλίνη μπορεί να έχει ως αποτέλεσμα την απενεργοποίηση της δράσης της κύριας κινάσης και την απορρύθμιση του κυτταρικού κύκλου καθώς, αν ο βαθμός συγγένειας των δύο μορίων είναι μεγάλος, ενδέχεται να μην υπάρχουν διαθέσιμα μόρια κυκλινών.

Μια αξιοσημείωτη ομάδα νέων μεταγράφων είναι τα κωδικά mRNAs, τα οποία δεν περιέχουν το εξώνιο 2, όπως τα *CDK4* μετάγραφα 18-21, που χαρακτηρίζονται από την παρουσία διαφορετικών κωδικονίων έναρξης και επομένως, διαθέτουν αισθητά διαφοροποιημένο ανοιχτό πλαίσιο ανάγνωσης (εικόνα 25). Ως εκ τούτου, η αμινοξική ακολουθία των πρωτεϊνών, που ενδέχεται να κωδικοποιούνται, διαφέρει σε αρκετά σημεία από την ακολουθία των αμινοξέων της κυρίαρχης CDK4 (εικόνα 24). Συγκεκριμένα, όπως παρουσιάζεται και στην εικόνα 29, για τις πρωτεϊνικές δομές που προκύπτουν από την έκφραση των μεταγράφων 18-21, η τρισδιάστατη δομή των μορίων είναι εντελώς διαφορετική στο τμήμα της αμινοτελικής περιοχής και απουσιάζει ο πλούσιος σε γλυκινη βρόγχος καθώς και η θέση δέσμευσης της κυκλίνης καθώς η παραγωγή τους εξαρτάται από την λειτουργία ενός νέου κωδικονίου έναρξης, το οποίο βρίσκεται στο 3^{ο} εξώνιο. Σε αυτή την περίπτωση, τα δυνητικά παραγόμενα πρωτεϊνικά μόρια αδυνατούν να συνδεθούν με κυκλίνες και επομένως, μειώνεται η πιθανότητα να δρουν ως κινάσες. Ωστόσο, η πραγματοποίηση της σύνθεσης και παραμονής των συγκεκριμένων πρωτεινών στο κύτταρο δίνει την δυνατότητα φωσφορυλίωσής τους, καθώς η περιοχή της θηλιάς της φωσφορυλίωσης παραμένει ακέραιη, γεγονός που ενδέχεται να οδηγεί σε δυσλειτουργία του κυτταρικού κύκλου στις διάφορες φάσεις του ή, ακόμα, να ρυθμίζει την ενεργοποίηση του κυτταρικού κύκλου και πολλαπλασιασμού.

Παρόλο που η παρούσα διπλωματική εργασία εστιάζει στην ανίχνευση των νέων εναλλακτικών μεταγράφων του γονιδίου *CDK4*, αξίζει να αναφερθεί ότι η υπόθεση της παρούσας διπλωματικής για την κωδικοποιητική ικανότητα των νέων μεταγράφων είναι σε πλήρη συμφωνία με προηγούμενες πρωτεομικές μελέτες, οι οποίες έχουν ήδη προσδιορίσει την ύπαρξη πολλαπλών CDK4 πρωτεϊνικών ισομορφών χρησιμοποιώντας τις μεθοδολογίες western blot και φασματομετρία μάζας. Συγκεκριμένα, ο Sun και οι συνεργάτες του, σχεδίασαν πολλαπλά αντισώματα, τα οποία στοχεύουν σε διαφορετικά σημεία της πρωτεΐνης, όπως τα sc-260 και sc-601, που στοχεύουν το καρβοξυτελικό άκρο, και ακολούθησαν εκτεταμένα πειράματα σε ανθρώπινες κυτταρικές σειρές με την χρήση της μεθοδολογίας western blot. Τα αποτελέσματα της πρωτεομικής τους ανάλυσης αποκάλυψαν, εκτός από την αναμενόμενη ζώνη των 33 kD, η οποία αντιστοιχεί στην κύρια CDK4 ισομορφή, την παρουσία τριών, επιπλέον, μικρότερων ζωνών μεταξύ των 24-28 kD. Τα συγκεκριμένα ευρήματα, παρόλο που απέδειξαν την ύπαρξη CDK4 ισομορφών, χαρακτηρίστηκαν ως «απροσδιόριστα» καθώς δεν ήταν γνωστή η παρουσία εναλλακτικών μεταγράφων με

κωδικοποιητική ικανότητα. Επιπλέον, σύμφωνα με την ίδια μελέτη, μία εκ των τριών «απροσδιόριστων» ζωνών αντιστοιχεί σε μια νέα ισομορφή CDK4, με μοριακό βάρος 25,9 kD, η οποία διαθέτει διαφορετικό κωδικόνιο έναρξης καθώς απουσιάζει το εξώνιο 2. Στην παρούσα μελέτη, το εναλλακτικό μετάγραφο *CDK4* v.18 στερείται πλήρως το εξώνιο 2 και διαθέτει ORF που κωδικοποιεί μια νέα CDK4 πρωτεΐνη με 25,9 kD. Συνεπώς, τα αποτελέσματα της νουκλεοτιδικής μας μελέτης είναι σε ομοφωνία με την πρωτεομική μελέτη των Sun και των συνεργατών του και επίσης, μπορούν να ερμηνεύσουν τις «απροσδιόριστες» ζώνες που εντοπίστηκαν μέσω western blot, καθώς διευκρινίζουν πλήρως την ύπαρξη κωδικοποιητικών mRNA του *CDK4*, τα οποία παράγουν CDK4 ισομορφές με μοριακό βάρος 22-31 kD (εικόνα 29). Πιο συγκεκριμένα, τα νέα μετάγραφα *CDK4* v.2 και v.4 κωδικοποιούν ισομορφές των 25,5 kD και 23,7 kD, αντίστοιχα, καθώς και τα *CDK4* μετάγραφα v.7 και v.20 αναμένεται να κωδικοποιούν πρωτεϊνικές ισομορφές μεταξύ 22-28 kD (27,4 kD και 24 kD, αντίστοιχα), γεγονός που καθιστά αυτά τα μετάγραφα ισχυρούς υποψήφιους για την παραγωγή των δύο πρωτεϊνών που δεν προσδιορίστηκαν.

Επιπλέον, η παρουσία των συγκεκριμένων εναλλακτικών μεταγράφων στα κύτταρα μπορεί να δικαιολογηθεί από την ύπαρξη των διλειτουργικών RNAs (bifunctional RNAs, biRNAs). Τα biRNAs είναι ένας νέος σχετικά τύπος μορίων, τα οποία χαρακτηρίζονται από την ιδιότητά τους να λειτουργούν ενίοτε ως κωδικά και άλλοτε ως μη κωδικά RNAs, και επομένως, έχουν διττό ρόλο και μπορούν είτε να εκτελούν σύνθεση πρωτεινών είτε να ενεργούν ως ρυθμιστικά στοιχεία [132, 133]. Ένας σημαντικός αριθμός, αυτών των πρόσφατα αναγνωρισμένων μορίων, έχει προσδιοριστεί με μεθόδους μαζικής παράλληλης αλληλούχησης, ενώ έχουν επιβεβαιωθεί από τις λειτουργίες διάφορων μορίων, όπως ορισμένα μεγάλα μη κωδικά μόρια RNA (long non-coding RNAs, lncRNAs), τα οποία διαθέτουν μικρού μεγέθους ORFs και υπό προϋποθέσεις κωδικοποιούν πεπτιδικές αλληλουχίες [133].

Στη συγκεκριμένη μελέτη ταυτοποιήθηκαν επιπλέον αρκετά κωδικά μετάγραφα, τα οποία διαθέτουν ανοιχτά πλαίσια ανάγνωσης μικρού μήκους (smORFs) και κωδικοποιούν πεπτίδια μικρότερα των 100 αμινοξέων, με μέσο μήκος 80 αμινοξέα. Σύμφωνα με υπάρχουσες βιβλιογραφικές αναφορές, υπάρχουν πολλά lncRNA μόρια, τα οποία έχουν την δυνατότητα να κωδικοποιούν τέτοιου είδους μικροπεπτίδα, το μήκος των οποίων δεν υπερβαίνει τα 100 αμινοξέα. Η έκφραση των συγκεκριμένων μορίων εξαρτάται από τον τύπο των κυττάρων, τον ιστό και το στάδιο ανάπτυξης στο οποίο

βρίσκονται τα κύτταρα [134, 135]. Πράγματι, ένα αυξανόμενο πλήθος επιστημονικών μελετών επιβεβαιώνει την ευρεία παρουσία των lncRNAs στο ανθρώπινο μεταγράφωμα, με αποτέλεσμα ο αριθμός τους να υπερβαίνει τον αριθμό των κωδικών mRNA μορίων που κωδικοποιούν πρωτεΐνες. Επιπλέον, ένας μεγάλος αριθμός μη κωδικών μορίων έχει smORFs τα οποία μπορούν, σε κατάλληλες συνθήκες, να επάγουν την σύνθεση μικροπεπτιδίων [136, 137]. Οι απαιτήσεις των κυττάρων κάθε χρονική στιγμή καθορίζουν την ισορροπία μεταξύ των επιπέδων των κωδικών και μη κωδικών μορίων RNA ανάλογα με την ανάγκη σύνθεσης πρωτεινών ή την ανάγκη κυτταρικής ρύθμισης ως μόρια - τελεστές [135].

Στη παρούσα μελέτη, οι τρισδιάστατες δομές των προβλεπόμενων μικροπεπτιδίων σχετίζονται με τις αλληλεπιδράσεις που δημιουργούνται μεταξύ αμινοξικών αλληλουχίων με ρυθμιστικά μόρια, όπως αυτά τα μικροπεττίδια. Είναι, επίσης, πιθανόν τα μικροπεπτίδια αυτά να συνδυάζονται με ρυθμιστικά lncRNAs και να ελέγχουν τη μεταγραφή και την μετάφραση βασικών μεταγράφων του γονιδίου σταθεροποιώντας τα μόρια που συμμετέχουν στους μηχανισμούς αυτούς. Τέλος, η ανίχνευση των παραγόμενων μικροπεπτιδίων χρήζει περαιτέρω διερεύνησης, προκειμένου να αποσαφηνιστεί ο ρόλος της λειτουργίας τους στους διάφορους ιστούς, καθώς η παρουσία τους μπορεί να οδηγεί σε φαινόμενα απώλειας ή προσθήκης λειτουργιών (loss / gain -of-function effects) και επομένως, στην εκδήλωση ασθενειών, όπως ο καρκίνος.

Η βιοπληροφορική ανάλυση των αποτελεσμάτων, που λήφθηκαν από τα πειράματα αλληλούχησης κατά τη διάρκεια της παρούσας μελέτης, αποκάλυψε την ύπαρξη συνολικά 10 εναλλακτικών μεταγράφων του *CDK4* τα οποία δεν χαρακτηρίζονται από την παρουσία ORF ή smORF. Η νουκλεοτιδική αλληλουχία καθενός από τα συγκεκριμένα μετάγραφα περιέχει σε κάποια πρώιμη θέση του μεταγράφου ένα PTC με αποτέλεσμα αυτά τα *CDK4* μετάγραφα πιθανότατα να μη διαθέτουν δυναμικό κωδικοποίησης πρωτεϊνών και τελικά να χαρακτηρίζονται ως lncRNAs (εικόνες 23 & 25). Η ανίχνευση των συγκεκριμένων lncRNAs μορίων αποτελεί ένα αρκετά ενδιαφέρον εύρημα, καθώς στην οικογένεια των ανθρώπινων *CDKs* γονιδίων έχει ταυτοποιηθεί ένας πολύ μικρος αριθμός lncRNAs, όπως για παράδειγμα το εναλλακτικό μετάγραφο 10 του γονιδίου *CDK7* (κωδικός πρόσβασης στη GenBank: NR_136690.2). Το συμπέρασμα που μπορεί να εξαχθεί, από την συγκεκριμένη ομάδα μεταγράφων, είναι ότι τα συγκεκριμένα εναλλακτικά μετάγραφα του *CDK4* αντιπροσωπεύουν μεγάλα μη κωδικά

μόρια RNA, lncRNAs, τα οποία εμπλέκονται σε πληθώρα κυτταρικών διεργασιών και τα οποία χαρακτηρίζονται ως μεσολαβητές που ρόλο έχουν την επίτευξη των διασυνδέσεων μεταξύ RNA και πρωτεινών. Επιπλέον, μπορεί να δρουν ως ρυθμιστές της γονιδιακής έκφρασης σε διάφορα επίπεδα και να ενισχύουν την ενεργοποίηση βιολογικών μηχανισμών των κυττάρων [138, 139].

Συνοψίζοντας, η παρούσα μελέτη αποσαφηνίζει πλήρως το γενωμικό προφίλ του γονιδίου *CDK4*. Με την εφαρμογή των μεθόδων μαζικής παράλληλης αλληλούχησης, και κυρίως, με την καινοτόμο τεχνολογία αλληλούχησης τρίτης γενιάς, με την χρήση της πλατφόρμας Oxford Nanopore Technologies®, ανιχνεύθηκαν και προσδιορίστηκαν τα, εως σήμερα, άγνωστα εναλλακτικά μετάγραφα του γονιδίου *CDK4,* το οποίο αποτελεί, χωρίς αμφιβολία, ένα βασικό παράγοντα που συμμετέχει στη κυτταρική ομοιόσταση και συμβάλλει στη ρύθμιση των μηχανισμών πολλαπλασιασμού και ανάπτυξης των κυττάρων. Η μελέτη που παρουσιάστηκε, διευκρινίζει για πρώτη φορά το περίπλοκο μεταγραφικό προφίλ του *CDK4* γονιδίου που αφορά στην κατανόηση των γενικών μηχανισμών εναλλακτικού ματίσματος που συμβαίνουν στα κύτταρα.

Παράλληλλα, στη παρούσα διπλωματική εργασία γίνεται προσπάθεια να απαντηθούν ερωτήματα για την τύχη των εναλλακτικών mRNA μεταγράφων, τα οποία παράγονται από ένα μόνο γονίδιο στα ευκαρυωτικά κύτταρα. Με βάση την κατηγοριοποίηση των εναλλακτικών μεταγράφων, που ταυτοποιήθηκαν σε διαφορετικές ρυθμιστικές και λειτουργικές ομάδες ανάλογα με τη δομη και τις αλληλεπιδράσεις τους, τα αποτελέσματα της παρούσας μελέτης ενισχύουν περαιτέρω την θέση ότι, ένα μόνο γονίδιο είναι ικανό να παράγει πλήθος εναλλακτικών μεταγράφων. Τα εναλλακτικά mRNAs ενός γονιδίου μπορεί να αποτελούν λειτουργικά στοιχεία έχοντας κωδικοποιητική δράση, αλλά επιπλέον, ορισμένα μπορεί να δρουν ως ρυθμιστικοί μεσολαβητές για την πραγματοποίηση συγκεκριμένων κυτταρικών διαδικασιών. Το ευρύ φάσμα των εναλλακτικών μεταγράφων του γονιδίου *CDK4,* το οποίο παρουσιάστηκε, αποτελεί το πρώτο βήμα για την συναρμολόγηση των άγνωστων «κομματίων του παζλ», τα οποία σχετίζονται με τις ακριβείς λειτουργίες της, άκρως πολύτιμης για το κύτταρο, κυκλινοεξαρτώμενης κινάσης 4 και την διερεύνηση των επιπτώσεων της απώλειας του μορίου αυτού στην κυτταρική ομοιόσταση και την εκδήλωση ασθενειών.

5. ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Dahm, R., *Friedrich Miescher and the discovery of DNA.* Dev Biol, 2005. **278**(2): p. 274-88.
2. Dahm, R., *Discovering DNA: Friedrich Miescher and the early years of nucleic acid research.* Hum Genet, 2008. **122**(6): p. 565-81.
3. Veigl, S.J., O. Harman, and E. Lamm, *Friedrich Miescher's Discovery in the Historiography of Genetics: From Contamination to Confusion, from Nuclein to DNA.* J Hist Biol, 2020. **53**(3): p. 451-484.
4. Portin, P., *The birth and development of the DNA theory of inheritance: sixty years since the discovery of the structure of DNA.* J Genet, 2014. **93**(1): p. 293-302.
5. Brush, S.G., *How theories became knowledge: Morgan's chromosome theory of heredity in America and Britain.* J Hist Biol, 2002. **35**(3): p. 471-535.
6. Avery, O.T., C.M. Macleod, and M. McCarty, *Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii.* J Exp Med, 1944. **79**(2): p. 137-58.
7. Hershey, A.D. and M. Chase, *Independent functions of viral protein and nucleic acid in growth of bacteriophage.* J Gen Physiol, 1952. **36**(1): p. 39-56.
8. Wilkins, M.H., A.R. Strokes, and H.R. Wilson, *Molecular structure of deoxypentose nucleic acids. 1953.* Nature, 2003. **421**(6921): p. 398-400; discussion 396.
9. Franklin, R.E. and R.G. Gosling, *Molecular configuration in sodium thymonucleate.* Nature, 1953. **171**(4356): p. 740-1.
10. Watson, J.D. and F.H. Crick, *The structure of DNA.* Cold Spring Harb Symp Quant Biol, 1953. **18**: p. 123-31.
11. Watson, J.D. and F.H. Crick, *Genetical implications of the structure of deoxyribonucleic acid.* Nature, 1953. **171**(4361): p. 964-7.
12. Portin, P., *The concept of the gene: short history and present status.* Q Rev Biol, 1993. **68**(2): p. 173-223.
13. Franca, L.T., E. Carrilho, and T.B. Kist, *A review of DNA sequencing techniques.* Q Rev Biophys, 2002. **35**(2): p. 169-200.
14. McGinn, S. and I.G. Gut, *DNA sequencing - spanning the generations.* N Biotechnol, 2013. **30**(4): p. 366-72.

15. Shendure, J., et al., *DNA sequencing at 40: past, present and future.* Nature, 2017. **550**(7676): p. 345-353.
16. Pichersky, E., *Terminal labeling of DNA for Maxam and Gilbert sequencing.* Methods Mol Biol, 1993. **23**: p. 247-53.
17. Pichersky, E., *Terminal labeling of DNA for Maxam and Gilbert sequencing.* Methods Mol Biol, 1996. **58**: p. 441-6.
18. Wu, R., *Nucleotide sequence analysis of DNA.* Nat New Biol, 1972. **236**(68): p. 198-200.
19. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors.* Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
20. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.
21. Staden, R., *A strategy of DNA sequencing employing computer programs.* Nucleic Acids Res, 1979. **6**(7): p. 2601-10.
22. Chen, C.Y., *DNA polymerases drive DNA sequencing-by-synthesis technologies: both past and present.* Front Microbiol, 2014. **5**: p. 305.
23. Ansorge, W.J., *Next-generation DNA sequencing techniques.* N Biotechnol, 2009. **25**(4): p. 195-203.
24. Pareek, C.S., R. Smoczynski, and A. Tretyn, *Sequencing technologies and genome sequencing.* J Appl Genet, 2011. **52**(4): p. 413-35.
25. Hall, N., *Advanced sequencing technologies and their wider impact in microbiology.* J Exp Biol, 2007. **210**(Pt 9): p. 1518-25.
26. Yohe, S. and B. Thyagarajan, *Review of Clinical Next-Generation Sequencing.* Arch Pathol Lab Med, 2017. **141**(11): p. 1544-1557.
27. Garrido-Cardenas, J.A., et al., *DNA Sequencing Sensors: An Overview.* Sensors (Basel), 2017. **17**(3).
28. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
29. Goodwin, S., et al., *Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome.* Genome Res, 2015. **25**(11): p. 1750-6.
30. Su, Z., et al., *Next-generation sequencing and its applications in molecular diagnostics.* Expert Rev Mol Diagn, 2011. **11**(3): p. 333-43.

31. Shokralla, S., et al., *Next-generation sequencing technologies for environmental DNA research.* Mol Ecol, 2012. **21**(8): p. 1794-805.
32. Ahmadian, A., M. Ehn, and S. Hober, *Pyrosequencing: history, biochemistry and future.* Clin Chim Acta, 2006. **363**(1-2): p. 83-94.
33. Harrington, C.T., et al., *Fundamentals of pyrosequencing.* Arch Pathol Lab Med, 2013. **137**(9): p. 1296-303.
34. Yan, J.B., et al., *Pyrosequencing is an accurate and reliable method for the analysis of heteroplasmy of the A3243G mutation in patients with mitochondrial diabetes.* J Mol Diagn, 2014. **16**(4): p. 431-9.
35. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.
36. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
37. Dunham, J.P. and M.L. Friesen, *A cost-effective method for high-throughput construction of illumina sequencing libraries.* Cold Spring Harb Protoc, 2013. **2013**(9): p. 820-34.
38. Morozova, O., M. Hirst, and M.A. Marra, *Applications of new sequencing technologies for transcriptome analysis.* Annu Rev Genomics Hum Genet, 2009. **10**: p. 135-51.
39. Machida, R.J. and Y.Y. Lin, *Four methods of preparing mRNA 5' end libraries using the Illumina sequencing platform.* PLoS One, 2014. **9**(7): p. e101812.
40. Turcatti, G., et al., *A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis.* Nucleic Acids Res, 2008. **36**(4): p. e25.
41. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.
42. Kumar, R., et al., *A High-Throughput Method for Illumina RNA-Seq Library Preparation.* Front Plant Sci, 2012. **3**: p. 202.
43. Slatko, B.E., A.F. Gardner, and F.M. Ausubel, *Overview of Next-Generation Sequencing Technologies.* Curr Protoc Mol Biol, 2018. **122**(1): p. e59.
44. Kircher, M., S. Sawyer, and M. Meyer, *Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform.* Nucleic Acids Res, 2012. **40**(1): p. e3.

45. Bentley, D.R., *Whole-genome re-sequencing.* Curr Opin Genet Dev, 2006. **16**(6): p. 545-52.
46. Balasubramanian, S., *Sequencing nucleic acids: from chemistry to medicine.* Chem Commun (Camb), 2011. **47**(26): p. 7281-6.
47. Quail, M.A., et al., *A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers.* BMC Genomics, 2012. **13**: p. 341.
48. Rothberg, J.M., et al., *An integrated semiconductor device enabling non-optical genome sequencing.* Nature, 2011. **475**(7356): p. 348-52.
49. Tops, B.B., et al., *Development of a semi-conductor sequencing-based panel for genotyping of colon and lung cancer by the Onconetwork consortium.* BMC Cancer, 2015. **15**: p. 26.
50. Merriman, B., et al., *Progress in ion torrent semiconductor chip based sequencing.* Electrophoresis, 2012. **33**(23): p. 3397-417.
51. Huang, X., et al., *A Dual-Mode Large-Arrayed CMOS ISFET Sensor for Accurate and High-Throughput pH Sensing in Biomedical Diagnosis.* IEEE Trans Biomed Eng, 2015. **62**(9): p. 2224-33.
52. Li, H., et al., *CMOS Electrochemical Instrumentation for Biosensor Microsystems: A Review.* Sensors (Basel), 2016. **17**(1).
53. Glenn, T.C., *Field guide to next-generation DNA sequencers.* Mol Ecol Resour, 2011. **11**(5): p. 759-69.
54. Gomez, J., et al., *Non optical semi-conductor next generation sequencing of the main cardiac QT-interval duration genes in pooled DNA samples.* J Cardiovasc Transl Res, 2014. **7**(1): p. 133-7.
55. Nakano, M., et al., *Single-molecule PCR using water-in-oil emulsion.* J Biotechnol, 2003. **102**(2): p. 117-24.
56. Kijas, J.M., et al., *Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles.* Biotechniques, 1994. **16**(4): p. 656-60, 662.
57. Balla, B., et al., *Fast and robust next-generation sequencing technique using ion torrent personal genome machine for the screening of neurofibromatosis type 1 (NF1) gene.* J Mol Neurosci, 2014. **53**(2): p. 204-10.

58. Harris, T.D., et al., *Single-molecule DNA sequencing of a viral genome.* Science, 2008. **320**(5872): p. 106-9.
59. Chaisson, M.J. and P.A. Pevzner, *Short read fragment assembly of bacterial genomes.* Genome Res, 2008. **18**(2): p. 324-30.
60. Pickrell, J.K., et al., *Understanding mechanisms underlying human gene expression variation with RNA sequencing.* Nature, 2010. **464**(7289): p. 768-72.
61. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
62. Kharchenko, P.V., M.Y. Tolstorukov, and P.J. Park, *Design and analysis of ChIP-seq experiments for DNA-binding proteins.* Nat Biotechnol, 2008. **26**(12): p. 1351-9.
63. Mooney, S.D., *Progress towards the integration of pharmacogenomics in practice.* Hum Genet, 2015. **134**(5): p. 459-65.
64. van El, C.G., et al., *Whole-genome sequencing in health care. Recommendations of the European Society of Human Genetics.* Eur J Hum Genet, 2013. **21 Suppl 1**: p. S1-5.
65. Choi, M., et al., *Genetic diagnosis by whole exome capture and massively parallel DNA sequencing.* Proc Natl Acad Sci U S A, 2009. **106**(45): p. 19096-101.
66. Ng, S.B., et al., *Exome sequencing identifies the cause of a mendelian disorder.* Nat Genet, 2010. **42**(1): p. 30-5.
67. Raffan, E. and R.K. Semple, *Next generation sequencing--implications for clinical practice.* Br Med Bull, 2011. **99**: p. 53-71.
68. Gulilat, M., et al., *Targeted next generation sequencing as a tool for precision medicine.* BMC Med Genomics, 2019. **12**(1): p. 81.
69. Ekblom, R. and J. Galindo, *Applications of next generation sequencing in molecular ecology of non-model organisms.* Heredity (Edinb), 2011. **107**(1): p. 1-15.
70. Veneziano, D., G. Nigita, and A. Ferro, *Computational Approaches for the Analysis of ncRNA through Deep Sequencing Techniques.* Front Bioeng Biotechnol, 2015. **3**: p. 77.
71. So, K.K., et al., *Whole Genome Chromatin IP-Sequencing (ChIP-Seq) in Skeletal Muscle Cells.* Methods Mol Biol, 2017. **1668**: p. 15-25.
72. Muhammad, II, et al., *RNA-seq and ChIP-seq as Complementary Approaches for Comprehension of Plant Transcriptional Regulatory Mechanism.* Int J Mol Sci, 2019. **21**(1).

73. Pages, A., et al., *The discovery potential of RNA processing profiles.* Nucleic Acids Res, 2018. **46**(3): p. e15.
74. Gong, L., et al., *Picky comprehensively detects high-resolution structural variants in nanopore long reads.* Nat Methods, 2018. **15**(6): p. 455-460.
75. van Dijk, E.L., et al., *The Third Revolution in Sequencing Technology.* Trends Genet, 2018. **34**(9): p. 666-681.
76. Xiao, T. and W. Zhou, *The third generation sequencing: the advanced approach to genetic diseases.* Transl Pediatr, 2020. **9**(2): p. 163-173.
77. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
78. Flusberg, B.A., et al., *Direct detection of DNA methylation during single-molecule, real-time sequencing.* Nat Methods, 2010. **7**(6): p. 461-5.
79. Loman, N.J. and A.R. Quinlan, *Poretools: a toolkit for analyzing nanopore sequence data.* Bioinformatics, 2014. **30**(23): p. 3399-401.
80. Loman, N.J., J. Quick, and J.T. Simpson, *A complete bacterial genome assembled de novo using only nanopore sequencing data.* Nat Methods, 2015. **12**(8): p. 733-5.
81. Lu, H., F. Giordano, and Z. Ning, *Oxford Nanopore MinION Sequencing and Genome Assembly.* Genomics Proteomics Bioinformatics, 2016. **14**(5): p. 265-279.
82. Petersen, L.M., et al., *Third-Generation Sequencing in the Clinical Laboratory: Exploring the Advantages and Challenges of Nanopore Sequencing.* J Clin Microbiol, 2019. **58**(1).
83. Quick, J., et al., *Real-time, portable genome sequencing for Ebola surveillance.* Nature, 2016. **530**(7589): p. 228-232.
84. Rang, F.J., W.P. Kloosterman, and J. de Ridder, *From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy.* Genome Biol, 2018. **19**(1): p. 90.
85. Thapar, R., et al., *RNA Modifications: Reversal Mechanisms and Cancer.* Biochemistry, 2019. **58**(5): p. 312-329.
86. Gott, J.M. and R.B. Emeson, *Functions and mechanisms of RNA editing.* Annu Rev Genet, 2000. **34**: p. 499-531.
87. Berget, S.M., C. Moore, and P.A. Sharp, *Spliced segments at the 5' terminus of adenovirus 2 late mRNA.* Proc Natl Acad Sci U S A, 1977. **74**(8): p. 3171-5.

88. Darnell, J.E., Jr., *Implications of RNA-RNA splicing in evolution of eukaryotic cells.* Science, 1978. **202**(4374): p. 1257-60.
89. Will, C.L. and R. Luhrmann, *Spliceosome structure and function.* Cold Spring Harb Perspect Biol, 2011. **3**(7).
90. Shi, Y., *Mechanistic insights into precursor messenger RNA splicing by the spliceosome.* Nat Rev Mol Cell Biol, 2017. **18**(11): p. 655-670.
91. Wahl, M.C., C.L. Will, and R. Luhrmann, *The spliceosome: design principles of a dynamic RNP machine.* Cell, 2009. **136**(4): p. 701-18.
92. Verma, B., et al., *Minor spliceosome and disease.* Semin Cell Dev Biol, 2018. **79**: p. 103-112.
93. Chen, M. and J.L. Manley, *Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches.* Nat Rev Mol Cell Biol, 2009. **10**(11): p. 741-54.
94. Baralle, F.E. and J. Giudice, *Alternative splicing as a regulator of development and tissue identity.* Nat Rev Mol Cell Biol, 2017. **18**(7): p. 437-451.
95. Modrek, B. and C. Lee, *A genomic view of alternative splicing.* Nat Genet, 2002. **30**(1): p. 13-9.
96. Adamopoulos, P.G., et al., *Unraveling novel survivin mRNA transcripts in cancer cells using an in-house developed targeted high-throughput sequencing approach.* Genomics, 2020.
97. Wang, E. and I. Aifantis, *RNA Splicing and Cancer.* Trends Cancer, 2020. **6**(8): p. 631-644.
98. Hallegger, M., M. Llorian, and C.W. Smith, *Alternative splicing: global insights.* FEBS J, 2010. **277**(4): p. 856-66.
99. Bonnal, S.C., I. Lopez-Oreja, and J. Valcarcel, *Roles and mechanisms of alternative splicing in cancer - implications for care.* Nat Rev Clin Oncol, 2020. **17**(8): p. 457-474.
100. Ladomery, M., *Aberrant alternative splicing is another hallmark of cancer.* Int J Cell Biol, 2013. **2013**: p. 463786.
101. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.
102. Wood, D.J. and J.A. Endicott, *Structural insights into the functional diversity of the CDK-cyclin family.* Open Biol, 2018. **8**(9).

