NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF CHEMICAL ENGINEERING

Prioritization of Pharmacological Compounds for Tumor Immunogenic Profile Manipulation using Next-Generation Sequencing Data

DIPLOMA THESIS

Katopodi Xanthi-Lida

SUPERVISORS

Assist. Prof. Vlachos Ioannis S.
BIDMC, HMS, Broad Institute of the MIT and Harvard

Prof. Boudouvis Andreas
National Technical University of Athens

Athens, October 2020

"Facts do not cease to exist
because they are ignored"

Aldous Huxley

Acknowledgements

The present study was carried out in collaboration with the Non Coding Research Lab at Beth Israel Deaconess Medical Center, HMS, Boston, MA, of Assistant Professor Ioannis S. Vlachos. The diploma thesis I am now called to deliver contains but a small fragment of my scientific journey as a member of the Non Coding Research Lab.

For this journey, I wholeheartedly thank my supervisor, Professor Ioannis Vlachos, for giving me the opportunity to expand my scientific horizons and work on science that strives to make a difference. I am truly grateful for our collaboration, his inspiring mentorship, his never-ending guidance and support, as well as for the pure bliss of our shared "Eureka!" moments.

Moreover, I would like to thank Professor Andreas Boudouvis, professor at the School of Chemical Engineering of N.T.U.A., for accepting to supervise this diploma thesis on behalf of the National Technical University of Athens. In actuality, I would like to genuinely thank him for his assistance, encouragement and trust throughout the years I spent as student at the School of Chemical Engineering.

Of course, this diploma thesis would have been entirely different if not for the extremely insightful time spent with all members of the Non Coding Research Lab, especially during our weekly lab meetings. I would like to particularly express my gratitude towards Dr. Yered Pita-Juárez, for his patience, kindness and willingness to assist me every time I reached out to him.

Furthermore, I am exceedingly grateful to my friends for their companionship and encouragement, as well as to my iGEM Athens family, for being a second scientific home to me; special thanks to Maria Litsa for her aesthetic interventions and grammar lessons.

Last but not least, I am forever indebted to my parents and family, without whom I would not be standing where I am.

So long, and thanks for all the fish,
X.L.K.

Abstract

Cancer is the second leading cause of death worldwide and one of the most well-researched medical topics. Nevertheless, the lack of effective treatment for the majority of tumor types is evident, while existing therapeutic approaches are unable to guarantee desired results. In the frame of this diploma thesis, a novel holistic approach is adopted for the discovery of drugs, compounds, and gene targets that can alter the immunogenic profile of tumors and expand the arsenal of immunotherapeutics.

The approach described in this study comes as a means to eradicate the current limitations of existing immunotherapies and improve their efficiency. In general, immunotherapy leverages components of the immune system in order to boost its ability to detect and destroy malignant cells. Unfortunately, only a small fraction of patients respond to immunotherapy, with the primary reason being the ability of tumor cells to bypass the immune system's control. Thus, it is essential for the improvement of existing immunotherapies that the tumor emanates its malignant nature in order for the immune system to detect and combat cancer cells. The proposed method is based on the existence of Tumor Specific Antigens (TSAs) that can elicit an immune response. The hypothesis made is that personalized gene targeting and drug administration are able to manipulate the antigenic profile of tumor cells as they will allow for the controlled generation of strongly immunogenic TSAs; such TSAs will act as targets for the immune system. Following this approach, tumors that were able to escape the control of the immune system can now be sensitized to immunotherapy.

For this purpose, a plethora of publicly available Next Generation Sequencing (NGS) studies were analyzed to reveal tumor-specific antigens whose production was induced following drug administration or gene targeting. An annotation-free and hypothesis-free approach to capture the expression and translation spaces was incorporated, allowing for unbiased characterization of the putative antigen spaces. After comparison of the two spaces to sets of data derived from healthy samples, results revealed significant cancer-specific effects on the transcriptome and translatome. Furthermore, the effect of treatment on the expression space was also evident on all three case studies included in the study, both in the number of treatment-specific transcripts and in the expression change between treatment and control. Last but not least, the implementation of a database structure was initiated which will enable efficient storing of analyzed results and cross-study comparisons.

Περιγραφή

Εδώ και τρεις δεκαετίες ο καρκίνος αποτελεί τη δεύτερη αιτία θανάτου παγκοσμίως, μετά τις καρδιαγγειακές παθήσεις. Το 2017 καταγράφονται 9.56 εκατομμύρια θάνατοι παγκοσμίως και 31 χιλιάδες στην Ελλάδα, ενώ το 2018 διαγιγνώσκονται 17 εκατομμύρια νέα περιστατικά καρκίνου ανά τον κόσμο. Αυτό που πιθανώς συγκλονίζει σε μεγαλύτερο βαθμό δεν είναι τα προαναφερθέντα νούμερα αυτά καθ' αυτά αλλά κάτι που κρύβεται από πίσω: η έλλειψη αποτελεσματικών θεραπειών για τα περισσότερα είδη καρκίνου ή, ακόμη, μιας καθολικής θεραπείας, αποτελεσματικής για όλους τους ασθενείς και τους τύπους.

Με τον όρο "καρκίνος" χαρακτηρίζεται μια ομάδα νοσημάτων με κοινό χαρακτηριστικό την υπερβολική, ανεξέλεγκτη και χωρίς προγραμματισμό ανάπτυξη και διαίρεση των κυττάρων του οργανισμού, με τις αιτίες εμφάνισης να εντοπίζονται σε κυτταρικό επίπεδο. Υπό φυσιολογικές συνθήκες, τα κύτταρα αναπτύσσονται, διαφοροποιούνται και εξειδικεύονται, και τέλος διαιρούνται προκειμένου να διατηρηθεί ο υγιής οργανισμός. Πιθανή εκτροπή από τη συγκεκριμένη φυσιολογική πορεία οδηγεί σε πλεονάζοντα κύτταρα και τη δημιουργία κυτταρικών όγκων. Ορισμένοι όγκοι, γνωστοί ως καλοήθεις, δεν είναι επικίνδυνοι για την υγεία και, επομένως, δεν χαρακτηρίζονται ως καρκινικοί. Αντίθετα, οι κακοήθεις όγκοι -καρκινικοί όγκοι, καρκινώματα ή νεοπλάσματα- θέτουν σε κίνδυνο την υγεία του ασθενούς ενώ έχουν την ιδιότητα να μεταπηδούν σε άλλους ιστούς του σώματος μέσω της διαδικασίας της μετάστασης. Εξαίρεση καρκίνου που δε σχηματίζει στέρεους όγκους (solid tumors) αποτελεί η λευχαιμία που προσβάλλει τα κύτταρα του αίματος.

Ο καρκίνος μπορεί να αναπτυχθεί σε όλους τους ιστούς του ανθρωπίνου σώματος και ως εκ τούτου προκύπτουν διαφορετικοί τύποι και μορφές καρκίνου. Πλέον, καταγράφονται περισσότεροι από 200 τύποι καρκίνου, ο καθένας εκ των οποίων αντιμετωπίζεται και θεραπεύεται με διαφορετικό τρόπο. Παρ' όλες τις διαφορές μεταξύ των τύπων καρκίνου, υπάρχει ένας μηχανισμός του οργανισμού που επιχειρεί να αντιμετωπίσει την εμφάνιση καρκινωμάτων: το ανοσοποιητικό σύστημα. Η αλληλεπίδραση του ανοσοποιητικού συστήματος με τον καρκίνο διακρίνεται σε τρεις φάσεις. Στην πρώτη φάση της εξάλειψης (elimination), τα καρκινικά κύτταρα αναγνωρίζονται από τα κύτταρα του ανοσοποιητικού και καταστρέφονται. Η αναγνώριση αποτελεί σημαντικό στάδιο και σημείο εκκίνησης για την φάση εξάλειψης, που διευκολύνεται από τη διαδικασία παραγωγής και παρουσίασης (presentation) των καρκινικών αντιγόνων (tumor antigens). Τα καρκινικά αντιγόνα αποτελούν ολιγοπεπτίδια που παράγονται στα καρκινικά κύτταρα, διαφέρουν από τα αντιγόνα των φυσιολογικών κυττάρων (αυτοαντιγόνα, self antigens), και όταν παρουσιάζονται στην επιφάνεια των κυττάρων προκαλούν, σε αντίθεση με τα αυτοαντιγόνα, ανοσολογική απόκριση. Με λίγα λόγια, τα καρκινικά αντιγόνα συνιστούν ένα καρκινικό "σήμα", προδίδοντας την καρκινική φύση του κυττάρου στο οποίο εντοπίζονται.

Η δεύτερη φάση αλληλεπίδρασης είναι αυτή της ισορροπίας (equilibrium), η οποία επέρχεται όταν η πλήρης εξάλειψη των καρκινικών κυττάρων δεν είναι πλέον εφικτή. Σε αυτή τη φάση, τα καρκινικά κύτταρα έχουν συσσωρεύσει γενετικές και επιγενετικές μεταλλάξεις, καθιστώντας τα πιο ανθεκτικά στην επίθεση του ανοσοποιητικού. Το ανοσοποιητικό σύστημα από την άλλη, περιορίζει την ανεξέλεγκτη διαίρεση και εξάπλωση των καρκινικών κυττάρων, αλλά δε δύναται να προκαλέσει την εξάλειψή τους. Το τρίτο και τελευταίο στάδιο είναι αυτό της διαφυγής (escape) κατά το οποίο το ανοσοποιητικό σύστημα χάνει πλήρως τον έλεγχο των καρκινικών κυττάρων. Τα καρκινικά κύτταρα, έχοντας συσσωρεύσει μεταλλάξεις που τα καθιστούν πιο ανθεκτικά, επιστρατεύουν κυτταρικούς και ανοσολογικούς μηχανισμούς προς όφελός τους, διαφεύγουν των αμυντικών μηχανισμών του οργανισμού και συνεχίζουν να αναπτύσσονται ανεξέλεγκτα.

Λόγω της σημασίας του ανοσοποιητικού συστήματος στην άμυνα του οργανισμού, και δη λαμβάνοντας υπ' όψιν τη δυναμική αλληλεπίδραση του με τον καρκίνο, η αξιοποίηση παραγόντων του ανοσοποιητικού για την καταπολέμηση καρκινωμάτων έχει αναδειχθεί ως η πιο επαναστατική μέθοδος θεραπείας του καρκίνου, γνωστή ως ανοσοθεραπεία. Προσεγγίσεις στην ανοσοθεραπεία περιλαμβάνουν τη χρήση αναστολέων σημείων ελέγχου, γενετικά τροποποιημένων Τ λεμφοκυττάρων, μονοκλωνικών αντισωμάτων και εμβολίων, με στόχο την ισχυρότερη ανοσολογική απόκριση του οργανισμού έναντι του καρκίνου. Οι ανοσοθεραπείες συνήθως επιφέρουν εξαιρετικά αποτελέσματα, με μακρόχρονη ίαση και λίγες παρενέργειες. Δυστυχώς, μόνο ένα μικρό ποσοστό των ασθενών ανταποκρίνεται στην ανοσοθεραπεία, γεγονός που αποδίδεται στην πολυπλοκότητα του ελέγχου και ρύθμισης των ανοσολογικών μηχανισμών αλλά και στην πολυπλοκότητα και ανομοιογένεια του ίδιου του καρκίνου και του καρκινικού περιβάλλοντος.

Στο επίκεντρο των μελετών για την υπέρβαση των περιορισμών της ανοσοθεραπείας βρίσκεται ο έγκυρος χαρακτηρισμός του αντιγονικού και ανοσολογικού προφίλ των καρκινικών κυττάρων. Σύγχρονες προσεγγίσεις περιλαμβάνουν την ανάλυση δεδομένων Αλληλούχισης Επόμενης Γενεάς (Next Generation Sequencing, NGS) για την εύρεση καρκινικών αντιγόνων και τη μελέτη αυτών ως προς την πιθανή ανοσολογική απόκριση του οργανισμού. Παράλληλα, ερευνώνται εις βάθος οι μηχανισμοί αναγνώρισης των καρκινικών κυττάρων από το ανοσοποιητικό σύστημα, οι κυτταρικοί και ανοσολογικοί μηχανισμοί που εμπλέκονται στην ανάπτυξη του καρκίνου και το καρκινικό περιβάλλον. Η πληρέστερη κατανόηση των παραμέτρων αυτών αξιοποιείται για την εξατομικευμένη προσέγγιση του καρκίνου μέσω του σχεδιασμού μιας θεραπείας συμβατής και αποτελεσματικής για τον εκάστοτε ασθενή, αλλά και για την ανακάλυψη νέων θεραπευτικών παραγόντων που δύνανται να βελτιώσουν την αποδοτικότητα της ανοσοθεραπείας για το σύνολο των ασθενών. Οι υφιστάμενες έρευνες, αν και επιχειρούν να απαντήσουν στο καίριο ερώτημα της βελτίωσης της αποδοτικότητας της ανοσοθεραπείας, κατά πλειοψηφία υστερούν στην ολιστική προσέγγιση του ζητήματος, συχνά αγνοώντας σημαντικές παραμέτρους.

Αντίθετα, η μέθοδος που αναλύεται στην παρούσα διπλωματική επιχειρεί να αποτελέσει πυλώνα για την εξάλειψη των υπαρχόντων περιορισμών στην ανοσοθεραπεία μέσω της αναζήτησης καθολικών σταθερών για όσο δυνατόν περισσότερους ασθενείς και τύπους καρκίνου. Έτσι, υιοθετείται μια αυτοματοποιημένη προσέγγιση του ζητήματος, μέσω της μελέτης φαρμάκων, ουσιών και γονιδίων-στόχων που δύνανται να μεταβάλλουν το ανοσολογικό προφίλ των καρκινικών κυττάρων, ενισχύοντας συνεπώς την αποδοτικότητα των ανοσοθεραπειών. Η μέθοδος που προτείνεται βασίζεται στην ύπαρξη καρκινικών αντιγόνων (tumor specific antigens - TSAs) που προκαλούν ανοσολογική απόκριση, και η βασική υπόθεση που πραγματοποιείται είναι η ακόλουθη: η χορήγηση φαρμακολογικών ουσιών ή/και η στόχευση γονιδίων σε ασθενείς με καρκίνο προκαλεί τη μεταβολή του αντιγονικού τους προφίλ επιτρέποντας την ελεγχόμενη παραγωγή ισχυρά ανοσογονικών αντιγόνων που θα αποτελέσουν στόχο για το ανοσοποιητικό σύστημα. Συνεπώς, με την παρούσα προσέγγιση δίδεται ένα επιπλέον εργαλείο για τη βελτίωση της αποδοτικότητας της ανοσοθεραπείας καθώς όγκοι που διέφευγαν του ελέγχου του ανοσοποιητικού συστήματος πλέον μετατρέπονται σε ισχυρά ανοσογονικοί και ευάλωτοι στην ανοσοθεραπεία.

Με τη μέθοδο αυτή επαναξιοποιείται πληθώρα ήδη υπαρχόντων πειραμάτων αλληλούχισης επόμενης γενεάς από *in vitro, in vivo* και κλινικές έρευνες. Σε αυτές περιλαμβάνονται περιπτώσεις υγιών και καρκινικών δειγμάτων, με ή χωρίς την στόχευση γονιδίων ή/και την χορήγηση φαρμακολογικών ουσιών. Στόχος είναι η ανάδειξη εκείνων των φαρμάκων και γονιδίων-στόχων που επιτρέπουν τον χειρισμό του αντιγονικού προφίλ των ασθενών και οδηγούν σε έντονα ανοσογονικούς όγκους, ευάλωτους στην ανοσοθεραπεία. Για την επίτευξη του σκοπού αυτού, το ζητούμενο προσεγγίζεται από δύο διευθύνσεις: την μεταγραφή-έκφραση και την μετάφραση.

Στην παρούσα μελέτη προτείνεται μια καθολική, αμερόληπτη προσέγγιση για την χαρτογράφηση τόσο του χώρου μεταγραφής όσο και του χώρου μετάφρασης, οδηγώντας σε χαρτογράφηση πιθανών αντιγονικών χώρων στα δύο αυτά επίπεδα. Αρχικά, η σύγκριση καρκινικών δειγμάτων με υγιή ανέδειξε την ύπαρξη μεταγραφικών και μεταφραστικών γεγονότων που απαντώνται μόνο στον καρκίνο. Παράλληλα, μέσω της ανάλυσης τριών διαφορετικών μελετών που περιλαμβάνουν χορήγηση φαρμακολογικής ουσίας ή/και στόχευση γονιδίων, υπογραμμίστηκε η επίδραση της εκάστοτε θεραπείας στην έκφραση, αποτελώντας ισχυρό επιχείρημα για την υπόθεση της χειραγώγησης του αντιγονικού προφίλ καρκινικών κυττάρων. Τέλος, παρουσιάζεται η δημιουργία μιας βάσης δεδομένων που θα επιτρέψει την αποθήκευση των επεξεργασμένων δεδομένων από τις μελέτες αλληλούχισης, καθώς και την σύγκριση μεταξύ αυτών προς ανάδειξη εκείνων των φαρμάκων που δύνανται να προκαλέσουν την ισχυρότερη μεταβολή στο ανοσολογικό προφίλ των καρκινικών κυττάρων.

Table of Contents

Section A
Theoretical Background

Chapter A.1. Introduction

The term *"cancer"*, as a medical term to refer to internal or external tumors, was first coined by the ancient Greek physician Hippocrates (460-370BC). Hippocrates studied the solid masses, and the finger-like projections of metastatic cancer cells most probably reminded him of crabs' legs and claws. The words he used to describe this medical condition were "καρκίνος" (carcinos) and "καρκίνωμα" (carcinoma), both of which refer to a crab in greek. The Roman encyclopaedist and physician Aulus Cornelius Celsus (25-50 BC) translated Hippocrates' term into "cancer", the roman for crab; hence the widespread medical term. Some centuries later, Claudius Galenus (Galen, 129-200 AD) adopted the greek word "όγκος" (oncos) for the tumors, the literal translation of which is "volume" but can in context refer to a certain -or increasing- quantity, and by extent to a developing mass. Galenus' term is still used in greek medicine for tumors, but also acted as the base for the name of cancer specialists: oncologists[1].