103. Bockstaele, L., et al., *Regulated activating Thr172 phosphorylation of cyclin-dependent kinase 4(CDK4): its relationship with cyclins and CDK "inhibitors".* Mol Cell Biol, 2006. **26**(13): p. 5070-85.

104. Kato, J.Y., et al., *Regulation of cyclin D-dependent kinase 4 (cdk4) by cdk4-activating kinase.* Mol Cell Biol, 1994. **14**(4): p. 2713-21.

105. Sherr, C.J., D. Beach, and G.I. Shapiro, *Targeting CDK4 and CDK6: From Discovery to Therapy.* Cancer Discov, 2016. **6**(4): p. 353-67.

106. Malumbres, M., et al., *Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6.* Cell, 2004. **118**(4): p. 493-504.

107. Sherr, C.J., *G1 phase progression: cycling on cue.* Cell, 1994. **79**(4): p. 551-5.

108. Day, P.J., et al., *Crystal structure of human CDK4 in complex with a D-type cyclin.* Proc Natl Acad Sci U S A, 2009. **106**(11): p. 4166-70.

109. Klein, M.E., et al., *CDK4/6 Inhibitors: The Mechanism of Action May Not Be as Simple as Once Thought.* Cancer Cell, 2018. **34**(1): p. 9-20.

110. Russo, A.A., P.D. Jeffrey, and N.P. Pavletich, *Structural basis of cyclin-dependent kinase activation by phosphorylation.* Nat Struct Biol, 1996. **3**(8): p. 696-700.

111. Shao, Z. and P.D. Robbins, *Differential regulation of E2F and Sp1-mediated transcription by G1 cyclins.* Oncogene, 1995. **10**(2): p. 221-8.

112. VanArsdale, T., et al., *Molecular Pathways: Targeting the Cyclin D-CDK4/6 Axis for Cancer Treatment.* Clin Cancer Res, 2015. **21**(13): p. 2905-10.

113. Murphy, C.G. and M.N. Dickler, *The Role of CDK4/6 Inhibition in Breast Cancer.* Oncologist, 2015. **20**(5): p. 483-90.

114. Rocca, A., et al., *Progress with palbociclib in breast cancer: latest evidence and clinical considerations.* Ther Adv Med Oncol, 2017. **9**(2): p. 83-105.

115. Zhu, J., et al., *On the nature of human housekeeping genes.* Trends Genet, 2008. **24**(10): p. 481-4.

116. Hounkpe, B.W., et al., *HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets.* Nucleic Acids Res, 2021. **49**(D1): p. D947-D955.

117. Tarze, A., et al., *GAPDH, a novel regulator of the pro-apoptotic mitochondrial membrane permeabilization.* Oncogene, 2007. **26**(18): p. 2606-20.

118. Kim, D., B. Langmead, and S.L. Salzberg, *HISAT: a fast spliced aligner with low memory requirements.* Nat Methods, 2015. **12**(4): p. 357-60.

119. Adamopoulos, P.G., M.C. Theodoropoulou, and A. Scorilas, *Alternative Splicing Detection Tool-a novel PERL algorithm for sensitive detection of splicing events, based on next-generation sequencing data analysis.* Ann Transl Med, 2018. **6**(12): p. 244.
120. Wick, R.R., L.M. Judd, and K.E. Holt, *Performance of neural network basecalling tools for Oxford Nanopore sequencing.* Genome Biol, 2019. **20**(1): p. 129.
121. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.
122. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic Acids Res, 2003. **31**(13): p. 3784-8.
123. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nat Methods, 2015. **12**(1): p. 7-8.
124. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
125. Ozsolak, F. and P.M. Milos, *Single-molecule direct RNA sequencing without cDNA synthesis.* Wiley Interdiscip Rev RNA, 2011. **2**(4): p. 565-70.
126. Adamopoulos, P.G., et al., *Identification of novel alternative transcripts of the human Ribonuclease kappa (RNASEK) gene using 3' RACE and high-throughput sequencing approaches.* Genomics, 2020. **112**(1): p. 943-951.
127. LeBlanc, V.G. and M.A. Marra, *Next-Generation Sequencing Approaches in Cancer: Where Have They Brought Us and Where Will They Take Us?* Cancers (Basel), 2015. **7**(3): p. 1925-58.
128. Thermann, R., et al., *Binary specification of nonsense codons by splicing and cytoplasmic translation.* EMBO J, 1998. **17**(12): p. 3484-94.
129. Zhang, J., et al., *At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation.* Mol Cell Biol, 1998. **18**(9): p. 5272-83.
130. Shi, M., et al., *Premature Termination Codons Are Recognized in the Nucleus in A Reading-Frame Dependent Manner.* Cell Discov, 2015. **1**.
131. Kanev, G.K., et al., *The Landscape of Atypical and Eukaryotic Protein Kinases.* Trends Pharmacol Sci, 2019. **40**(11): p. 818-832.
132. Hube, F. and C. Francastel, *Coding and Non-coding RNAs, the Frontier Has Never Been So Blurred.* Front Genet, 2018. **9**: p. 140.

133. Nam, J.W., S.W. Choi, and B.H. You, *Incredible RNA: Dual Functions of Coding and Noncoding.* Mol Cells, 2016. **39**(5): p. 367-74.
134. Choi, S.W., H.W. Kim, and J.W. Nam, *The small peptide world in long noncoding RNAs.* Brief Bioinform, 2019. **20**(5): p. 1853-1864.
135. Robinson, E.K., S. Covarrubias, and S. Carpenter, *The how and why of lncRNA function: An innate immune perspective.* Biochim Biophys Acta Gene Regul Mech, 2020. **1863**(4): p. 194419.
136. Dhamija, S. and M.B. Menon, *Non-coding transcript variants of protein-coding genes - what are they good for?* RNA Biol, 2018. **15**(8): p. 1025-1031.
137. Ransohoff, J.D., Y. Wei, and P.A. Khavari, *The functions and unique features of long intergenic non-coding RNA.* Nat Rev Mol Cell Biol, 2018. **19**(3): p. 143-157.
138. Kitagawa, M., et al., *Cell cycle regulation by long non-coding RNAs.* Cell Mol Life Sci, 2013. **70**(24): p. 4785-94.
139. Chen, L.L. and G.G. Carmichael, *Decoding the function of nuclear long non-coding RNAs.* Curr Opin Cell Biol, 2010. **22**(3): p. 357-64.

ΠΕΡΙΛΗΨΗ

Το γονίδιο *CDK4* είναι μέλος της οικογένειας των κυκλινοεξαρτώμενων κινασών, η οποία έχει εξαιρετικά σημαντικό ρόλο στα μονοπάτια σηματοδότησης του κυττάρου, τη ρύθμιση της μεταγραφής και την κυτταρική διαίρεση. Η ελαττωματική δράση του συμπλόκου που δημιουργεί η κινάση με την κυκλίνη D1, ενδέχεται να οδηγεί σε ενίσχυση του κυτταρικού πολλαπλασιασμού και επομένως να εμπλέκεται στην καρκινογένεση. Παρά το γεγονός ότι ο βιολογικός ρόλος του γονιδίου *CDK4* έχει μελετηθεί σε μεγάλο βαθμό, ο μηχανισμός που αφορά την ωρίμανση των πρόδρομων mRNA μορίων, τόσο σε φυσιολογικές όσο και παθολογικές καταστάσεις, δεν έχει διερευνηθεί. Επομένως, η ταυτοποίηση πιθανών νέων εναλλακτικών μεταγράφων του γονιδίου *CDK4*, ειδικά εκείνων που κωδικοποιούν για πρωτεινες, θα μπορούσε να οδηγήσει στο χαρακτηρισμό νέων διαγνωστικών ή/και προγνωστικών βιοδεικτών ή νέων θεραπευτικών στόχων.

Στην παρούσα διπλωματική εργασία, σχεδιάστηκε και εφαρμόστηκε μια στοχευμένη μέθοδος αλληλούχησης με τη χρήση της τεχνολογίας του νανοπόρου, με την οποία επιτεύχθηκε εξαιρετικά υψηλό βάθος ανάγνωσης και εκτεταμένη διερεύνηση νέων πιθανών *CDK4* mRNA μορίων. Εξαιτίας του αυξημένου ποσοστού λαθών που συμβαίνουν κατά την αλληλούχηση μέσω νανοπόρων, η πειραματική επιβεβαίωση των αποτελεσμάτων πραγματοποιήθηκε με αντίδραση αλληλούχησης νέας γενιάς και πιο συγκεκριμένα με χρήση της τεχνολογίας αλληλούχησης μέσω ημιαγωγού.

Η παρούσα διπλωματική εργασία αποσαφηνίζει για πρώτη φορά το περίπλοκο μεταγραφικό προφίλ του ανθρώπινου γονιδίου *CDK4*, αποκαλύπτωντας την παρουσία άγνωστων εναλλακτικών μεταγράφων που προέρχονται από νέα γεγονότα εναλλακτικής συρραφής. Παράλληλα, η μελέτη της μεταφραστικής ικανότητας των νέων μεταγράφων οδήγησε στο συμπέρασμα πως η πλειοψηφία των νέων μορίων mRNAs διαθέτουν ανοιχτό πλαίσιο ανάγνωσης και συνεπώς προβλέπεται να κωδικοποιούν νέες πρωτεϊνικές ισομορφές. Επιπλέον, πραγματοποιήθηκε μελέτη του προφίλ έκφρασης των νέων *CDK4* μεταγράφων σε ανθρώπινες καρκινικές κυτταρικές σειρές με τη χρήση ποσοτικής PCR σε πραγματικό χρόνο. Το ευρύ φάσμα των εναλλακτικών μεταγράφων του γονιδίου *CDK4* (*CDK4* v.2 – v.42) αποτελεί το πρώτο βήμα για την συναρμολόγηση των κομματίων που θα αποκαλύψουν την εμπλοκή της κινάσης στην κυτταρική ομοιόσταση και την παθοφυσιολογία.

ABSTRACT

CDK4 is a member of the cyclin-dependent kinases, a family of protein kinases with outstanding roles in signalling pathways, transcription regulation and cell division. Defective or overactivated CDK4/cyclin D1 pathway leads to enhanced cellular proliferation, thus being implicated in human cancers. Although the biological role of CDK4 has been extensively studied, its pre-mRNA processing mechanism under normal or pathological conditions is neglected. Thus, the identification of novel *CDK4* mRNA transcripts, especially protein-coding ones, could lead to the identification of new diagnostic and/or prognostic biomarkers or new therapeutic targets.

In the present study, instead of using the “gold-standard” direct RNA sequencing application, a targeted nanopore sequencing approach was designed and employed, which offered a tremendously higher sequencing depth and enabled the thorough investigation of new putative *CDK4* mRNAs. Due to the notable error rates of nanopore sequencing, validation of the results was carried out with next-generation sequencing based on the semi-conductor technology.

The present study elucidates for the first time the complex transcriptional landscape of the human *CDK4* gene, highlighting the existence of previously unknown *CDK4* transcripts with new alternative splicing events and protein-coding capacities. The relative expression levels of each novel *CDK4* transcript in human malignancies were elucidated with custom qPCR-based assays. The presented wide spectrum of *CDK4* transcripts (*CDK4* v.2 – v.42) is only the first step to distinguish and assemble the missing pieces regarding the exact functions and implications of this fundamental kinase in cellular homeostasis and pathophysiology.

ΠΑΡΑΡΤΗΜΑ

Στο πλαίσιο της παρούσας διπλωματικής εργασίας πραγματοποιήθηκε η παρακάτω δημοσίευση σε διεθνές επιστημονικό περιοδικό με κριτές:

Adamopoulos PG, **Athanasopoulou K**, Tsiakanikas P, Scorilas A. A comprehensive nanopore sequencing methodology deciphers the complete transcriptional landscape of cyclin dependent kinase 4 (CDK4) in human malignancies. ***FEBS J***. 2021; doi: 10.1111/febs.16201.

The FEBS Journal

FEBS PRESS science publishing by scientists

A comprehensive nanopore sequencing methodology deciphers the complete transcriptional landscape of cyclin-dependent kinase 4 (CDK4) in human malignancies

Panagiotis G. Adamopoulos, Konstantina Athanasopoulou, Panagiotis Tsiakanikas and Andreas Scorilas

Department of Biochemistry and Molecular Biology, National and Kapodistrian University of Athens, Athens, Greece

Keywords
alternative splicing; CDK4; cyclin-dependent kinases; long-read sequencing; nanopore sequencing

Correspondence
A. Scorilas, Department of Biochemistry and Molecular Biology, National and Kapodistrian University of Athens, Panepistimiopolis, 15701 Athens, Greece
Tel: +30 2107274306
E-mail: ascorilas@biol.uoa.gr

(Received 5 May 2021, revised 2 August 2021, accepted 14 September 2021)

doi:10.1111/febs.16201

Cyclin-dependent kinase 4 (CDK4) is a member of the cyclin-dependent kinases, a family of protein kinases with outstanding roles in signaling pathways, transcription regulation, and cell division. Defective or overactivated CDK4/cyclin D1 pathway leads to enhanced cellular proliferation, thus being implicated in human cancers. Although the biological role of CDK4 has been extensively studied, its pre-mRNA processing mechanism under normal or pathological conditions is neglected. Thus, the identification of novel *CDK4* mRNA transcripts, especially protein-coding ones, could lead to the identification of new diagnostic and/or prognostic biomarkers or new therapeutic targets. In the present study, instead of using the 'gold standard' direct RNA sequencing application, we designed and employed a targeted nanopore sequencing approach, which offers higher sequencing depth and enables the thorough investigation of new mRNAs of any target gene. Our study elucidates for the first time the complex transcriptional landscape of the human *CDK4* gene, highlighting the existence of previously unknown *CDK4* transcripts with new alternative splicing events and protein-coding capacities. The relative expression levels of each novel *CDK4* transcript in human malignancies were elucidated with custom qPCR-based assays. The presented wide spectrum of *CDK4* transcripts (*CDK4* v.2–v.42) is only the first step to distinguish and assemble the missing pieces regarding the exact functions and implications of this fundamental kinase in cellular homeostasis and pathophysiology.

Introduction

It is well-documented that in the eucaryotic cells, the processing of precursor mRNAs (pre-mRNAs) towards mature mRNAs is a mandatory step before their final translation into peptides [1]. The principal modifications during pre-mRNA processing include intron removal, exon joining, formation of the 5′ prime cap (5′-cap), and the addition of the poly(A) tail in the 3′ end [2]. Pre-mRNA splicing is regulated by a multicomponent ribonucleoprotein (RNP) complex, composed of five distinct snRNPs and multiple peptides, which is commonly addressed as the spliceosome. The spliceosome is charged with the accurate recognition of splice sites, the subsequent intron excision, and the differential splicing of exons, a

Abbreviations
aa, amino acids; ATCC, American Type Culture Collection; CDKs, cyclin-dependent kinases; IGV, Integrative Genomics Viewer; lncRNAs, long noncoding RNAs; NGS, next-generation sequencing; NMD, nonsense-mediated mRNA decay; ORF, open-reading frame; pre-mRNAs, precursor mRNAs; PTC, premature termination codon; qPCR, quantitative PCR; RNP, ribonucleoprotein; smORFs, small open-reading frames; SRA, sequence read archive; TGS, third-generation sequencing; T_m, melting temperature.

procedure called the alternative splicing, which is the most essential mechanism for the generation of numerous mature mRNA transcript variants from a single gene [3]. The resultant production of multiple mRNAs from a single gene leads to the translation of various protein isoforms with distinct structural characteristics and biological attributes, thus greatly contributing to both transcriptome and proteome complexity [4]. Since the vast majority (almost 95%) of human multi-exon genes are subjected to alternative splicing, this mechanism has emerged as a dynamic rheostat of gene expression, tightly regulating not only how many transcript variants will be transcribed from each gene in specific timepoint and/or developmental stage but also their expression profile and abundancy in different tissues [5,6]. Alternative splicing dictates the production of mRNA transcripts through a wide repertoire of mechanisms, including the exon skipping (also known as cassette-type alternative exon), mutually exclusive exons, intron retention, alternative promoter and transcription termination usage, alternative selection of 5′ or 3′ splice sites, and alternative polyadenylation [7]. An increasing number of studies have confirmed that mRNA splicing is a fundamental mechanism that is firmly associated with tumorigenesis, cancer progression, and metastasis, since defects in the alternative splicing machinery can lead to the expression of tumor-specific alternative transcripts [8–10]. Therefore, the identification and complete characterization of novel mRNA transcript variants of human genes constitute a considerable challenge in the field of modern genomics. Especially in terms of cancer-related genes, the in-depth knowledge of their transcriptional profile could be very significant, since several of these mRNA transcripts may possess diagnostic and/or prognostic attributes, or even represent promising therapeutic targets [11].