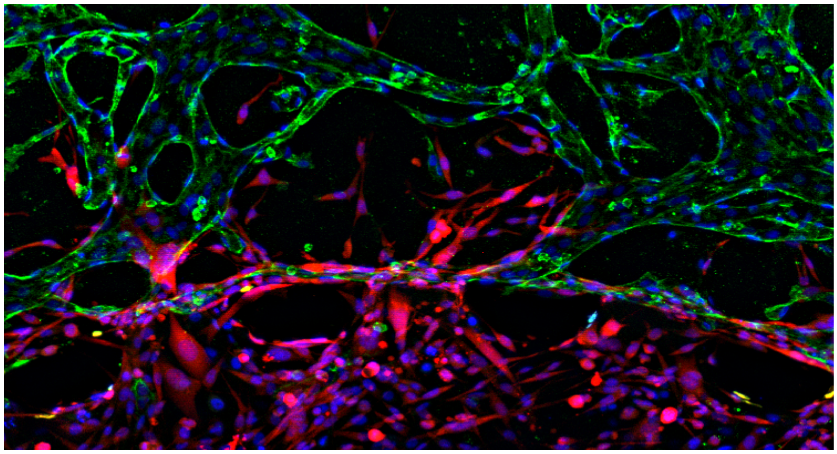

Figure A.1: Cancer cells under the microscope[2]

A.1.1. Cancer

The medical term *"cancer"* refers to a collection of related diseases that are all characterized by an excessive and uncontrollable division of abnormal cells. The initiation of this uncontrollable growth of cells may take place anywhere in the human body, leading to the formation of malignant tumors[1,3]. Cancer cells also have the ability to spread to different tissues during a process called metastasis, affecting not only neighboring tissue but even distant parts of the body[1,4].

Cancer is categorized in several groups with regards to the tissue that gives rise to it. Carcinoma originates from the skin or in tissues that line internal organs. Sarcoma originates from bone, cartilage, fat, muscle, blood vessels, or other connective or supportive tissue. Central nervous system cancers originate from brain and spinal cord tissues. Hematologic malignancies include leukemia, lymphoma and multiple myeloma. Leukemia originates from the blood and bone marrow, while lymphoma and myeloma originate from immune cells[3].

The origin of cancer lies primarily on a molecular level as a patient's DNA has or accumulates mutations that were not repaired. The vast majority of mutations that act as triggers for cancer occur on genes that encode functional proteins, which in turn control how cells function. Therefore, such mutations will disrupt the mechanisms that the cells employ to grow and divide[1]. Of course, any living human accumulates mutations on their DNA throughout their life, however the majority does not lead to cancer.

Genetic mutations appear during one's life as a result of environmental exposures, including exposure to carcinogenic substances or chemicals, or to radiation -UV rays from the sun or other manmade sources of radiation. Moreover, cancer-causing mutations might be the result of unhealthy habits, such as tobacco usage, alcohol consumption or even an unhealthy diet. There have also been several cases of infectious agents -viruses, bacteria and parasites- associated with cancer. Last but not least, it should be noted that genetic mutations that cause cancer can also be inherited from one's parents[1,5].

Each cancer patient bears their own unique genetic profile associated with their disease. Furthermore, clonal evolution of tumors suggests that a single malignant cell giving rise to a population of cancer cells causes tumor heterogeneity in pathology and molecular/genetic profiles. This concept is often linked to Darwinian selection at the micro-level, as intratumor cellular diversity gives rise to cancer cell populations with different abilities regarding expansion and proliferation, as well as resistance to therapy. Taking these into account, it's not far-fetched to assume that each patient's cancer is different[6].

A.1.2. Cancer in Numbers

Cancer is the second leading cause of death in the world, following cardiovascular diseases. 2017 marks 9.6 million deaths caused by cancer worldwide, of which 700 thousand were documented in the United States and 31 thousand in Greece[7].

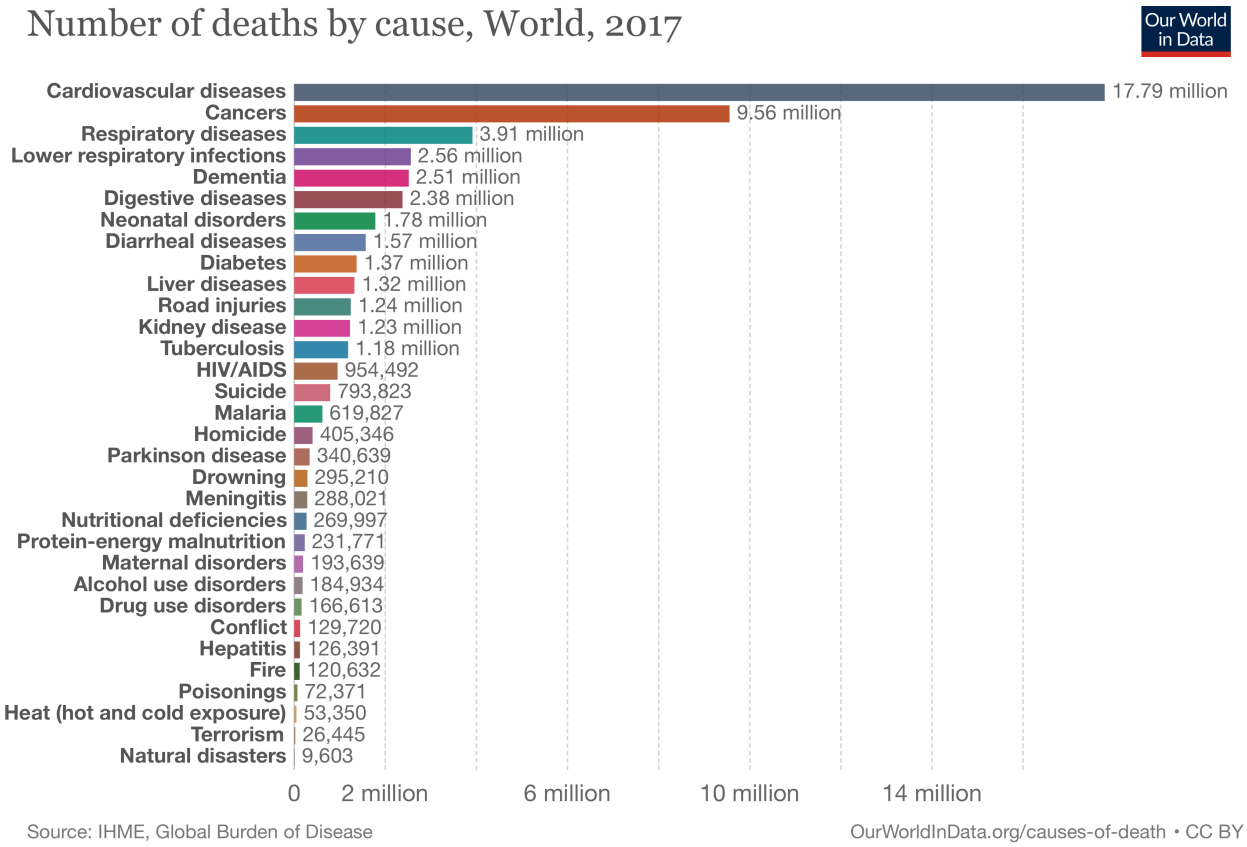

Figure A.2: Number of deaths by cause worldwide for 2017[6].

2018 marks 18 million new cases of cancer worldwide, with the United States bearing the first place in cancer occurrence since the 1990s with approximately 4 times more incidents per 100,000 population than the world average[7,8].

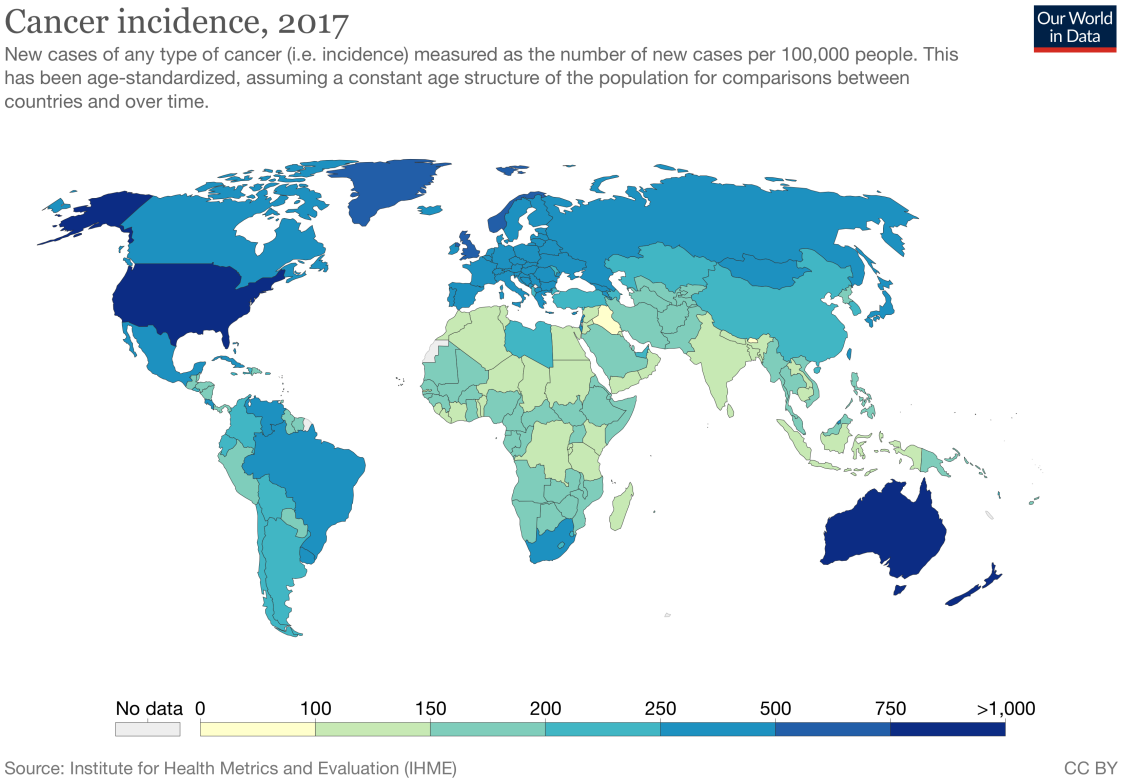

Figure A.3: World map of cancer incidence per 100,000 people for 2017[8].