During the last decade, considerable advances in sequencing technology have transformed our perspective regarding the dynamics of human transcriptome, not only by identifying new RNA molecules with diverse attributes in both physiological and pathological states [12,13] but also by detecting novel alternative splice variants firmly associated with particular diseases [14]. Under that prism, next-generation sequencing (NGS) has emerged, without any doubt, as the most powerful research tool that enables the identification of novel alternative splicing events, even those with significantly decreased abundancy compared to the predominant ones, that traditional cloning and sequencing techniques were incapable of detecting. Despite the tremendous increase in both sequencing depth and coverage, NGS technology requires a clonal amplification of the target DNA that may incorporate systematic errors during downstream analysis, whereas the signal production from multiple clones results in limited read lengths (typically 150–400 nucleotides), an approach known as 'short-read' high-throughput sequencing. Consequently, the analysis of NGS datasets requires the use of large computational resources and specialized bioinformatic tools that will enable the alignment of the short reads to the reference genome or sequence, but most importantly the assembly of the short reads for the formation of larger contigs.

However, more recently, third-generation sequencing (TGS), also referred as single-molecule sequencing, has revolutionized the scientific field of sequencing technology, facilitating our understanding regarding the abundance, diversity, and molecular features of human genome and transcriptome [15]. TGS constitutes a state-of-the-art sequencing technology providing a real-time analysis [16] in which ultra-long reads can be sequenced with or without PCR amplification or fragmentation of the template. These capabilities ensure an increased sequencing depth, a better coverage, and most importantly an unbiased (PCR-free) overview of the genomic or transcriptomic profile. At the same time, the analysis of the raw data does not require sophisticated software and bioinformatics algorithms, since assembly is often not necessary [17]. The most prominent disadvantage of TGS technology compared with NGS is the production of datasets with consistently higher error rates during basecalling. However, for error-tolerant applications, such as the sequencing of whole genomes and transcriptomes as well as the identification of full-length mRNA transcripts in single sequencing reads, the long-read sequencing offers incomparable advantages. Moreover, as TGS evolves in terms of sequencing platforms and software for raw signal processing, we could speculate that TGS will constitute the method of choice even for the detection of SNPs and/or epigenetic modifications [18,19]. Therefore, exploiting the benefits of TGS to study the alternative splicing of cancer-related genes would be a major advancement for the implementation of modern transcriptomics in the tailored cancer diagnostics.

A human gene that has not been under investigation for alternative splice variants is the *CDK4* (also known as *CMM3* and *PSK-J3*), which is located on chromosome 12q14.1. *CDK4* belongs to the family of cyclin-dependent kinase (*CDK*) genes, which encode members of CDK protein family. The members of the family are characterized as serine/threonine protein kinases [20], which are activated only after their interaction with specific cyclin subunits, in order to form

complexes that are involved in fundamental cellular processes [21,22]. Previous published work has only focused on the exact function of *CDK4* in the cellular homeostasis, whereas its thorough mRNA structure has been completely neglected. A quick exploration of GenBank® records regarding *CDK4* reveals only a single annotated mRNA transcript, namely *CDK4* v.1 (GenBank® accession number: NM_000075.4). This gene consists of eight exons and seven intervening introns, encoding a polypeptide (CDK4 is.1) of 303 amino acids (aa). The annotated open-reading frame (ORF) involves a start codon that is located 19 nt from the 5′ end of exon 2 and a stop codon that resides 93 nt from the 5′ end of exon 8. The encoded CDK4 protein is a significant regulator of cell division and differentiation in higher eukaryotic cells [23,24]. In terms of structure, the first 1-96 aa residues of the CDK4 protein form the 5-stranded β-sheet N-terminal region, while the remaining residues (97-303 aa) compose the mainly α-helical C-terminal region, resulting in the typical two-lobed structure observed in the vast majority of CDK family members [25]. The N-terminal contains three significant domains, the cyclin-binding site domain (amino acid sequence PISTVRE), the glycine-rich inhibitory element (G-loop) that is present at the most members of CDK family (GXGXXG motif, GVGAYG in CDK4 member) and the K35 region, which is responsible for protein's interactions and is characterized as binding site [25]. The cyclin-binding site is associated with cyclin D1 to form and activate the D/CDK4 complex at specific intervals during the cell cycle [26]. This complex predominantly phosphorylates retinoblastoma tumor-suppressor protein (Rb) to promote dissociation of the latter from several transcription factors and especially E2F, resulting in the expression of E2F targets, which are crucial for DNA replication [27,28]. As a result, the D/CDK4 complex maintains cellular homeostasis by bridging numerous extracellular signaling pathways to the cell cycle [25,29] in order to regulate the G1-S transition. Moreover, the end of N-terminal (residues 93–96), also called hinge, consists of 4 amino acids 'FEHV', bridges the two lobes of the protein, and is conserved in all typical kinases. On the other side, the C-terminal (residues 97–303) contains regions and domains crucial for the activation of the protein. In brief, besides D140 that is connected with the activation site, the amino acids DFG (residues 158–160) and APE (residues 182–184) constitute the motifs that enclose the activation loop. The sequence between the conserved DFG and APE motifs contains the key region for the substrate phosphorylation, also called T-loop (sequence QMALTPVVVTLW), which is responsible for the functionality of the whole protein [22,25,30], especially due to the residue T172 that is required for the enzymatic activity.

Summarizing, CDK4 is a key driver of numerous cell cycle transitions directly affecting cell growth, apoptosis, and angiogenesis. Defective or overactivated CDK4/cyclin D1 pathway may be responsible for enhanced cellular proliferation, suggesting that CDK4 could be a promising target for the development of novel versatile anticancer therapies [20,24]. However, toward that aim, there is still limited evidence regarding their thorough pre-mRNA processing and the actual mature mRNAs that are synthesized under normal and/or pathological conditions. Thus, the identification of novel *CDK4* mRNA transcripts, especially protein-coding ones, is of high significance, as it could lead to the identification of new diagnostic and/or prognostic biomarkers or new promising therapeutic targets [28,31]. Toward this direction, in the present study we meticulously explored and fully characterized a substantial number of previously unknown, widely expressed *CDK4* mRNA transcripts bearing novel splicing events, using an *in-house* developed nanopore sequencing approach.

Results

Alignment of nanopore long-read sequences unveils novel *CDK4* mRNA transcripts

Based on the visualization of the successfully aligned nanopore sequencing reads as well as the obtained analysis results from our *in-house* developed generic splicing tool 'ASDT', all exon/intron boundaries of the annotated *CDK4* mRNA transcript (*CDK4* v.1, NM_000075.4) were confirmed. However, the bioinformatic analysis also unveiled a total of 18 novel splice junctions between annotated exons of *CDK4* (Table 1). In addition, due to the long-read sequencing methodology that was implemented, the derived sequencing reads bearing the novel splice junctions covered the entire cDNA region from the first until the last exon of the gene and therefore did not necessitate any assembly, since they already represented novel full-length *CDK4* mRNA transcripts (Figs S1 and S2). Intriguingly, taking together all the findings, an extraordinary number of 41 novel *CDK4* mRNA transcripts (*CDK4* v.2–v.42) were successfully aligned to the reference human genome and their exon/intron boundaries were fully characterized (Fig. 1). Of note, all the 18 novel splice junctions of the *CDK4* gene that were unveiled by nanopore sequencing were also confirmed with targeted semiconductor sequencing approach (Fig. S3).

Table 1. Summarizing overview of the annotated and novel splice junctions of *CDK4* gene that were detected as well as the number of nanopore and NGS reads covering and therefore verifying each junction.

	Splice junction between known exons	Nanopore sequencing reads confirming each splice site	Ion Torrent sequencing reads confirming each splice site
Annotated	*Exon 1–Exon 2*	62381	94800
	Exon 2–Exon 3	171786	158610
	Exon 3–Exon 4	172207	211454
	Exon 4–Exon 5	128893	207149
	Exon 5–Exon 6	102348	156111
	Exon 6–Exon 7	175501	150030
	Exon 7–Exon 8	199642	168203
Novel	*Exon 1–Exon 3*	10908	10670
	Exon 1–Exon 4	516	655
	Exon 1–Exon 5	4467	6520
	Exon 1–Exon 6	2025	3488
	Exon 1–Exon 7	54	61
	Exon 1–Exon 8	197	895
	Exon 2–Exon 4	115	207
	Exon 2–Exon 5	151	308
	Exon 2–Exon 6	42	73
	Exon 2–Exon 8	15	23
	Exon 3–Exon 5	1329	2142
	Exon 3–Exon 6	84	157
	Exon 3–Exon 8	41	62
	Exon 4–Exon 6	281	567
	Exon 4–Exon 8	114	199
	Exon 5–Exon 7	1343	885
	Exon 5–Exon 8	1062	1642
	Exon 6–Exon 8	2875	3718

As shown by the visualization of the aligned reads with IGV, the identified transcripts can be categorized into two major groups in terms of structure, based on whether they share the annotated initiation codon residing on exon 2 (*CDK4* v.2–v.17) or lack the entire sequence of exon 2, thus utilizing alternative initiation start codons (*CDK4* v.18–v.42).

Structural analysis of novel *CDK4* mRNA transcripts sharing the annotated initiation codon

Nanopore sequencing revealed a total of 16 novel *CDK4* mRNA transcripts (*CDK4* v.2–v.17), which contain the annotated translation start codon located on exon 2. Based on the obtained sequencing reads, these splice variants are derived from novel cassette exon(s) events, and therefore, they lack one or more annotated exons from their respective mRNA sequences. Interestingly, ORF prediction analysis indicated that all these mRNA transcripts, besides *CDK4* v.11 and v.12, have ORFs, and as a result, they are predicted to encode novel protein isoforms (Fig. 2). Furthermore, *CDK4* v.2–v.10 contain both annotated exons 2 and 3, and consequently, the respective protein isoforms share the same N-terminal domain (1-96 aa) with the annotated *CDK4* v.1. On the contrary, the rest ORF-containing transcripts (*CDK4* v.13–v.17) lack the entire sequence of exon 3 and although most of them have ORFs, the deduced protein isoforms possess a differentiated N-terminal region. Additionally, it should be mentioned that since all the novel splicing events of *CDK4* v.2–v.17 are located downstream of exon 2, their respective C-terminal regions demonstrate notable structural variations.

Based on their structure, *CDK4* v.2, v.3, and v.4 can be characterized as the most promising novel mRNA transcripts in terms of encoding new functional protein isoforms. In detail, even though these transcripts have truncated cDNA sequences as compared to the annotated *CDK4* v.1 due to their existing cassette exon events, they share the exact same nucleotide sequence with *CDK4* v.1 until the 3′ end of exon 5 (Fig. 2). Consequently, their predicted protein isoforms (ORF lengths 230aa, 286aa, and 213aa, respectively) contain all the crucial domains of the annotated CDK4 is.1 that are mandatory for its proper functionality, including the glycine-rich loop (GVGAYG), the cyclin-binding site (PISTVRE), the DFG-APE motif, and the activation loop (T-loop).

Furthermore, the identified transcripts *CDK4* v.5 and v.6 are further truncated, since they lack the entire exon 5 (Fig. 2). Both *CDK4* v.5 and v.6 are expected to be protein-coding mRNAs, due to the presence of ORFs in their mRNA sequence (ORF length 181aa and 204aa, respectively). Interestingly, their predicted encoded isoforms include the canonical glycine-rich loop and cyclin-binding site domains, but at the same time, they contain only the DFG motif and a non-canonical phosphorylation domain, which is caused by the absence of the coding exon 5 (Fig. S4). Four additional novel *CDK4* mRNA transcripts (*CDK4* v.7–v.10) lacking the entire exon 4 were identified. These mRNAs are produced by alternative splicing of exon 3 with distant exons and all of them have ORFs, thus being expected to encode new isoforms (Fig. 2). However, once again the absence of coding exon 4 suggests that the expected isoforms have defective DFG-APE motifs as well as a noncanonical phosphorylation domain.

Next, the new transcript *CDK4* v.11 is the only identified transcript in which the alternative splicing of exon 2 with exon 4 is present. This novel splicing

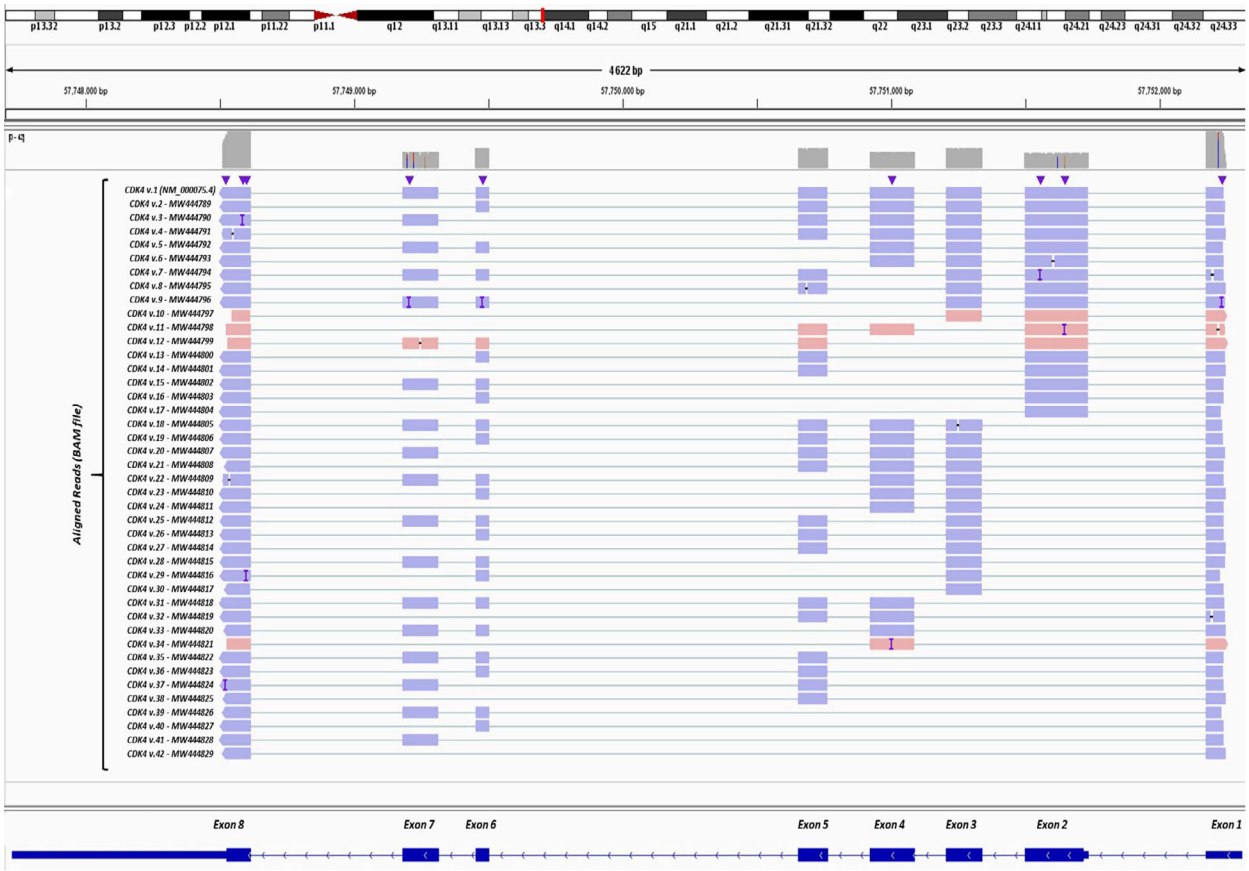


Fig. 1. Visualization of the mapped nanopore sequencing reads representing the annotated (*CDK4* v.1) and the identified novel *CDK4* mRNA transcripts (*CDK4* v.2–v.42) that comprise new exon/intron boundaries, using IGV. The loaded BAM file was derived from the alignment of the spliced long-read sequences to the human reference genome (GRCh38) with Minimap2 aligner. For visual reasons, two colors are used to demonstrate the direction of each aligned sequencing read (Plus strand: orange, Minus strand: cyan). The variant number of each identified transcript as well as the GenBank® accession number is shown next to the respective mapped sequencing read.

event, however, leads to the production of a premature termination codon (PTC), which strongly suggests that *CDK4* v.11 represents a noncoding RNA, candidate for nonsense-mediated mRNA decay (NMD) pathway. In addition, our findings confirmed three new transcript variants, *CDK4* v.12–v.14, which are characterized by the simultaneous absence of both exons 3 and 4 (Fig. 2). Besides *CDK4* v.12 that also contains a PTC, the existence of two distinct ORFs to the nucleotide sequences of *CDK4* v.13 and v.14 directly supports that they are actually protein-coding mRNAs, although the predicted encoded proteins completely lack both the DFG-APE motifs and the activation loop domain. Another isoform with similar characteristics is expected to be encoded from the next identified protein-coding transcript, *CDK4* v.15, although its nucleotide sequence is further truncated (Fig. S4).

Finally, our results support the existence of two additional mRNA transcripts (*CDK4* v.16 and v.17) that are characterized by significantly truncated mRNA sequences, bearing alternative splicing of exon 2 with distal exons of the gene (Fig. 2). It should be mentioned although, that due to their shorter nucleotide sequences, both *CDK4* v.16 and v.17 contain small open-reading frames (smORFs) with lengths less than 100 codons (92aa and 75aa, accordingly), and therefore, their protein-coding capacity as well as the functionality of the potentially encoded protein merit further investigation.

Structural analysis of novel *CDK4* mRNA transcripts with alternative initiation start codons

Besides the aforementioned findings, the applied targeted long-read sequencing methodology also unveiled a total of 13 mRNA transcripts (*CDK4* v.18–v.30) comprising the novel splice junction between exons 1

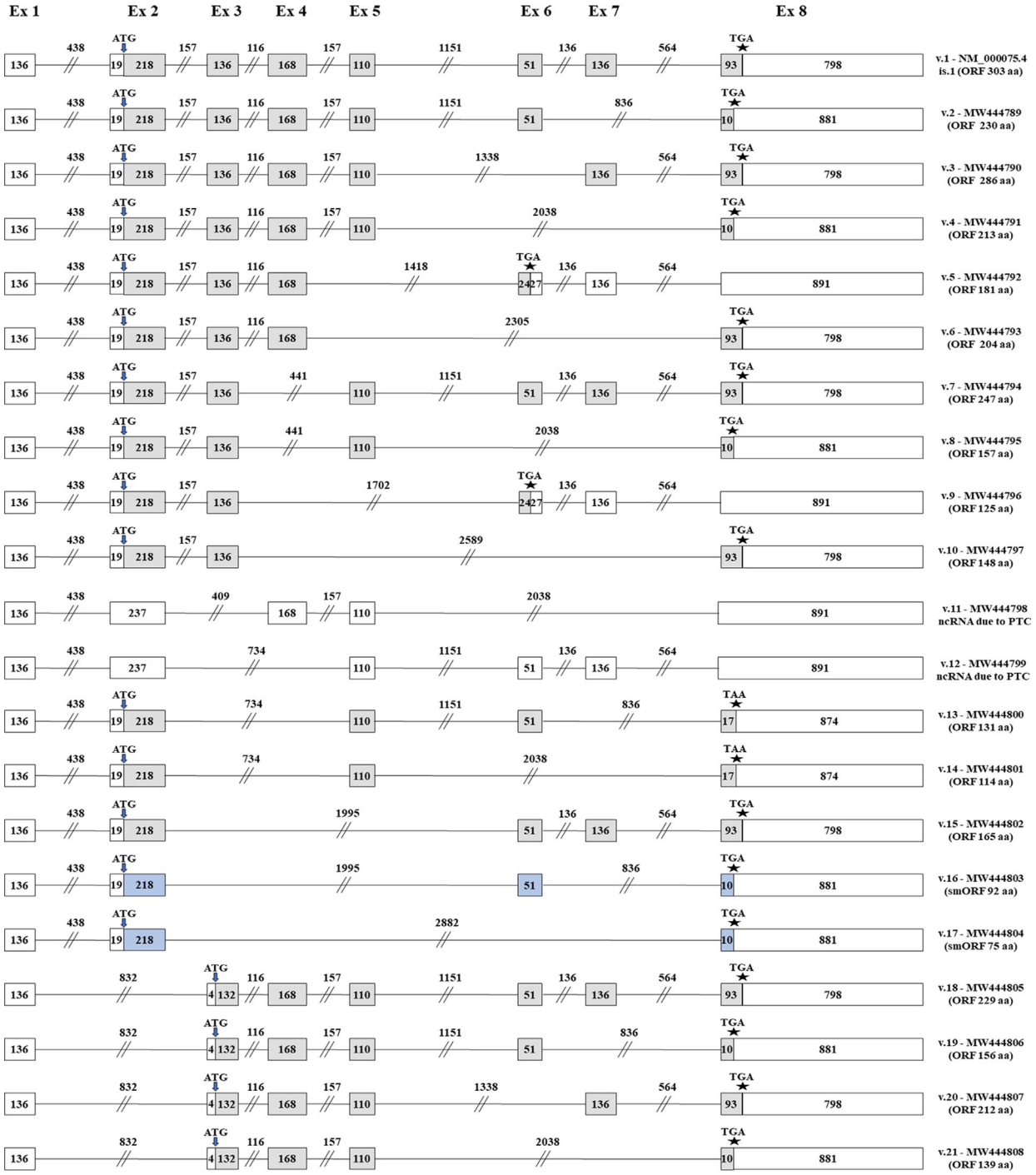


Fig. 2. Structural demonstration of the presented novel transcripts *CDK4* v.2–v.21. Exons are depicted as boxes and introns as lines; gray boxes represent the coding sequences of a transcript that contains an ORF, whereas white boxes correspond to the noncoding regions of each transcript. Blue boxes are used to represent coding sequences that lead to small ORFs (<100 aa). Numbers inside boxes and above lines indicate the length of each exon or intron in nucleotides. Arrows exhibit the position of the initiation codon (ATG), while asterisks (*) indicate the position of the stop codon. The splice variant number, the ORF length (only for protein-coding transcripts), and the GenBank® accession number are demonstrated next to each transcript.

and 3, thus completely lacking exon 2 that encompasses the annotated initiation codon (Fig. 3). In detail, apart from *CDK4* v.18 that lacks exon 2 but contains all the other exons of the gene, the rest 12 splice variants are derived from multiple simultaneous exon skipping events, besides the skipping of exon 2, leading to notably truncated alternative transcript variants as compared to the annotated *CDK4* v.1. Furthermore, it should be noted that for each *CDK4* v.18–v.30 mRNA transcript, all the described exon skipping events that were detected downstream of exons 1/3 splicing junction were also present in the cDNA sequences of the previously described transcripts that are characterized by the annotated initiation start codon (*CDK4* v.2–v.17) and no additional splicing event was detected. At the protein level, since the annotated start codon residing in exon 2 is absent from *CDK4* v.18–v.30, ORF query indicated that the most prevalent initiation codon resides in exon 3 and more specifically to the 5th nt from its 5′ end (Fig. 3). Provided that this alternative initiation codon is utilized for translation, the currently described transcript variants, besides *CDK4* v.28 and v.29, have ORFs, and hence, they are most likely protein-coding mRNAs. However, the predicted isoforms are expected to demonstrate notable structural variations with CDK4 is.1, due to the absence of the coding sequence of exon 2, which encodes 74 aa of the total 96 aa of the N-terminal regions of CDK4 is.1. Consequently, the deduced protein isoforms share only the last 22 aa of the curated N-terminal region that is encoded by exon 3, and hence, they lack both the canonical glycine-rich loop and the cyclin-binding site (Fig. S4).

Furthermore, we identified 12 novel *CDK4* mRNA transcripts (*CDK4* v.31–v.42), which according to the acquired experimental nanopore sequencing reads comprise new alternative splicing events between distant exons of the gene. Specifically, four of these transcripts (*CDK4* v.31–v.34) lack both exons 2 and 3 from their nucleotide sequences, since exon 1 is alternatively spliced with exon 4. Similarly, four additional transcripts (*CDK4* v.35–v.38) are characterized by the alternative splicing between exons 1 and 5, which involves the simultaneous skipping of three consecutive exons (exons 2, 3, and 4). Finally, we identified 4 new *CDK4* mRNA transcripts with significantly truncated cDNA sequences. In brief, *CDK4* v.39 and v.40 share the new alternative splicing event between exons 1 and 6, while *CDK4* v.41 involves the splicing of exon 1 with exon 7 and *CDK4* v.42 is produced from the direct splicing of exon 1 with the last exon of the gene, exon 8.

Expression profile of the novel *CDK4* mRNA transcripts

Based on the results from the applied quantitative PCR (qPCR) assays, the novel *CDK4* mRNA transcripts described in the current study demonstrate a wide expression pattern, being expressed in most of the human tissues that were investigated. Of note, the presented transcripts are not only expressed in human malignancies with reportedly high *CDK4* expression, since their mRNAs were also detected in HEK-293 cell line (Fig. 4). Additionally, even though all transcripts are detected in most tissues, their expression levels are notably differentiated. In detail, our qPCR analysis supports that the novel transcript *CDK4* v.7 is the most overexpressed mRNA in the 16 human malignancies that were investigated, whereas transcripts *CDK4* v.37 and v.41 are hardly detectable in most tissues (Figs 4 and 5). Furthermore, transcripts *CDK4* v.2, v.10, and v.17 are among the most overexpressed mRNAs along with v.7 in all cancer types and normal embryonic kidney.

However, it should be mentioned that due to the significant number of the newly identified transcripts and to the subtle differences in the splicing events between them, in some cases the specific amplification of a single novel *CDK4* transcript with a single qPCR or PCR assay is rather impossible. Consequently, two pairs of transcripts, v.4 / v.8, and v.21 / v.27 are co-amplified with the same pair of primers, and therefore, their expression levels are not discriminated with qPCR. The same limitation also occurs in two sets of four transcripts, *CDK4* v.18, v.22, v.25, and v.28 and *CDK4* v.19, v.23, v.26, and v.29 (Figs 4 and 5). The four transcripts of each set are also co-amplified, and therefore, their separate quantification levels remain unclear. Finally, it should be mentioned that even though their discrimination can be achieved with nested PCR, this approach will introduce bias in the interpretation of each variant quantification level.

Discussion

In the present study, we developed a comprehensive TGS-based methodology using the nanopore sequencing platform MinION Mk1C, aiming the in-depth detection and identification of novel *CDK4* mRNA transcripts. The obtained datasets resulted in the identification of 41 *CDK4* mRNA transcripts originating from multiple combinations of exon skipping events, which are the most dominant alternative splicing mechanism [32]. Notably, despite our research was mainly focused on the detection of

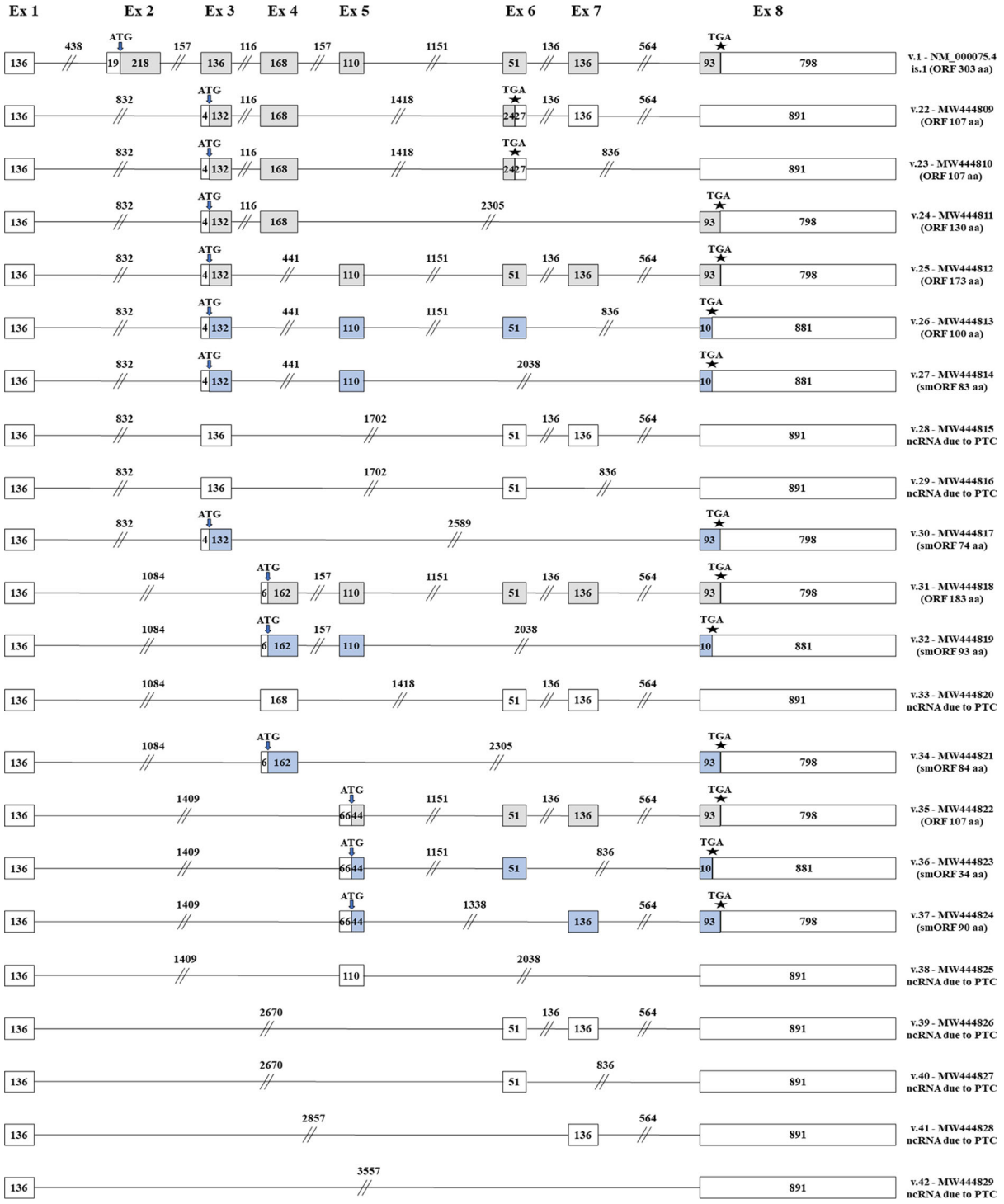


Fig. 3. Structural demonstration of the presented novel transcripts *CDK4* v.22–v.42. Exons are depicted as boxes and introns as lines; gray boxes represent the coding sequences of a transcript that contains an ORF, whereas white boxes correspond to the noncoding regions of each transcript. Blue boxes are used to represent coding sequences that lead to small ORFs (<100 aa). Numbers inside boxes and above lines indicate the length of each exon or intron in nucleotides. Arrows exhibit the position of the initiation codon (ATG), while asterisks (*) indicate the position of the stop codon. The splice variant number, the ORF length (only for protein-coding transcripts), and the GenBank® accession number are demonstrated next to each transcript.