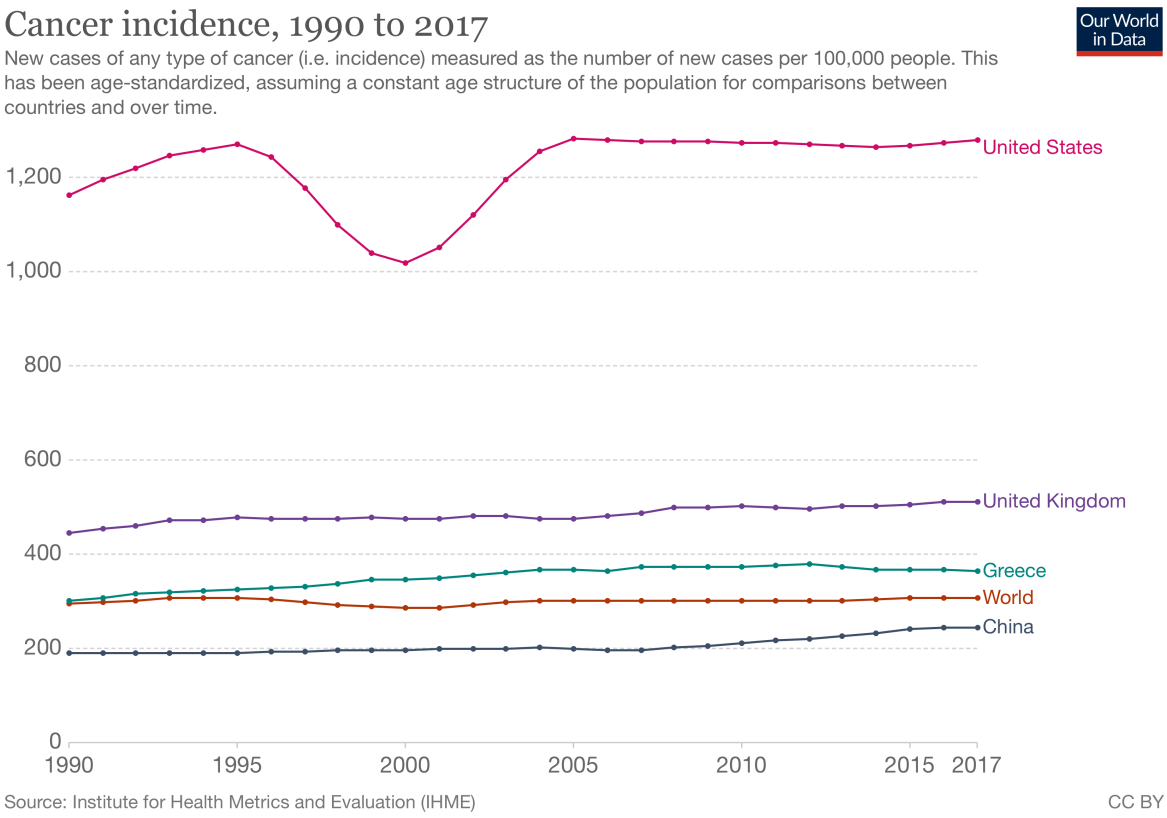

Figure A.4: Chart of cancer incidence per 100,000 people for 2017, including statistics for worldwide occurrences as for the United States, United Kingdom, China and Greece[8].

A.1.3. Next Generation Sequencing

In 2003, one of the most rigorous and insightful international projects was completed -The Human Genome Project. World-known scientists from all around the globe joined forces in order to unveil the truth of our own existence through own main task: deciphering the human DNA. This endeavor included the sequencing of the whole human genome -i.e. the uncovering of the order of the nucleotides that comprise the chromosomes of humans- as well as the identification of genes, their location on the genome, and their functions[9].

Nowadays, almost 20 years later, sequencing has become yet a standard procedure for multiple biological projects. The dawn of Next Generation Sequencing (NGS) ushered a new era of gaining insight on the genomic, transcriptomic -and many more "-omic"- profiles of organisms, cells, clinical samples. This second generation of sequencing methods allows for massively parallel deep sequencing of DNA[10,11]. Among the most prominent NGS experimental procedures are the following:

- **Whole Genome Sequencing (WGS)**: The technique by which the whole genome of an organism is sequenced. This includes all the chromosomes and mitochondria of an organism, and, in the case of plants, of the chloroplasts.

- **Whole Exome Sequencing (WES)**: The technique by which all protein coding regions of the genome are sequenced. Those regions are called exons, therefore this technique was called *exome* sequencing.

- **RNA Sequencing (RNA-Seq)**: The technique by which the RNA is sequenced. RNA is the transcribed version of the DNA, with multiple functions inside the cell. Messenger RNA (mRNA) acts as an intermediate between the DNA and protein synthesis, while other types of RNA may have structural functions, act as transporters, exhibit various regulatory functions, and many more. RNA-Seq is capable of capturing the majority of RNA types that are transcribed in a sample, or, with proper sample preparation, selectively sequence a certain type of RNA.

- **Ribosome Profiling (Ribo-seq)**: The technique by which actively translated mRNA molecules are sequenced. RNA is sequenced here as well, however the sample preparation and RNA targeting procedure varies greatly compared to RNA-seq; thus, Ribo-seq holds a different place in the panel of NGS techniques.

Chapter A.2. Cancer Immunology

The immune system combats numerous threats: exogenous, such as bacteria, viruses and other pathogens, as well as endogenous, such as cancer. It is the major defense mechanism of the human body, with extraordinary cells and mechanisms at its disposal, and is comprised of two arms: the innate and adaptive immunity. One could envision the innate immunity as a first line of defense against external or internal enemies, whereas the adaptive immunity is a more sophisticated and well-trained army of cells. But how does the immune system combat caner?

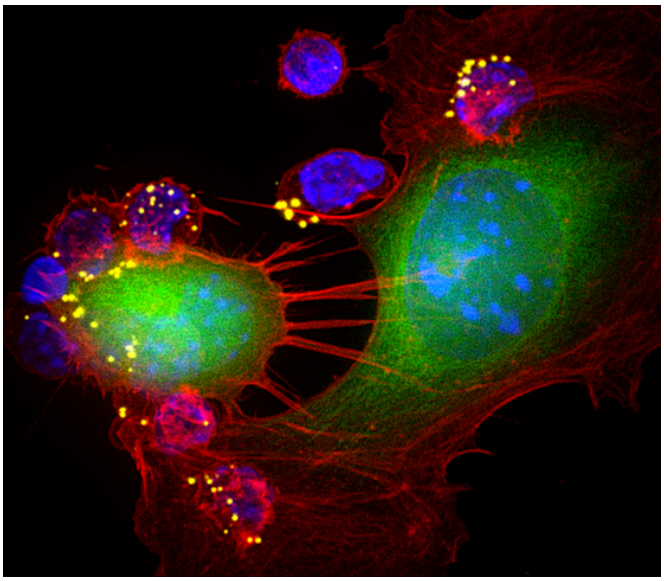

Figure A.5: Multiple T-cells (blue) attacking B16F10 tumor cells (green)[12].

A.2.1. The Immune System versus Cancer

The modus operandi of the immune system is based on a dynamic interaction between its two arms, the innate and the adaptive -or acquired- immunity. Innate immune responses are considered ancient defense mechanisms, relying on the recognition of typical, pathogen-associated motifs via a set of receptors. These receptors, termed pattern recognition receptors (PRRs), are non-specific, and bind to the pathogen-associated molecular patterns (PAMPs) or the damage-associated molecular patters (DAMPs), eliciting a swift immune response[13,14].

On the contrary, acquired immunity revolves around the fundamental ability to specifically detect an extraordinarily diverse set of molecules (i.e. antigens) and rapidly induce an immune response. This ability, termed immunological memory, is based upon prior encounter with the particular threat, to which end the innate immunity plays a crucial role[14]. In brief, innate immune cells encounter exogenous or endogenous threats, gather information on their macromolecular profile, and then pass the acquired knowledge to the adaptive immune cells; those cells in turn are trained on the new agenda, begin patrolling for potential breaches in defense, and upon contact with the known enemy, they are activated, attacking and eradicating the threat.