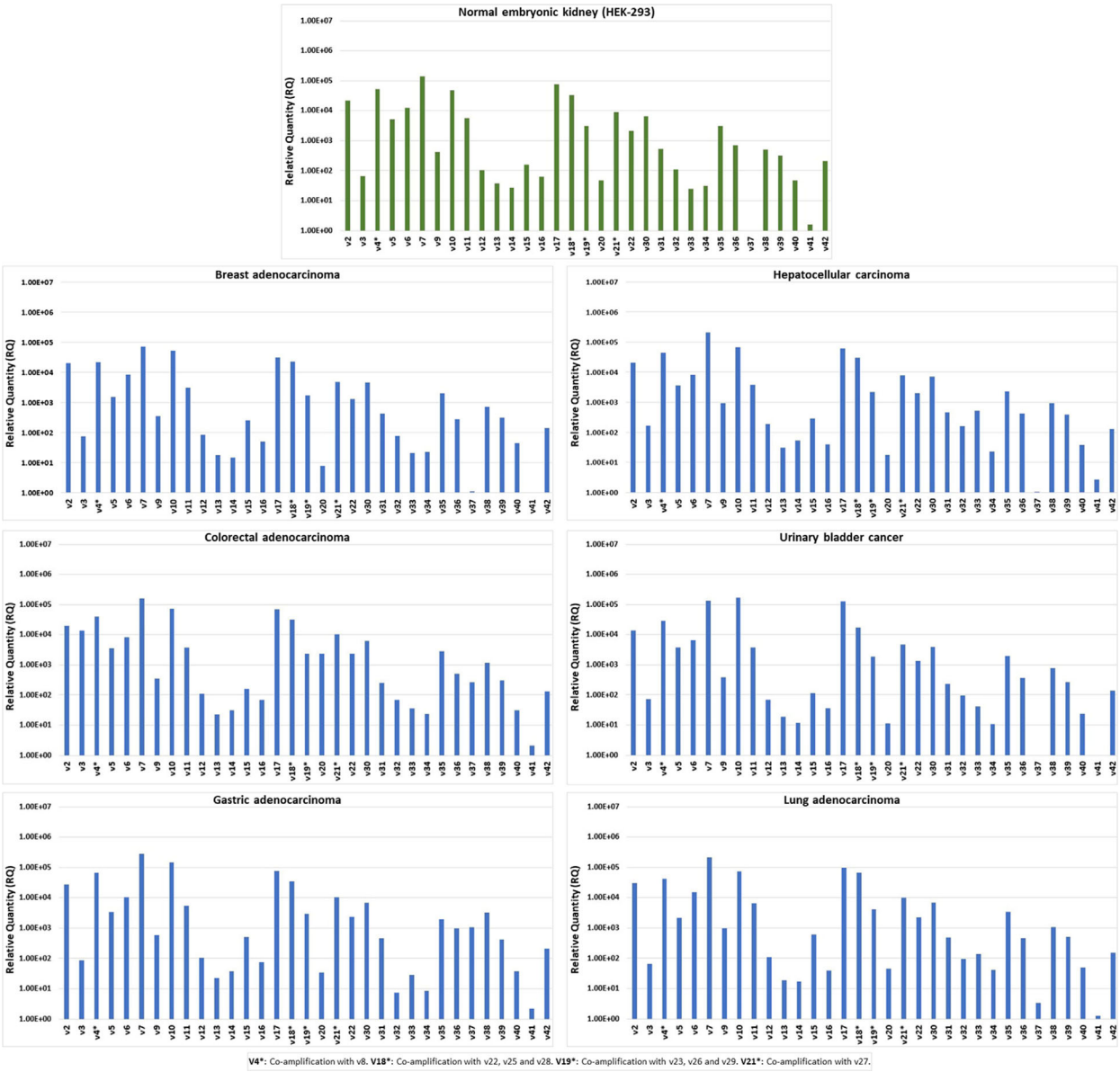


Fig. 4. Barplots demonstrating the relative expression levels of each novel *CDK4* transcript using qPCR in the normal embryonic kidney cell line HEK-293 as well as in six human malignancies with reportedly high *CDK4* expression. The expression levels of each transcript (or set of transcripts) were calculated in relevance to the corresponding mRNA expression of the housekeeping gene *GAPDH*. The relative quantity of each novel *CDK4* variant is presented as variant copies / 10^6 *GAPDH* copies (*Y*-axis). The symbol * represents transcripts that are not specifically amplified in the qPCR assay, but are co-amplified with other variants, due to their subtle splicing differences.

novel splicing events occurring during RNA processing, the identification of the presented *CDK4* transcripts was achieved by applying a targeted DNA-seq by ligation assay, instead of using the 'gold standard' direct RNA sequencing application [33]. The implementation of the presented targeted DNA-seq by ligation approach offered a tremendous sequencing depth and amount of biological information for analysis regarding *CDK4*, allowing the thorough identification even of low-frequent yet full-length mRNA transcripts in single reads. On the contrary, the presented novel *CDK4* mRNA transcripts are not expected to be identified with the established direct RNA sequencing application, which

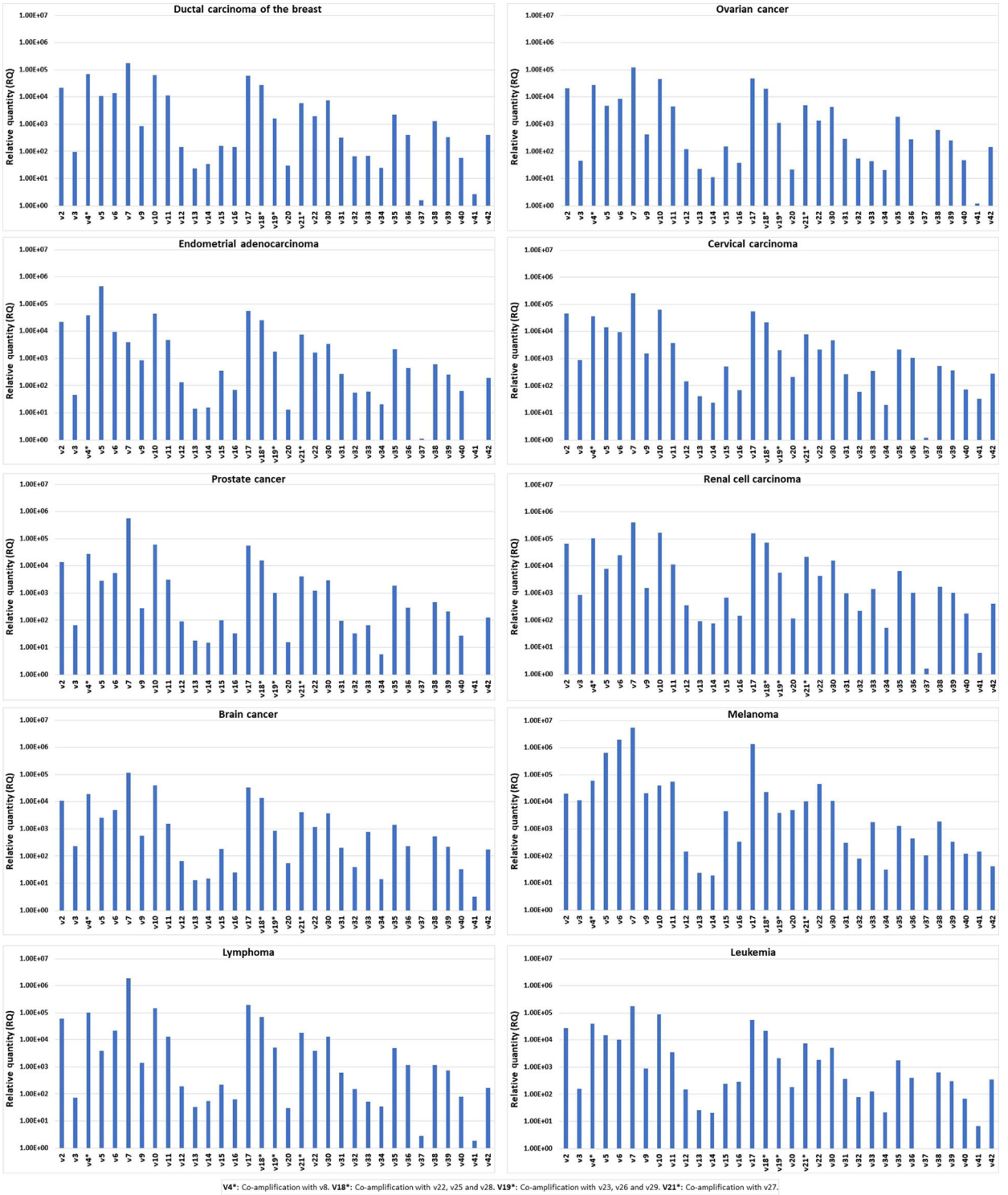


Fig. 5. Barplots demonstrating the relative expression levels of each novel *CDK4* transcript using qPCR in ten human malignancies that were investigated in the current work. The expression levels of each transcript (or set of transcripts) were calculated in relevance to the corresponding mRNA expression of the housekeeping gene *GAPDH*. The relative quantity of each novel *CDK4* variant is presented as variant copies / 10^6 *GAPDH* copies (*Y*-axis). The symbol * represents transcripts that are not specifically amplified in the qPCR assay, but are co-amplified with other variants, due to their subtle splicing differences.

may be the most suitable application for whole transcriptome sequencing and mRNA expression profiling, but fails to identify novel transcripts, due to the seriously decreased sequencing depth in terms of a specific gene. To further validate this conclusion, publicly available RNA-seq experiments from TGS datasets in the Sequence Read Archive (SRA) database were utilized to inquire the potential existence of the novel *CDK4* mRNA transcripts described in the current study. Our *in silico* analysis confirmed that none of the presented transcripts was detected in nanopore direct RNA-seq reads from several human cell lines (e.g., HepG2 and U87), thus highlighting the significant advantage of the presented TGS approach in terms of sensitivity.

Unequivocally, the detection and characterization of the presented *CDK4* mRNA transcripts is not surprising, since most of the members of the human *CDK* gene family are subjected to alternative splicing and hence produce multiple splice variants. The identified *CDK4* splice variants can either be protein-coding mRNAs, thus having the potential to encode new protein isoforms or represent noncoding RNAs, due to the existence of PTCs. For this purpose, *in silico* determination of ORF was performed for each mRNA sequence, based on the existing literature supporting that mRNAs with PTC residing from approximately 50 nt upstream of the last exon junction and anywhere downstream this limit are usually stable [34–36]. According to these criteria, the identified transcripts were categorized in potential protein-coding mRNAs with ORFs, mRNAs with smORFs and therefore ambiguous protein-coding capacity as well as ncRNAs with PTCs (Figs 2 and 3).

The main *CDK4* transcript (CDK4 v.1) encodes the cyclin-dependent kinase 4, a protein that is characterized by all the conserved domains and fundamental features of a typical eukaryotic kinase [37]. Undoubtedly, all the generated *CDK4* mRNA transcripts that contain exon 2 (*CDK4* v.2-v.17) and therefore share the annotated translation start codon have a high potential to encode protein isoforms similar to CDK4 is.1. However, the most promising novel transcripts that fulfill all the required criteria for encoding functional CDK4 isoforms are *CDK4* v.2, v.3, and v.4. This is inferred from the fact that the nucleotide sequences of these mRNA transcripts encode protein isoforms, which maintain all the crucial domains of the annotated CDK4 is.1 (Fig. S4). Based on the determined ORFs for *CDK4* v.2–v.4, the respective novel CDK4 isoforms share an identical to CDK4 is.1 N-terminal region (1-96 aa), and as a result, they most likely possess the kinase catalytic activity, due to the presence of the glycine-rich loop, which is responsible for the guidance of the ATP at the position of phosphorylation. Furthermore, due to the presence of the conserved motif 'PISTVRE', the isoforms encoded by *CDK4* v.2–v.4 have the potential to bind with cyclin D1, thus leading to the formation of CDK4/cyclin D1 complexes, which constitutes the first step in the kinase activation [20,25]. The assumption that *CDK4* v.2–v.4 encode novel isoforms with cyclin-binding and kinase activity is further enhanced by the 3D protein prediction models, which clearly indicate that these protein isoforms exhibit a significant structural resemblance with the annotated CDK4 is.1. In particular, the aforementioned CDK4 isoforms share the typical bilobal structure that is observed not only in CDK4 is.1, but also in other kinases as well (Fig. 6). Taken together all these considerations, it can be assumed that *CDK4* v.2–v.4 have the potential to encode CDK4 isoforms with the same binding affinity to cyclin as CDK4 is.1. Under these circumstances, these putative proteins could develop a stable interaction with cyclin, resulting in the activation of the catalytic activity of D–CDK4 complexes. Also, it should be mentioned that the phosphorylation loop is located within the respective C-terminal regions of *CDK4* v.2–v.4 (Fig. S4), a fact which is crucial for the functionality of the protein since it is implicated in the adoption of the active conformation of DFG motif and regulates the activity of the kinase. The simultaneous activation of these identical proteins in cells may operate as a signal for over activation and cell proliferation, which may lead to carcinogenesis or cancer progression, and therefore, their functional role merits investigation.

Another issue raised by the current study is the functional role of the rest protein-coding *CDK4* mRNAs. For instance, *CDK4* v.5 and v.6 (ORF length 181 and 204 aa, respectively) seem to encode isoforms that include all the domains of the N-terminal region that are critical for cyclin-binding and the typical kinase activity, while based on their 3D prediction models, they share the typical two-lobed structure of the annotated CDK4 is.1 (Fig. 6). However, their activation segment in the C-lobe that spans from the DFG motif to the APE motif and contains the so-called T-loop is expected to be partially disordered due to the complete absence of APE motif (Fig. S4), which strongly suggests that although they may interact with cyclins, their enzymatic activity is lost or significantly reduced. Similar findings are also observed in *CDK4* v.7 and v.8, which exhibit a bilobal structure highly similar to CDK4 is.1, contain the same N-lobe with CDK is.1, but their activating site between DFG and

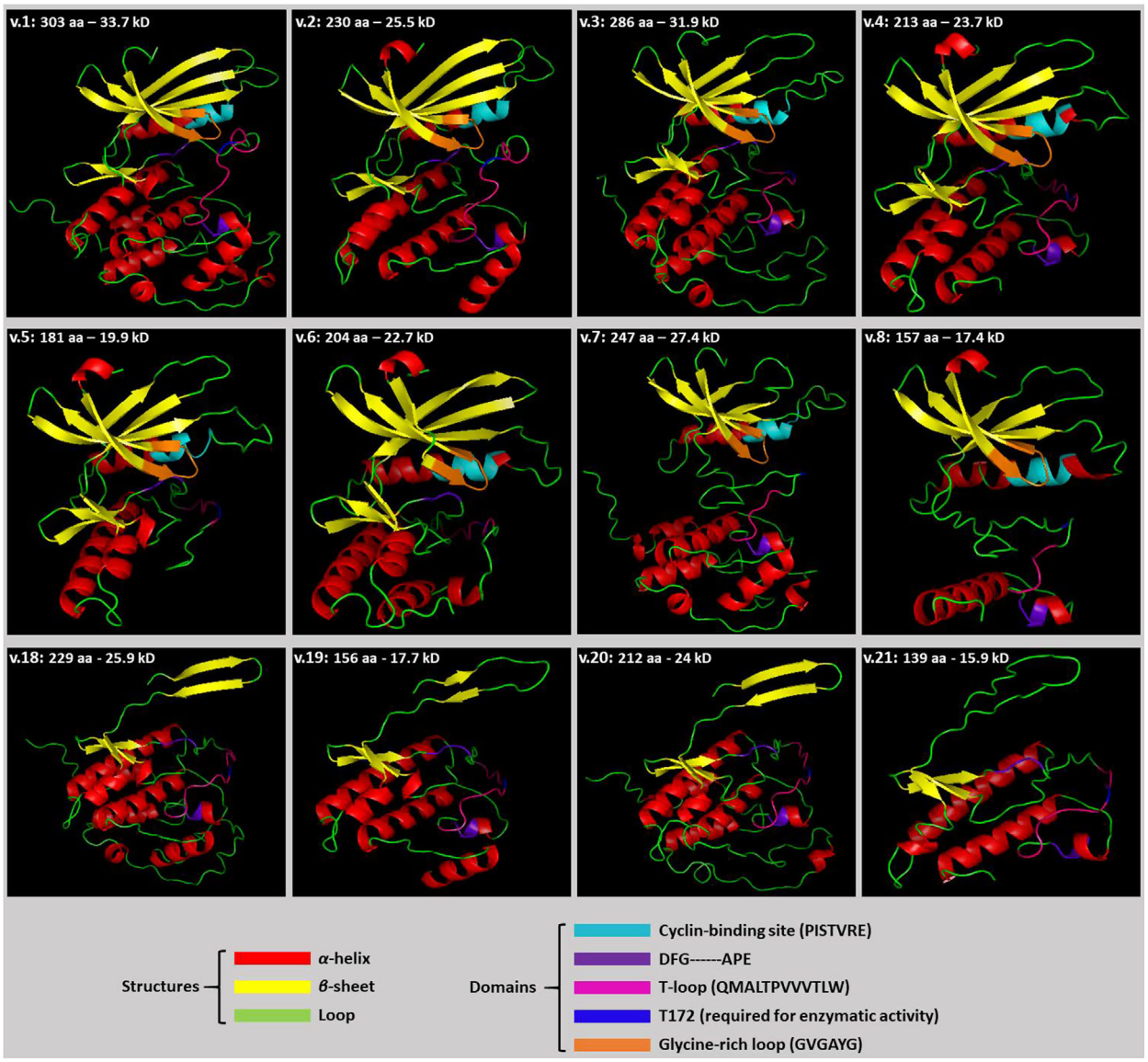


Fig. 6. Predicted 3D structure models of the most promising novel *CDK4* mRNA transcripts in terms of protein-coding capacity. Each domain is demonstrated in different colors for viewing purposes. For each protein isoform, only the 3D structure with the highest confidence score is depicted.

APE motifs is disturbed in terms of integrity (Fig. 6). The rest transcripts sharing the annotated initiation codon at exon 2 present further alternations regarding their corresponding protein structures, and therefore, their functional role is even more complex to be defined. Furthermore, the rest isoforms that are encoded by *CDK4* transcripts lacking exon 2 (e.g., *CDK4* v.18–v.21) do not possess the cyclin-binding site and the glycine-rich inhibitory element, and as illustrated in the 3D prediction models, they exhibit a completely different N-terminal folding (Fig. 6). This incident supports that they are rather incapable of binding to cyclins or possessing kinase activities.

Although our study focuses on the detection of alternative *CDK4* splice variants at the mRNA level, our hypothesis about their protein-coding capacity is in total agreement with previous proteomic studies that have already identified multiple CDK4 isoforms using both western blot and LS-MS/MS methodologies [38]. In detail, Sun *et al.* utilized multiple antibodies including sc-260 and sc-601, which target the C terminus of mouse and human CDK4, respectively, sc-

536636 that specifically binds to 270–290 aa of human CDK4 and sc-23896 that targets the first 20 aa of the protein and performed extensive western blotting experiments in human cell lines, which are also included in the present study. Besides the 33 kD band that corresponds to the CDK4 is.1, western blotting and LC-MS/MS results of this study revealed 3 additional bands related to different CDK4 proteins at 24–28 kD. However, these bands strongly suggested the existence of truncated novel CDK4 proteins, but were characterized as 'unspecified', since there was not any identified nucleotide sequence corresponding to a novel transcript [38]. Nevertheless, findings of the same study revealed that one of these additional bands corresponds to a new 25.9 kD CDK4 isoform that completely lacked the annotated exon 2 as derived from partial transcript sequencing. In our study, *CDK4* v.18 completely lacks exon 2 and has an ORF that encodes a novel CDK4 protein with 25.9 kD, which strongly supports that *CDK4* v.18 encodes the new isoform that corresponds to this specific band. As a result, the sequencing findings of the present work not only are unanimous with the study of Sun *et al*. but also interpret the rest two unexpected bands detected by western blotting, since our study elucidates the existence of protein-coding mRNAs of *CDK4*, which encode truncated proteins as compared to CDK4 is.1, having molecular mass between 22 and 31 kD (Fig. 6). More specifically, the novel transcripts *CDK4* v.2 and v.4 of the current work encode protein isoforms of 25.5 kD and 23.7 kD, respectively, which are predicted to maintain the bilobal structure of CDK4 is.1, and therefore, they are the most promising transcripts to represent the unspecified bands (Fig. 6). In addition, CDK4 v.7 and v.20 are expected to encode protein isoforms between 22 and 28 kD (27.4 kD and 24 kD, accordingly), and therefore, they are strong candidates for being new protein-coding transcripts. Nevertheless, it should be mentioned that due to the subtle alternative splicing events of the *CDK4* transcripts as well as the confirmed transcriptional and protein multiplicity of CDK4, the utilization of any anti-CDK4 antibody is rather incapable of discriminating a specific new CDK4 protein.

Additionally, many of the transcript variants described in the current study contain ORFs with lengths < 100 aa (smORFs). Based on the existing knowledge, many long noncoding RNAs (lncRNAs) have the prospective to encode micropeptides of < 100 aa and their expression depends on the cell type or the stage of development [39,40]. In fact, an increasing number of scientific studies have confirmed that lncRNAs in the human transcriptome outnumber the protein-coding mRNAs, while a significant proportion of lncRNAs harbor small ORFs, which can be translated into micropeptides [41,42]. The balance between noncoding and coding RNA levels is modulated depending on whether they need to be translated into protein or whether they must act as RNA effectors [41]. Besides the identified mRNAs with ORFs or smORFs, our analysis unveiled a total of 10 novel *CDK4* transcripts containing PTCs, which presumably lack any protein-coding potential. This finding is quite surprising, since only a few lncRNAs have been identified in the human *CDK* gene family (e.g., *CDK7* transcript variant 10, GenBank accession number: NR_136690.2). This evidence points toward the idea that these splice variants represent lncRNAs, which are implicated in plethora of cellular processes characterized as mediators of interlinkages between RNA and proteins, gene expression regulators, and booster activators of biological mechanisms [43,44].

Into cytoplasm, CDK4/6 and its binding partner, cyclin D1, form strong complexes that translocate to nucleus. Normally, the cyclin D-CDK4/6 interactions regulate the progression from G1 to S phase during cell proliferation [29]. Thus, monitoring the activation of this complex is a crucial approach in cancer therapy. Palbociclib, ribociclib, and abemaciclib, that are used in clinical practice, are common drugs that disrupt the signaling pathway, mediated by the CDK4/6, by weakening the affinity between the subunits of the protein and inhibiting the assembly of the complex [24,31,45]. However, the confirmed transcriptional and proteomic multiplicity of CDK4 may lead to the generation of CDK4 isoforms that are resistant to CDK4/6 inhibitors, thus reducing the efficiency of the drug and to the translation of isoforms that may play catalytic role in tumor development. Moreover, NPCD, another CDK4/6 inhibitor, has the potential to regulate cell proliferation and apoptosis by mediating the activation of CDK4/6. Based on the study of Sun *et al.* that reveal the presence of uncharacterized CDK4 isoforms, NPCD provokes the inactivation of the whole subfamily of CDK4 proteins [38]. Thus, the expression of multiple CDK4 proteins may limit the therapeutic efficiency of CDK4/6 inhibitors by altering the specificity of binding to their target.

Summarizing, we have developed a cutting-edge sequencing approach to unveil the previously unknown aspects of the human *CDK4* gene, by exploiting the cutting-edge technology of nanopore sequencing. Our study elucidates for the first time the complex transcriptional landscape of the human *CDK4* gene and tries to provide explanations for the fate of the multiple mRNAs that are produced by

alternative splicing in the eukaryotic cells. The presented wide spectrum of *CDK4* alternative spliced variants is only the first step to distinguish and assemble the missing pieces regarding the exact functions and implications of this fundamental kinase in cellular homeostasis and pathology.

Materials and methods

Cell culture and total RNA extraction

The present work was carried out using an established panel of 52 human cell lines that originate from 17 distinct human tissues and were the following: MCF-7, SK-BR-3, BT-20, MDA-MB-231, MDA-MB-468 (breast adenocarcinoma), BT-474, T-47D, ZR-75-1 (ductal carcinoma of the breast), OVCAR-3, SK-OV-3, ES-2, MDAH-2774 (ovarian cancer), Ishikawa, SK-UT-1B (endometrial adenocarcinoma), HeLa, SiHa (cervical carcinoma), PC-3, DU 145, LNCaP (prostate cancer) T24, RT4 (urinary bladder cancer), ACHN, 786-O, Caki-1 (renal cell carcinoma), Caco-2, DLD-1, HT-29, HCT 116, SW 620, COLO 205, RKO (colorectal cancer), AGS (gastric adenocarcinoma), HepG2, HuH-7 (hepatocellular carcinoma), U87 MG, U-251 MG, D54, H4, SH-SY5Y (brain cancer), A549 (lung adenocarcinoma), FM3, MDA-MB-435S (melanoma), Raji, Daudi, U-937 (lymphoma), K-562, HL-60, Jurkat, REC-1, SU-DHL-1, GRANTA-519 (leukemia), and HEK-293 (normal embryonic kidney). All the above cell lines were cultured based on the American Type Culture Collection (ATCC) protocols.

The TRIzol Reagent (Ambion™, Thermo Fisher Scientific Inc., Waltham, MA, USA) was employed for total RNA extraction from each human cell line. All total RNA samples were appropriately diluted in THE RNA Storage Solution (Ambion™), while the assessment of their concentration and purity was performed spectrophotometrically at 260 and 280 nm, using a BioSpec-nano Micro-volume UV-Vis Spectrophotometer (Shimadzu, Kyoto, Japan).

First-strand cDNA synthesis

An amount of 2 μg total RNA from each human cell line was used as template for the reverse transcription reaction. First-stranded cDNA synthesis was carried out using an oligo-dT-adapter as RT primer, designed to anneal in the 3′ poly(A) tail of the mRNA transcripts. The nucleotide sequence of the oligo-dT-adapter was the following: 5′-GCGAGCACAGAATTAATACGACTCACTATAGGTT TTTTTTTTTVN-3′ (where V = G, A, C and N = G, A, T, C). Briefly, for each RT reaction an initial 4.5 μL cDNA synthesis mixture containing 2 μL total RNA (2 μg), 1 μL oligo-dT-adapter (10 μM), and 1.5 μL deionized H_2O was incubated in a hot-lid Veriti™ 96-Well Fast Thermal Cycler (Applied Biosystems, Waltham, MA, USA) at 72 °C for 3 min and 42 °C for 2 min and then immediately placed on ice.

Then, the total cDNA synthesis mixture volume was adjusted to 10 μL, by adding the following reagents: 1 μL nuclease-free H_2O, 2 μL 5X First-Stranded Buffer, 0.25 μL DTT (100 mM), 1 μL dNTP mix (10 mM each), 0.25 μL (10 U) RNaseOUT™ inhibitor (Invitrogen™, Thermo Fisher Scientific Inc.), and 1 μL (100 U) SMARTScribe™ Reverse Transcriptase (Takara Bio Inc., Otsu, Shiga, Japan). The RT reaction was carried out at 42 °C for 90 min in the thermal cycler. Finally, the reaction was terminated by heating the mixture at 70 °C for 10 min. The quality control of the produced cDNAs was assessed using the human glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) gene as housekeeping. The 52 created cDNA samples were diluted 1 : 10 in nuclease-free H_2O and were properly mixed to generate 17 cDNA pools, based on the tissue of origin/type of malignancy as described previously. These cDNA pools were used as templates for the downstream PCRs.

Specific amplification of *CDK4* mRNA transcripts

For the specific amplification of *CDK4* mRNA transcripts, we developed and employed a touchdown PCR-based assay, which was performed using the 17 cDNA pools as templates, aiming to increase both sensitivity and specificity as well as the PCR yield [46]. For this purpose, two gene-specific primers for *CDK4* were designed using the primer-BLAST designing tool [47]. In brief, a forward gene-specific primer (Ex1F: 5′-GTGTATGGGGCCGTAGGAAC-3′) was designed to target the first exon of *CDK4*, while the reverse gene-specific primer (Ex8R: 5′-AGCCACTCCATTGCTCACTC-3′) was designed to anneal in the last annotated exon (exon 8) and more specifically at the annotated translation termination codon. In detail, the applied assay was performed in reaction volumes of 25 μL that contained KAPA Taq Buffer A (Kapa Biosystems, Inc., Woburn, MA, USA) including $MgCl_2$ at a final concentration of 1.5 mM, 0.2 mM dNTP mix, 0.4 μM of each primer, and 1 U of KAPA Taq DNA Polymerase (Kapa Biosystems Inc.), in a Veriti™ 96-Well Fast Thermal Cycler (Applied Biosystems™). In addition, the cycling protocol of the touchdown PCR was the following: an initial denaturation step at 95 °C for 3 min, followed by 35 cycles of 95 °C for 30 s, 65 °C (auto-ΔTa: −0.3 °C/cycle) for 30 s, 72 °C for 2 min, and a final extension step at 72 °C for 2 min. The assessment of the applied PCR assay was performed with electrophoresis of the derived PCR products in agarose gels.

The derived amplicons were mixed to generate a final PCR product, which was purified with the NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel GmbH & Co. KG, Duren, Germany). The purified PCR product was the starting material for the subsequent DNA-seq library preparation. This approach enabled the simultaneous identification of amplified *CDK4* transcripts from all 52 cell lines that were investigated.

Targeted nanopore sequencing

An initial amount of 1μg purified PCR product corresponding to the amplified *CDK4* mRNA transcripts was used as input for the targeted DNA-seq library preparation workflow. Nanopore sequencing was carried out on a MinION Mk1C sequencer (Oxford Nanopore Technologies Ltd., Oxford, UK), using a FLO-MIN106D flow cell with R9.4.1 chemistry and the Ligation Sequencing Kit (SQK-LSK109, ONT) following the manufacturer's instructions. Briefly, the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Inc., Ipswich, MA, USA) was employed for the end repair process, and the Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) were used for the nucleic acid purification steps, whereas the Quick T4 Ligase (New England Biolabs, Inc) enabled the adapter ligation. The library was sequenced in a 3 h single run, leading to the generation of 1.78 million reads with a median length of 0.8 kb.