The explanation above is but a simplification of the sophisticated approach with which the immune system coordinates its defense against cancer and all other potential threats. In order for this defense machinery to be successful, it requires a great amount of fine-tuning between a variety of different cells that belong to both branches of the immune system. Certain types of cells that partake in this procedure are of particular notice, especially in cancer immunology, and will be concisely discussed.

Prior to that, another important term should be introduced: the term "antigen", which more often than not is associated with the term "antibody" as its counterpart, referring to a molecule that induces the production of antibodies. However, this tautologic definition of antigens is not entirely accurate. At a molecular level, an antigen is a molecule that can be recognized by -i.e. bound specifically to- the antigen-binding domain of an "antigen receptor" -i.e. an antibody or T-cell receptor (TCR)[14]. Again, this second definition is tautological as well; yet, it implies a one-on-one relationship between an antigen and a receptor, while not being limited to exogenous or endogenous threats, or even antibodies. In fact, the second definition also introduces self antigens normally produced in healthy cells, which serve as a sanity check for the immune system.

With regards to endogenous and exogenous threats, antigens in essence refer to products that are not normally produced by a healthy organism, hence are indicators of external intruders or internal abnormalities, serving the self vs non-self distinction. As such, tumor antigens are a significant signal produced by cancer cells, emanating their malignant nature. At the same time, they act as means to train the adaptive immune cells on identifying tumor cells. The latter is facilitated with the help of antigen presenting cells (APCs). APCs are mostly innate immune cells, which differ significantly from each other

both in antigen uptake -the acquisition of antigens- and effector functions -inducing certain abilities to the cells they are presenting the antigen to[14].

In order to induce an immune response against cancer, there are two phases involved: the priming and effector phases. An interesting example of antigen presenting cells that participate in this process are dendritic cells (DCs). During the priming phase, immature DCs obtain the tumor antigens of dying cancer cells, exhibiting phagocytic behavior towards apoptotic cells[13,14]. Cell death induces danger signals which function as stimulants to increase immunogenicity and, alongside antigens, lead to the maturation of dendritic cells. Upon maturation, DCs travel to the lymph nodes where they train the lymphocytes, T cells and B cells, the adaptive immune cells responsible for carrying out adaptive immunity. Trained lymphocytes carry antigen receptors specific for a given antigen, namely antigen-specific receptors, and proliferate only after exposure to said antigens[13].

Subsequently, the effector phase is initiated, with the continuous interaction between innate and adaptive immunity. During this phase, cytotoxic T cells (cytotoxic T lymphocytes, CTLs) are the main effector cells, charged with the responsibility to recognize, attack and eradicate tumor cells. Moreover, T helper cells are also developed during the priming phase, and they are responsible for mentainance and effectiveness of cytotoxic T cells[13,14,15].

A.2.2. Cancer Immunosurveillance and Immunoediting

In 1957, Thomas[16] and Burnet[17] introduced the **Immunosurveillance Hypothesis**, a notion that portrays the active role of the immune system in monitoring the development of tumors by recognizing and eliminating malignant cells[18]. This theory has since been reviewed and extended to a new notion termed **Cancer Immunoediting**, as strong experimental evidence suggests that the immune system does not only combat tumor cells, while at some cases is manipulated to enable their proliferation. Although this might seem counter-intuitive -as the sole purpose of the immune system is to attack potential threats-, the progressive evolution of tumors enables them to avoid components with anti-tumor properties and hijack a plethora of immune pathways and mechanisms to mask their malignant nature[18,19]. The dynamic interaction between the immune system and cancer, as portrayed in Cancer Immunoediting theory, is divided in three phases: elimination, equilibrium, and escape[13,19].

During the elimination phase -the Immunosurveillance Hypothesis analogue-, the malignant cells are recognized and destroyed by components of the immune system. Recognition of tumor cells is a most important step as well as the initiation point during the elimination phase. It involves the appearance of threat signals that are either secreted by tumor cells -e.g. interferons (IFNs), signaling molecules that are produced by cells in response to abnormal behavior and activate immune cells- or presented on the surface of

tumor cells[13]. The latter category of molecules includes antigens which are the leading signals in distinguishing between normal cells -"self"- and tumor cells -"non-self". Cell antigens, be it self or tumor, are peptides typically 9 to 11 amino acids long that are displayed on the cell surface; in cancer they are termed tumor antigens or neoantigens. The production, presentation, and recognition of tumor antigens will be discussed later in more detail. However, it should be noted that neoantigens are significantly different from self antigens, and that fact from an immunological perspective ensures that the components of the immune system will be able to distinguish them from self antigens and recognize them as threat signals[20]. Following tumor cell recognition, the immune cells are activated to respond and eradicate the malignant transformed cells.

If the immune system fails to destroy tumor cells during the elimination phase, the equilibrium phase is established. During this state, the tumor further evolves and mutates into more resilient forms which the immune system can restrict but cannot destroy. In a sense, this phase serves as a functional dormancy state between the immune system and tumor, all the while both sides seek routes for asserting dominance over one another. The equilibrium phase has been described as a long-lasting period throughout which no clinical manifestations of cancer are reported[13].

The third phase of escape is established as soon as the immune system is incapable of sustaining the dynamic equilibrium with the tumor, and/or cancer cells manage to evade the immune system's control by developing or utilizing cellular mechanisms to their advantage. Up to this point, cancer cells have accumulated numerous genetic and epigenetic alterations and have hijacked a plethora of mechanisms to escape immune control, leading to clinical manifestations of cancer.

A.2.3. Antigen Processing and Presentation Machinery

Antigen presentation on the cell surface involves several steps, all of which take place on a molecular level, with the process being common for both normal and tumor cells. All antigens, self and non-self, are the processed products of proteins expressed in the cells. Antigen processing and presentation involves the following steps[21]:

a. Proteins intended for degradation -mainly because they are deemed unneeded or damaged- are tagged during a biological process called ubiquitination, during which the small molecule ubiquitin is attached to the protein. Ubiquitin acts as a signal for the protein-transport machinery of the cell to move said protein for degradation[22].

b. The tagged proteins are moved to and degraded into smaller peptides by the proteasome -a protein complex found in the cytosol of the cells.

c. The peptides bind to HSP90 on the cytosol -a chaperone protein that acts as a protein stabilizer and aids protein degradation[21,23].

d. The peptides are transported to the endoplasmic reticulum (ER) by the TAP1-TAP2 protein complex, also known as the transporter associated with antigen processing (TAP)[21,24].

e. In the endoplasmic reticulum, the peptides are trimmed to appropriate length, 9 to 11 amino acids long, by ER aminopeptidases associated with antigen processing (ERAP).

f. The trimmed peptides, still in the endoplasmic reticulum, bind to newly synthesized major histocompatibility complex (MHC) class I molecules with the help of chaperone proteins[13,21,24].

g. The MHC-I-antigen complex is transported to the cell surface and the peptide is thus presented.

As briefly mentioned above, the major histocompatibility complex (MHC) class I molecule serves as the platform where the antigens bind to, and thus can be presented on the cell surface after the complex is transported there. MHC-I molecules are therefore regarded as extremely crucial components of the antigen presentation machinery.

Once the antigen is bound and presented by MHC-I, cytotoxic T lymphocytes (cytotoxic T cells or TCLs) bind to the antigen and recognize its self or non-self nature. In the case of non-self antigens -or of self antigens in autoimmune diseases- the T cells are activated and attack the cell that presented the non-self antigen. Specifically for cancer, the tumor-restricted expression of tumor antigens -a.k.a. neoantigens- guarantees the targeted activity of T cells against the tumor cells and subsequent absence of activity against normal cells[20].

A.2.4. Sources of Tumor Specific Antigens

In healthy cells, self antigens are the processed products of proteins normally expressed by those cells. As such, self antigens do not elicit an immune response as the immune cells have been trained to tolerate them and not be activated by them. However, in cancer cells tumor-specific antigens (TSAs) fall under at least one of the following general categories which describe their non-self nature:

a. They are the processed products of proteins or peptides not normally expressed in healthy cells.

b. They are the processed products of proteins or peptides normally expressed in tissues different from the ones in question.

c. They are the processed products of proteins or peptides normally expressed in healthy cells but are over-expressed in the tumor cells in question.