Validation of nanopore findings with NGS

To ensure that the obtained sequencing data were reproducible, an NGS based on a semiconductor sequencing technology was applied using the same starting material, thus enabling the direct comparison of nanopore and NGS datasets. The Ion Xpress™ Plus Fragment Library Kit (Invitrogen™, Thermo Fisher Scientific Inc.) was employed for the preparation of the DNA-seq library, using 1 μg of purified PCR product mix as input. Enzymatic fragmentation, adapter ligation, nick-repair, and purification of the ligated DNA were implemented following the manufacturer's protocol. Bead-based size selection of the created DNA-seq library was performed using the KAPA Pure Beads (Kapa Biosystems Inc.) in the recommended ratio of fragmented dsDNA:beads to enrich the library for 300-400 bp fragments. The size-selected library was quantified using the Ion Library TaqMan™ Quantitation Kit (Invitrogen™, Thermo Fisher Scientific Inc.) in an ABI 7500 Fast Real-Time PCR System (Applied Biosystems™). The sequencing template was created with emulsion PCR on an Ion OneTouch™ 2 System (Invitrogen™, Thermo Fisher Scientific Inc.), using the Ion PGM™ Hi-Q™ View OT2 kit (Invitrogen™, Thermo Fisher Scientific Inc.), strictly based on the instructions of the manufacturer. Next, the Ion OneTouch ES™ instrument (Invitrogen™, Thermo Fisher Scientific Inc.) was used for the template enrichment. Ultimately, semiconductor sequencing methodology was carried out in an Ion 316™ Chip v2 using an Ion PGM™ System (Invitrogen™, Thermo Fisher Scientific Inc.) and the Ion PGM™ Hi-Q™ View Sequencing kit.

Post-processing and bioinformatic analysis

The primary data analysis of nanopore sequencing, including basecalling, adapter trimming, and quality assessment, was performed with Guppy [48]. Nanopore sequencing reads were separated into the 'pass' and 'fail' folder, based on their quality scores. Only the sequencing reads existing in the 'pass' folder were used in the downstream analyses. As a next step, the generated FASTQ files containing the nanopore raw sequencing data were aligned to the human reference genome (GRCh38), using the general-purpose Minimap2 aligner [49], whose parameters were adapted to perform spliced alignment. The aligned spliced long reads were included in the generated SAM file, which was converted to BAM file, using the SAMtools program [50]. Mapped sequencing reads were visualized with the Integrative Genomics Viewer (IGV) software for the detection of the splice acceptor and donor sites [51]. Besides mapping with Minimap2, the detection of alternative splicing events in the created FASTQ file was also implemented with the *in-house* developed algorithm 'ASDT', which was designed by members of our group as a generic splicing tool capable of identifying alternative splicing events and cryptic exons from high-throughput sequencing datasets [52].

mRNA expression analysis with quantitative RT-PCR

To investigate the expression levels of the identified *CDK4* mRNA transcripts in the established panel of human cell lines, SYBR-Green fluorescent-based qPCR assays were developed and implemented using variant-specific primers (Table S1). Each primer pair was designed to specifically target and therefore amplify a single novel *CDK4* transcript variant (Table S2). Additionally, the expression analysis with qPCR-based assays was performed using the previously generated 17 cDNA pools as templates. Consequently, the expression profile of each *CDK4* transcript was elucidated in breast adenocarcinoma, ductal carcinoma of the breast, ovarian cancer, endometrial adenocarcinoma, cervical carcinoma, prostate cancer, urinary bladder cancer, renal cell carcinoma, colorectal cancer, gastric adenocarcinoma, hepatocellular carcinoma, brain cancer, lung adenocarcinoma, melanoma, lymphoma, leukemia, and normal embryonic kidney. Finally, the human *GAPDH* mRNA was utilized as the endogenous reference control for normalization purposes.

All RT-qPCR assays were conducted on a 7500 Fast Real-Time PCR System (Applied Biosystems™) and were implemented in 10-μL reactions, which contained 5 μL of the 2X Kapa SYBR® Fast qPCR Master Mix (Kapa Biosystems, Inc.), 1 μL of each primer (2 μM), and 1 μL of cDNA template. The applied thermal protocol included an initial denaturation step at 95 °C for 3 min, which was followed by 40 cycles of 95 °C for 3 s and 60 °C for 30 s.

Visualization and functional prediction of the deduced novel CDK4 protein isoforms

All the identified *CDK4* mRNA transcripts were tested whether they are characterized by ORF, thus having the potential to

encode novel protein isoforms, or represent NMD candidates. In order to determine the putative amino acid sequence encoded by each transcript variant, we used ExPASy, an online tool allowing the translation of a nucleotide (DNA/RNA) sequence to a protein sequence [53]. Finally, for the mRNA transcripts predicted to encode protein isoforms, 3D structure models were generated using the I-TASSER server, an online tool for 3D structure model construction [54,55], and were visualized with PyMOL molecular visualization system for their thorough investigation.

Acknowledgements

This work was supported by the Bodossaki Foundation (Athens, Greece) with a postdoctoral fellowship to Dr. Panagiotis G. Adamopoulos. The research presented was carried out within the framework of a Stavros Niarchos Foundation grant to the National and Kapodistrian University of Athens (grant ID 16785).

Conflict of interest

The authors declare no conflict of interest.

Author contributions

PGA conceptualized the study, contributed to methodology and bioinformatic analysis, and prepared the original draft. KA contributed to methodology, curated the data, and prepared the original draft. PT contributed to bioinformatic analysis and curated the data. AS conceptualized, supervised, and critically reviewed the study.

Data accessibility

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

1 Thapar R, Bacolla A, Oyeniran C, Brickner JR, Chinnam NB, Mosammaparast N & Tainer JA (2019) RNA modifications: reversal mechanisms and cancer. *Biochemistry* **58**, 312–329.
2 Gott JM & Emeson RB (2000) Functions and mechanisms of RNA editing. *Annu Rev Genet* **34**, 499–531.
3 Nilsen TW & Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**, 457–463.
4 Black DL (2000) Protein diversity from alternative splicing: a challenge for bioinformatics and post-genome biology. *Cell* **103**, 367–370.
5 Adamopoulos PG, Tsiakanikas P, Adam EE & Scorilas A (2020) Unraveling novel survivin mRNA transcripts in cancer cells using an in-house developed targeted high-throughput sequencing approach. *Genomics* **113**, 573–581.
6 Wang E & Aifantis I (2020) RNA splicing and cancer. *Trends Cancer* **6**, 631–644.
7 Pal S, Gupta R & Davuluri RV (2012) Alternative transcription and alternative splicing in cancer. *Pharmacol Ther* **136**, 283–294.
8 Ghigna C & Paronetto MP (2020) Alternative splicing: recent insights into mechanisms and functional roles. *Cells* **9**, 2327.
9 Bonnal SC, Lopez-Oreja I & Valcarcel J (2020) Roles and mechanisms of alternative splicing in cancer - implications for care. *Nat Rev Clin Oncol* **17**, 457–474.
10 Ladomery M (2013) Aberrant alternative splicing is another hallmark of cancer. *Int J Cell Biol* **2013**, 463786.
11 Wahl MC, Will CL & Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701–718.
12 Pages A, Dotu I, Pallares-Albanell J, Marti E, Guigo R & Eyras E (2018) The discovery potential of RNA processing profiles. *Nucleic Acids Res* **46**, e15.
13 Ge P & Zhang S (2015) Computational analysis of RNA structures with chemical probing data. *Methods* **79–80**, 60–66.
14 Ule J & Blencowe BJ (2019) Alternative splicing regulatory networks: functions, mechanisms, and evolution. *Mol Cell* **76**, 329–345.
15 Gong L, Wong CH, Cheng WC, Tjong H, Menghi F, Ngan CY, Liu ET & Wei CL (2018) Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat Methods* **15**, 455–460.
16 van Dijk EL, Jaszczyszyn Y, Naquin D & Thermes C (2018) The third revolution in sequencing technology. *Trends Genet* **34**, 666–681.
17 Midha MK, Wu M & Chiu KP (2019) Long-read sequencing in deciphering human genetics to a greater depth. *Hum Genet* **138**, 1201–1215.
18 Xiao T & Zhou W (2020) The third generation sequencing: the advanced approach to genetic diseases. *Transl Pediatr* **9**, 163–173.
19 Yang Y, Sebra R, Pullman BS, Qiao W, Peter I, Desnick RJ, Geyer CR, DeCoteau JF & Scott SA (2015) Quantitative and multiplexed DNA methylation analysis using long-read single-molecule real-time bisulfite sequencing (SMRT-BS). *BMC Genom* **16**, 350.
20 Wood DJ & Endicott JA (2018) Structural insights into the functional diversity of the CDK-cyclin family. *Open Biol* **8**, 180112.
21 Bockstaele L, Kooken H, Libert F, Paternot S, Dumont JE, de Launoit Y, Roger PP & Coulonval K (2006) Regulated activating Thr172 phosphorylation of cyclin-dependent kinase 4(CDK4): its relationship with

cyclins and CDK "inhibitors". *Mol Cell Biol* **26**, 5070–5085.
22 Kato JY, Matsuoka M, Strom DK & Sherr CJ (1994) Regulation of cyclin D-dependent kinase 4 (cdk4) by cdk4-activating kinase. *Mol Cell Biol* **14**, 2713–2721.
23 Malumbres M, Sotillo R, Santamaria D, Galan J, Cerezo A, Ortega S, Dubus P & Barbacid M (2004) Mammalian cells cycle without the D-type cyclin-dependent kinases Cdk4 and Cdk6. *Cell* **118**, 493–504.
24 Sherr CJ, Beach D & Shapiro GI (2016) Targeting CDK4 and CDK6: from discovery to therapy. *Cancer Discov* **6**, 353–367.
25 Day PJ, Cleasby A, Tickle IJ, O'Reilly M, Coyle JE, Holding FP, McMenamin RL, Yon J, Chopra R, Lengauer C *et al.* (2009) Crystal structure of human CDK4 in complex with a D-type cyclin. *Proc Natl Acad Sci USA* **106**, 4166–4170.
26 Sherr CJ (1994) G1 phase progression: cycling on cue. *Cell* **79**, 551–555.
27 Shao Z & Robbins PD (1995) Differential regulation of E2F and Sp1-mediated transcription by G1 cyclins. *Oncogene* **10**, 221–228.
28 VanArsdale T, Boshoff C, Arndt KT & Abraham RT (2015) Molecular pathways: targeting the cyclin D-CDK4/6 axis for cancer treatment. *Clin Cancer Res* **21**, 2905–2910.
29 Klein ME, Kovatcheva M, Davis LE, Tap WD & Koff A (2018) CDK4/6 inhibitors: the mechanism of action may not be as simple as once thought. *Cancer Cell* **34**, 9–20.
30 Russo AA, Jeffrey PD & Pavletich NP (1996) Structural basis of cyclin-dependent kinase activation by phosphorylation. *Nat Struct Biol* **3**, 696–700.
31 Murphy CG & Dickler MN (2015) The Role of CDK4/6 inhibition in breast cancer. *Oncologist* **20**, 483–490.
32 Chen M & Manley JL (2009) Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**, 741–754.
33 Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I & Wilson AC (2019) Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* **10**, 754.
34 Thermann R, Neu-Yilik G, Deters A, Frede U, Wehr K, Hagemeier C, Hentze MW & Kulozik AE (1998) Binary specification of nonsense codons by splicing and cytoplasmic translation. *EMBO J* **17**, 3484–3494.
35 Zhang J, Sun X, Qian Y, LaDuca JP & Maquat LE (1998) At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mRNA: a possible link between nuclear splicing and cytoplasmic translation. *Mol Cell Biol* **18**, 5272–5283.
36 Zhang J, Sun X, Qian Y & Maquat LE (1998) Intron function in the nonsense-mediated decay of beta-globin mRNA: indications that pre-mRNA splicing in the nucleus can influence mRNA translation in the cytoplasm. *RNA* **4**, 801–815.
37 Kanev GK, de Graaf C, de Esch IJP, Leurs R, Wurdinger T, Westerman BA & Kooistra AJ (2019) The landscape of atypical and eukaryotic protein kinases. *Trends Pharmacol Sci* **40**, 818–832.
38 Sun Y, Lou X, Yang M, Yuan C, Ma L, Xie BK, Wu JM, Yang W, Shen SX, Xu N *et al.* (2013) Cyclin-dependent kinase 4 may be expressed as multiple proteins and have functions that are independent of binding to CCND and RB and occur at the S and G 2/M phases of the cell cycle. *Cell Cycle* **12**, 3512–3525.
39 Choi SW, Kim HW & Nam JW (2019) The small peptide world in long noncoding RNAs. *Brief Bioinform* **20**, 1853–1864.
40 Robinson EK, Covarrubias S & Carpenter S (2020) The how and why of lncRNA function: An innate immune perspective. *Biochim Biophys Acta Gene Regul Mech* **1863**, 194419.
41 Dhamija S & Menon MB (2018) Non-coding transcript variants of protein-coding genes - what are they good for? *RNA Biol* **15**, 1025–1031.
42 Ransohoff JD, Wei Y & Khavari PA (2018) The functions and unique features of long intergenic non-coding RNA. *Nat Rev Mol Cell Biol* **19**, 143–157.
43 Kitagawa M, Kitagawa K, Kotake Y, Niida H & Ohhata T (2013) Cell cycle regulation by long non-coding RNAs. *Cell Mol Life Sci* **70**, 4785–4794.
44 Chen LL & Carmichael GG (2010) Decoding the function of nuclear long non-coding RNAs. *Curr Opin Cell Biol* **22**, 357–364.
45 Pandey K, An HJ, Kim SK, Lee SA, Kim S, Lim SM, Kim GM, Sohn J & Moon YW (2019) Molecular mechanisms of resistance to CDK4/6 inhibitors in breast cancer: a review. *Int J Cancer* **145**, 1179–1188.
46 Korbie DJ & Mattick JS (2008) Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat Protoc* **3**, 1452–1456.
47 Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S & Madden TL (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics* **13**, 134.
48 Wick RR, Judd LM & Holt KE (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129.
49 Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100.
50 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R & Genome Project Data Processing, S (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
51 Thorvaldsdottir H, Robinson JT & Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178–192.

52 Adamopoulos PG, Theodoropoulou MC & Scorilas A (2018) Alternative Splicing Detection Tool-a novel PERL algorithm for sensitive detection of splicing events, based on next-generation sequencing data analysis. *Ann Transl Med* **6**, 244.
53 Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, de Castro E, Duvaud S, Flegel V, Fortier A, Gasteiger E *et al*. (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* **40**, W597–603.
54 Yang J & Zhang Y (2015) I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res* **43**, W174–W181.
55 Yang J, Yan R, Roy A, Xu D, Poisson J & Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7–8.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Raw nanopore sequencing reads corresponding to the novel transcripts *CDK4* v.2 - v.19.
Fig. S2. Raw nanopore sequencing reads corresponding to the novel transcripts *CDK4* v.20 - v.42.
Fig. S3. Indicative sequencing reads derived from semi-conductor sequencing technology on an Ion PGM™ platform, confirming the existence of 18 novel splice junctions between the curated exons of the human *CDK4* gene.
Fig. S4. Amino acid sequences of the putative protein isoforms encoded by the presented novel protein-coding *CDK4* mRNA transcripts.
Table S1. Primers used in RT-qPCR assays for the expression analysis of the identified *CDK4* transcript variants.
Table S2. Variant-specific primer pairs used for the RT-qPCR expression analysis of each identified *CDK4* mRNA transcript.