A more detailed depiction of the putative sources of tumor specific antigens is given in the figure below:

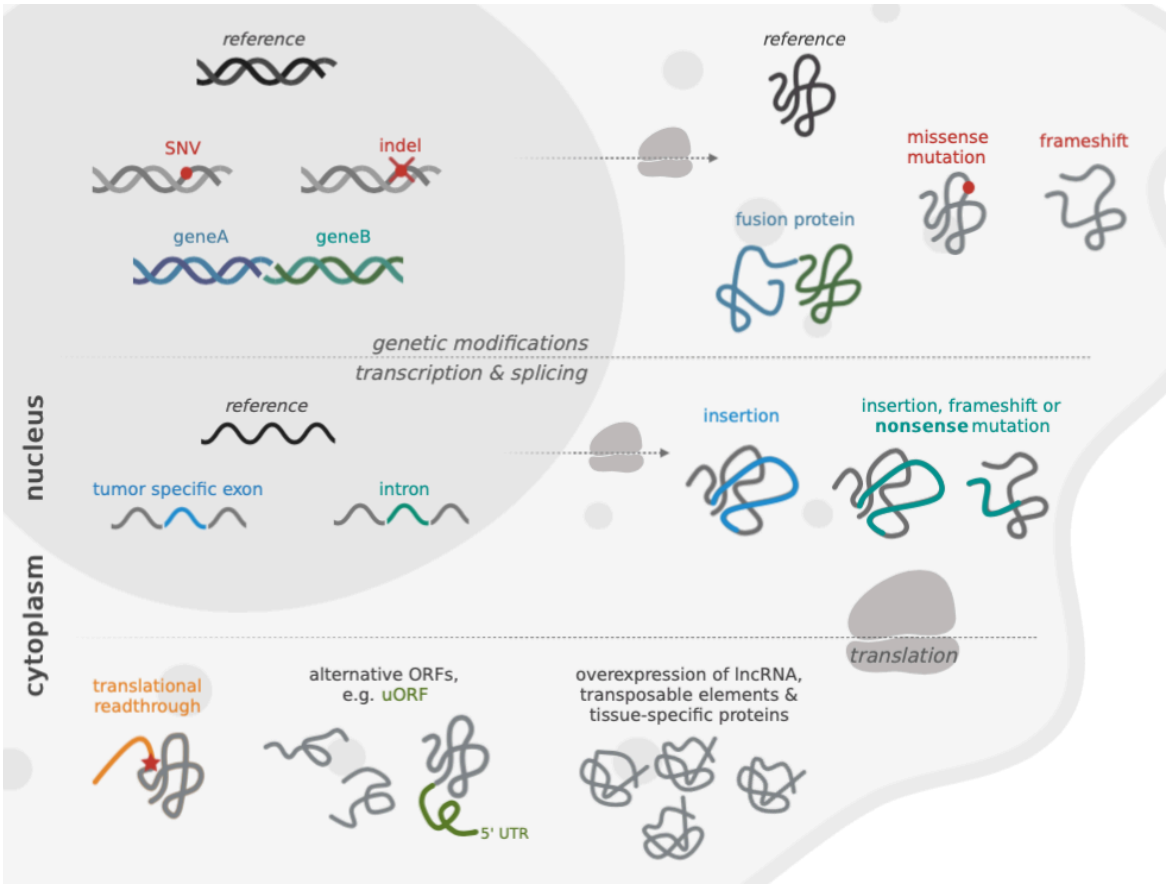

Figure A.6: Schematic of a cancer cell and the possible sources of neoantigens. These include non-synonymous mutations, protein fusions, cancer-specific exon inclusion, intron retention, translational tread through, expression of retroviral elements, expression of Cancer Testis Antigens (CTAs), alternative ORFs -to name the most well-documented ones.

A.2.5. Immune Escape Mechanisms

Immune escape is of particular interest from both clinical and therapeutic perspectives. Known escape mechanisms include the following[13]:

a. Defective tumor antigen processing and presentation machinery.
b. Lack of immune activating pathways.
c. Presence of mechanisms that inhibit immune response and induce an immunosuppressive state.
d. Development of resistant tumor cells as a result of accumulated genetic mutations.

The aforementioned mechanisms may shape immune response by affecting three distinct areas of immune regulation: the tumor's antigenicity (a), immunogenicity (b and c), and microenvironment (c). The accumulation of genetic mutations that leads to resistant tumor cells (d) is an umbrella case that spans over all three aspects, and sometimes could even be considered the end result of successful immune escape action of tumor cells.

Antigenicity

The ability of the immune system to distinguish between healthy and tumor cells is of paramount importance for an effective response against malignant cells. As discussed in previous sections, immune cells are able to do so via the recognition of non-self tumor antigens found on the surface of cancer cells. Therefore, retention of antigenicity is crucial for the immune system. On the contrary, loss of antigenicity serves as a way for cancer to escape immune surveillance and proliferate.

Loss of antigenicity may occur due to immune selection of cancer cells -i.e. the process by which cancer cells evolve and adapt to escape immune surveillance. Thus, cytotoxic T cells are not able to recognize cancer cells and attack them. However, antigenicity might also be compromised due to the accumulation of defects in the antigen presentation machinery that lead to partial or total loss of antigen presentation. Therefore, even if a tumor expresses sufficient antigens that may elicit an immune response, the immune system's ability to detect those cells and eradicate them also depends on the functionality of the MHC-antigen complex. In fact, defects of the antigen presentation machinery have been associated with a number of common solid tumors, including melanoma, breast, lung, renal, prostate and bladder cancers[18,25].

Immunogenicity

Even if tumors retain their antigenicity, enabling recognition by immune cells, they can still escape immune surveillance by decreasing their immunogenicity. This can be achieved via pathways that induce the downregulation of immune responses. Proteins CTLA4 (cytotoxic T-lymphocyte-associated protein 4) and PD-1 (programmed cell death protein 1) and the latter's ligand, PD-L1 (programmed death-ligand 1) have been linked with pathways that suppress the immune system. These molecules exhibit inhibitory abilities, actively obstructing anti-tumor T cell responses[18,26].

Tumor Microenvironment

The tumor microenvironment (TME) refers to normal cells, blood vessels, and molecules that surround and sustain malignant cells. Research evidence suggests that tumor proliferation or eradication relies on a dynamic interaction between the tumor cells and the tumor microenvironment, while TME is also said to shape therapeutic responses and resistance[27,28].

Tumors retaining antigenicity and immunogenicity may still be able to escape immune surveillance by altering the tumor microenvironment into one that suppresses infiltration of leukocytes. Tumor infiltrating lymphocytes (TILs) are essentially T cells that pervade tumor tissue, recognize malignant cells and proceed with elimination, thus being an important pillar in immune-mediated eradication of cancer. The immunosuppressive tumor microenvironment that some tumors establish may lead to reduced invasion of TILs, inactivation of TILs, or even manipulation of TILs to the tumor's advantage[18,29].

Chapter A.3. Cancer Immunotherapy

In 2018, the Nobel Prize in Physiology or Medicine was awarded jointly to James P. Allison and Tasuku Honjo "for their discovery of cancer therapy by inhibition of negative immune regulation," marking the beginning of a new era in cancer immunotherapy[30]. The two researchers had been studying CTLA-4 and PD-1, respectively, since the 1990s, revealing their corresponding roles on suppressing the activity of cytotoxic T cells. Reverse engineering this knowledge, they both worked on the concept of removing the breaks that keep the T cells inactive by inhibiting the action of CTLA-4 and PD-1. This breakthrough expanded the arsenal of immunotherapeutics and allowed for new approaches in cancer treatment, with authorities having since approved the use of immune checkpoint blockade for numerous different types of cancer with outstanding results[31].

Figure A.7: James P. Allison (left) and Tasuku Honjo (right), the 2018 Nobel laureates in Physiology or Medicine[30].

A.3.1. Types of Immunotherapies

Cancer immunotherapies leverage components of the immune system in order to boost its ability to recognize, attack, and destroy cancer cells. Types of immunotherapies include the following[32,33]:

- **Monoclonal Antibodies (MABs)**: MABs are immune system proteins produced in the lab. One category of MABs attaches itself to the surface of cancer cells, facilitating the recognition of said cancer cells by the immune system. Another category acts as switches that regulate pathways and machinery used by cancer cells to proliferate. A third category of MABs may attach to both cancer and T cells, assisting cancer cells eradication by T cells[34,35].

- **T cell Transfer Therapy**: The goal of this therapy is to provide the patient's immune system with more robust T cells that will be able to destroy malignant cells. There are two types of T cell transfer therapies: tumor infiltrating lymphocytes (TILs) therapy and CAR-T cells therapy. During TIL therapy, which is based on the ability of TILs to recognize tumor cells, TILs are extracted from the TME, cultivated to increase their numbers, and then re-inserted intravenously to the patient's blood stream. On the other hand, CAR T cell therapy includes the modification of T cells in order to be able to produce a chimeric antigen receptor (CAR) which exhibits higher specificity in recognizing a particular antigen[26,36,37].

- **Immune System Modulators**: This type of molecules are based on proteins normally produced by the immune system, including cytokines. As an immunotherapy method, immune system modulators are either natural or artificial, and are used to stimulate a more robust immune response[38,39].

- **Immune Checkpoint Inhibitors (ICI)**: The work of Allison and Honjo expanded the potential of cancer immunotherapies by introducing a revolutionary approach. Immune checkpoints, such as CTLA-4 and PD-1, are proteins normally expressed in immune cells, and their role is to suppress strong immune responses so that cytotoxic T cells do not destroy healthy cells. By releasing those immune breaks, T cells are authorized for an all-out cancer cell eradication. This can be achieved with molecules that inhibit the action of the immune checkpoints; therefore this type of therapy was named immune checkpoint inhibition[31,40].

- **Cancer Vaccines**: As all types of vaccines, cancer vaccines utilize molecules or cells that train the immune system to recognize and attack a threat -in an upcoming encounter- or stimulate it into action -if the disease is present. Cancer vaccines include preventive and treatment vaccines, that either contain tumor antigens, whole cancer cells, or immune cells vaccine. Cancer vaccines are still available as part of clinical trials[20,41].

A.3.2. Advantages

The crux in cancer immunotherapy is how normal defense mechanisms of a patient can be used, or further accentuated to eliminate cancer cells. Hence, the focus of cancer treatment has been shifted from the tumor to the host's immune system. This aspect in itself serves as a revolutionary path in cancer therapies, owning to its universal character.

Immunotherapies have demonstrated their efficacy for a plethora of cancer types, and particularly so against cancers that have been resistant to chemotherapy or radiation therapy, e.g. melanoma, with outstanding results and long-term survival rates. The latter is based on the ability of the immune system to form an immunologic memory, which is further amplified through immunotherapy. This has been apparent in patients with metastatic cancers as well; metastatic cancers are considered an incurable disease for the majority of patients, with immunotherapy managing, in several cases, to confront the complicated nature of metastasis[26,42].

Furthermore, the side effects of immunotherapy are usually mild, especially compared to chemotherapies or radiation therapies which most certainly will expose the patient's healthy cells to additional perils[42].

A.3.3. Limitations

Pitfalls in cancer immunotherapy with regards to side effects may include overstimulation or misdirection of the action of the immune system, with symptoms varying from fever and inflammation to more severe conditions that resemble autoimmune diseases[42]. At the same time, currently available immune therapies are costly, and the therapeutic agents are often associated with significant toxicity[43].

However, the major challenge in cancer immunotherapy emerges with respect to its efficacy and consistency across the majority of cancer patients and cancer types. Even if immunotherapies have demonstrated outstanding results, they have done so only for a small fraction of the patients. Moreover, in several cases, patients have exhibited acquired resistance to immunotherapy, with responders relapsing after a period of response. Such facts may not come as a surprise, taking into account the highly complex nature of cancer -across patients with different genomic profiles and also given the intratumor diversity-, the exceedingly regulated nature of the immune system, as well as the dynamic interaction between the two, with the latter sometimes leading to immune-protected tumors[26,44].

The inability to predict whether a certain type of immune therapy will be of benefit for a specific patient is evident, and alarming. The fact that the medical community is not able to distinguish the appropriate approach for the treatment of a patient, poses an obstacle of paramount importance. This lack unveils the need to determine the mechanisms of

tumor immune escape and/or how the immune system selects for proliferating tumors. Furthermore, it calls for the discovery of additional biomarkers that will point to the suitable treatment. For example, thus far, high expression of PD-L1 has been a strong indicator of response to anti-PD-1 therapy; however, this has not been the case for all tumors, and PD-1 is but a single molecule that may act as a biomarker[18,44,45].

At the same time, genomic instability due to accumulation of non-synonymous mutations, may lead to production of immunogenic tumor antigens, while several studies have linked the high Tumor Mutation Burden (TMB) with response to immunotherapy. However, a majority of those antigens do not seem to lead to the induction of T cell responses. This observation leads to two main conclusions: First, the tumor mutation burden cannot always be an indicator of response to therapy. Secondly, naturally occurring tumor antigens may not be sufficient for proper activation of an immune response, or the continuous exchange between the tumor and the immune system may have lead to selection of the tumor antigen repertoire that will not elicit such a response, or tumor may actively block antigen presentation. Again, the need for understanding the underlying mechanisms of immune escape is brought forth, alongside the urgency for biomarkers of diagnostic, predictive, prognostic and/or therapeutic value[20,45].

A.3.4. Anticipated Innovations

All of the limitations brought up in the previous paragraphs point to one of the most anticipated innovations in cancer treatment: the personalization of immunotherapy that will allow for efficient, targeted treatments with long-lasting results. Since each patient's tumor is unique, the identification of all prominent factors that enable tumor progression and shape the immune response will allow for the selection of the most suitable therapy, or even a combination of therapies.

In order to facilitate the personalized aspect of cancer treatment, a number of other innovations are expected to occur: First of all, the unveiling of the immune escape mechanism and pathways that may bring forth therapeutic targets or prognostic biomarkers. Furthermore, a more accurate charting of the antigenic and immunogenic profile of tumor cells, allowing for the utilization of tumor antigens as therapeutic targets. Last but not least, the discovery of therapeutic approaches that might synergistically enhance each other; for example, there have been indications of T cell checkpoint inhibition therapies being accentuated by tumor antigen-reactive T cells, and vice versa[20,45].

All of the above innovations are expected to lead to better treatment outcomes, higher long-term survival rates and lower relapse rates, reduced treatment cost and toxicity, and, overall, to efficient and consistent methods to combat cancer once and for all.

A.3.5. Manipulation of the Immunogenic Profile

In order to unveil the underlying "structure" of cancer immunity, several research groups have deliberately tinkered with pathways, mechanisms, and molecules involved with cancer. The end goal is to reveal potential therapeutic targets or prognostic markers from the observed outcomes. Other research groups, at the same time, have discovered various cancer-specific phenomena that shed light on the production of tumor antigens.

Falling under the latter category, it has been shown that aberrant splicing related to accumulation of mutations in genes associated with the spliceosome has led to intron retention events in cancer. This may lead to the production of immunogenic antigens, since intron retention events do not often occur in healthy cells[46]. Furthermore, it has been shown that endogenous retroelements may be a primary source of tumor antigens in cancer cells. Dysregulation of pathways that suppress their expression in healthy tissue may fall apart in the malignant environment of cancer cells, leading to the production of endogenous viral proteins, and subsequently of immunogenic antigens[47]. More related results can be found in paragraph A.2.4.

In addition, multiple pathways and their components have been studied over the years due to their implication in cancer progression. For example, components of RNA methylation and demethylation pathways, such as the enzyme FTO, have been linked to tumorigenesis[48]. Another important enzyme studied by many groups is PRMT5, associated with methylation and splicing. Combination of immune checkpoint inhibition with pharmacological inhibition of PRMT5 showed limited tumor growth in mouse models, suggesting that inhibition of PRMT5 in a clinical context may enhance an anti-tumor immune response[49]. A handful of studies have also demonstrated how targeting specific pathways, mechanisms, and molecules may lead to a change in the immunogenic profiles of cells. Among the mechanisms studied, is the DNA mismatch repair (MMR) system, responsible for the repair of erroneous incorporation of bases that might arise during DNA replication. MMR-deficient and MMR-defective tumors, as well as purposeful inactivation of MMR with the clinical agent temozolomide, showed increased mutational burden, which in turn lead to production of tumor antigens[50].

All of the above are, milder or stronger -as in the case of MMR inactivation-, indicators that alteration of the immunogenic profile is possible via interference with drugs or compounds, or by targeting specific genes and pathways that have been related to cancer. This interference seems to be a pillar in inducing production of antigens that may act as targets for the cytotoxic T cells.

Purpose of this Study

The essential reason behind patients not being able to respond to cancer immunotherapy primarily lies with the exceedingly diverse nature of cancer, both across patients and across cells that comprise the same tumor, as well as its ability to escape immune surveillance, and proliferate. The immune escape mechanisms, already known to a certain degree, need to be circumvented or confronted in order for the immunotherapy to show its true potential in eradicating tumors.

State of the art approaches focus on two main aspects: First, targeting of mechanisms, cellular procedures, and molecules that are known to be implicated in cancer, in order to uncover their therapeutic potential; Secondly, characterization of the tumor immunogenic landscape, primarily by focusing on specific potential sources of tumor antigens. Both approaches are partially restricting, and rely on hypotheses that have been suggested by previous studies, all the while possibly disregarding the importance of the different immune escape mechanisms that might be at play.

This study introduces a holistic approach for the accurate characterization of the antigenic profile of cancer cells, which revolves around two axes: expression and translation. First of all, the proposed method goes beyond putative genomic sources of antigens, extracting products of any genomic, transcriptomic, translatomic, or other origin, procedure, or anomaly. Furthermore, this study focuses on how the antigenic profile can be manipulated, searching for antigens that are produced solely in cancer cells and exclusively after drug administration or gene expression perturbation. The analysis of compounds or genes does not rely on prior knowledge of a relationship between said compounds or genes with cancer. On the contrary, all publicly available datasets with cases of drug administration or gene targeting in tumor cells are potential sources of information for the suggested method. Analysis of such a vast set of datasets allows for the generation of universal, hypothesis-free metrics that can highlight prominent antigens, and prioritize compounds or genes that can be used to alter the antigenic profile of tumors.

Section B
Methods

The methods implemented below were developed, tested, and discussed in the Non-coding Research Lab of Assistant Professor Ioannis S. Vlachos at Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA. The analyses were performed on a High Performance Computing environment, on Harvard's O2 Server.

B.1. Capturing the Kmerome

Analysis of RNA-Seq studies reveals information on the transcriptome, the actively transcribed RNA molecules for each sample. In this study, the goal is to go beyond the expression space and search for abundance of kmers of specific lengths. Antigens are usually 9 to 11 amino acids long, which correspond to RNA sequences of lengths 27, 30 and 33, respectively. Therefore, the kmers of interest that may lead to antigens are 27, 30, and 33 nucleotides long; such kmer spaces are generated and analyzed below. B.1. Panel of Normals

B.1.1. The Panel of Normals

The Genotype-Tissue Expression (GTEx) project is an ongoing endeavor to study tissue-specific gene expression and regulation. Healthy samples from all human tissues are included in the database, containing WGS, WES, and RNA-Seq experiments[51]. In order to highlight the tumor-specific nature of certain transcripts that may prove to have an antigenic potential, a control baseline needs to be created. Normally expressed genes and transcripts from GTEx RNA-Seq samples are the means to create a Panel of Normals, which provides the transcriptomic basis of non-tumor transcripts and acts as a first layer of filter to yield tumor-specific antigens.

The following tissues were utilized in creating the PoNs from GTEx samples:

Tissue	Sex	Number of Samples	Total Number of Samples
Adipose Tissue	Male	5	10
	Female	5	
Adrenal Gland	Male	5	10
	Female	5	
Blood Vessel	Male	5	10
	Female	5	
Brain	Male	5	10
	Female	5	
Breast	Male	5	10
	Female	5	
Colon	Male	5	10
	Female	5	

Esophagus	Male	5		10
	Female	5		
Heart	Male	5		10
	Female	5		
Lung	Male	5		10
	Female	5		
Muscle	Male	5		10
	Female	5		
Nerve	Male	5		10
	Female	5		
Ovary	Female	5		5
Pancreas	Male	5		10
	Female	5		
Pituitary	Male	5		10
	Female	5		
Prostate	Male	5		5
Salivary Gland	Male	5		10
	Female	5		
Skin	Male	5		10
	Female	5		
Small Intestine	Male	5		10
	Female	5		
Spleen	Male	5		10
	Female	5		
Stomach	Male	5		10
	Female	5		
Thyroid	Male	5		10
	Female	5		
Uterus	Female	5		5
Vagina	Female	5		5
			TOTAL	***210***

Table B.1: Composition of the Panel of Normals with regards to the tissue origin of the GTEx RNA-Seq samples used.

A total of 210 GTEx RNA-Seq samples were analyzed, taken from 23 tissue types and both sexes, apart from sex-specific tissues -ovary, prostate, uterus, vagina. **Testis samples were excluded** from the creation of the PoNs, as testis transcripts have been identified as immunogenic antigens in tumors originating from tissues different from testis -Cancer Testis Antigens, CTAs.

The samples above were analyzed as follows:

- Downloaded from dbGaP, the database of Genotypes and Phenotypes.
- Quality control was performed following best practices[52]:
    - Initial quality control of reads using FastQC[54] and MultiQC[55].
    - Trimming of adapter sequences using Cutadapt[56].
    - Secondary quality control of trimmed reads using FastQC[54] and MultiQC[55].
- Unstranded studies were converted to stranded to resolve read orientation. **Homebrew pipeline***.
- Reverse reads were reverse complemented to resolve sense orientation. **Homebrew pipeline***.
- Reads were kmer-ized using Jellyfish[53] to nucleotide sequences of lengths lengths 27, 30, and 33 bases, to capture the 9-11a.a. length of the antigens.
- Panels of Normals for different kmer lengths (k = 27, 30, and 33) were created by concatenating results from all samples to a single file containing unique instances of all kmers found on healthy GTEx samples.

B.1.2. Analysis of RNA-Seq Studies

RNA-Seq studies of samples treated with pharmacological compounds or subjected to gene expression perturbation (gene knockdown, gene knock-out, up-regulation) were analyzed to capture the transcriptomic change following said intervention. The samples were analyzed as follows:

- Downloaded from SRA, the Sequence Read Archive.
- Quality control was performed following best practices[52]:
    - Initial quality control of reads using FastQC[54] and MultiQC[55].
    - Trimming of adapter sequences using Cutadapt[56].
    - Secondary quality control of trimmed reads using FastQC[54] and MultiQC[55].
- Unstranded studies were converted to stranded to resolve read orientation. **Homebrew pipeline***.

* acknowledgements to **Dr. Yered Pita-Juárez**, Ioannis Vlachos Non-coding Research Lab

- Reverse reads were reverse complemented to resolve sense orientation. **Homebrew pipeline***.
- Reads were kmer-ized using Jellyfish[53] to nucleotide sequences of lengths lengths 27, 30, and 33 bases, to capture the 9-11a.a. length of the antigens.

B.1.3. Generation of Universal Metrics

Kmer-ization of reads allowed for the extraction of the following sets of kmers per treatment, for each kmer length (k = 27, 30, and 33), which can be used downstream for further analysis:

1. Treatment kmers minus Panel of Normals kmers.
2. Treatment kmers minus Control kmers.
3. Treatment kmers minus Panel of Normals kmers **and** Control kmers.

The sets of kmers above were subjected to the following data handling, with the use of R programming language for statistical computing[57], for statistical comparisons and generation of the most fitting metrics:

- Normalization of counts by number of reads sequenced -post-adapter trimming and post- quality control trimming of reads.
- Normalization of counts by number of bases sequenced -post-adapter trimming and post-quality control trimming of reads.
- Fold Change (FC) of the treatment groups with respect to the control groups.
- log2 Fold Change (log_2FC) between treatment and control groups.
- Filtering of treatment kmers:
    1. $log_2FC > 3$
    2. Average coverage per treatment group >= 3 instances per million reads

* acknowledgements to **Dr. Yered Pita-Juárez**, Ioannis Vlachos Non-coding Research Lab

B.2. Capturing the ORFome

Analysis of Ribo-Seq studies reveals information on the translatome, the actively translated RNA molecules for each sample. This approach sheds light on which RNA molecules are actually translated to proteins or peptides. As not everything that is transcribed is going to be translated, the Ribo-Seq analysis acts as an extra layer of filtering kmers for putative antigens.

The Open Reading Frame (ORF) corresponds to the DNA sequences that have the ability to be translated, as they include a start codon -a triplet of bases, AUG- which the ribosome recognizes and, hence, initiates translation. The approach described below, associates Ribo-Seq reads with the statistically most probable ORF coordinates they originated from.

B.2.1. Analysis of Ribo-Seq Studies

Ribo-Seq studies containing healthy samples, as well as cancer samples and cancer samples treated with pharmacological compounds or subjected to gene expression perturbation (gene knockdown, gene knock-out, up-regulation) were analyzed to capture the translatome of healthy, cancer and cancer-treated samples. The samples were analyzed as follows:

- 45 datasets in total were manually curated.
- Samples were downloaded from SRA, the Sequence Read Archive.
- Samples were underwent quality control and analyzed with a **homebrew pipeline*** that utilizes the PRICE method for inference of the active Open Reading Frames (ORFs)[58].
    - Owing to the highly heterogeneous nature of Ribo-Seq library preparations, one needs to take into account how samples were prepared and modify the analysis, especially at the adapter trimming step.

B.2.2. Panels

From the 45 different studies analyzed, 3 different panels were created:

1. Healthy Panel: Containing the ORFs of all analyzed healthy samples. This is an analogue of the Panel of Normals but for actively translated RNA molecules/ ORFs. Healthy samples treated with pharmacological compounds or subjected to gene expression perturbation **were not** included in this panel.
2. Cancer Panel: Containing the ORFs of all analyzed cancer samples.

* acknowledgements to Dr. Yered Pita-Juárez, Ioannis Vlachos Non-coding Research Lab

3. Cancer Plus Panel: Containing the ORFs of all analyzed cancer samples that were treated with pharmacological compounds or were subjected to gene expression perturbation.

Panel	Number of Samples
Healthy	209
Cancer	*73*
Cancer Plus	*57*

Table B.2: Number of Ribo-Seq samples used for the generation of the 3 ORF Panels

To create the aforementioned panels the following steps were carried out:

- Curation of the Ribo-Seq datasets led to categorization of samples based on disease state -healthy or cancer- and treatment methods -control or treated.
- For each sample, ORFs were filtered for False Discovery Rate (FDR) of 10% or less, in order to extract those that are statistically more probable to be translated and exclude false positives or noise.
- Results were gathered per panel type to allow for downstream statistical analysis.

Comparisons between panels was facilitated with the following data handling approaches, and with the use of R[57].

- A list of ORF coordinates from all samples per panel was created, allowing for the comparison of overlapping ORFs between panels.
- A table containing the raw counts for all samples per panel was created. Normalized versions of the table were created with regards to a) the number of reads mapped on the **genome**, and b) the number of reads mapped on the **transcriptome**.
    - The average coverage per ORF per panel was calculated, to allow for direct comparisons.

B.3. Kmer Database

Analysis of dozens of different studies which correspond to hundreds of different samples, and comparisons between them with the end goal of bringing forth the difference in the expression space and kmerome requires fast and robust kmer algebra. This is not an easy feat to do manually, and requires for structures that support such large-scale queries.

To this end, a PostgreSQL Database was created in order to host the results of the analyses described above. The database is designed to include sample metadata as well as kmer post-analysis metadata.

To construct the Kmer DB, the following tools were implemented:

- A first DB schema was created with dbdiagram.io[59], an online tool which allows for user-friendly database schema design and extraction of the SQL code required to build said database.
- PostgreSQL, a free and open-source relational database management system was chosen to host the designed DB[60].
- SQL code extracted from dbdiagram.io was revised and run to create the database. Further improvements to the database structure were carried out with Navicat, a user-friendly Graphical User Interface (GUI) that assist database management[61].

The Database schema is presented below:

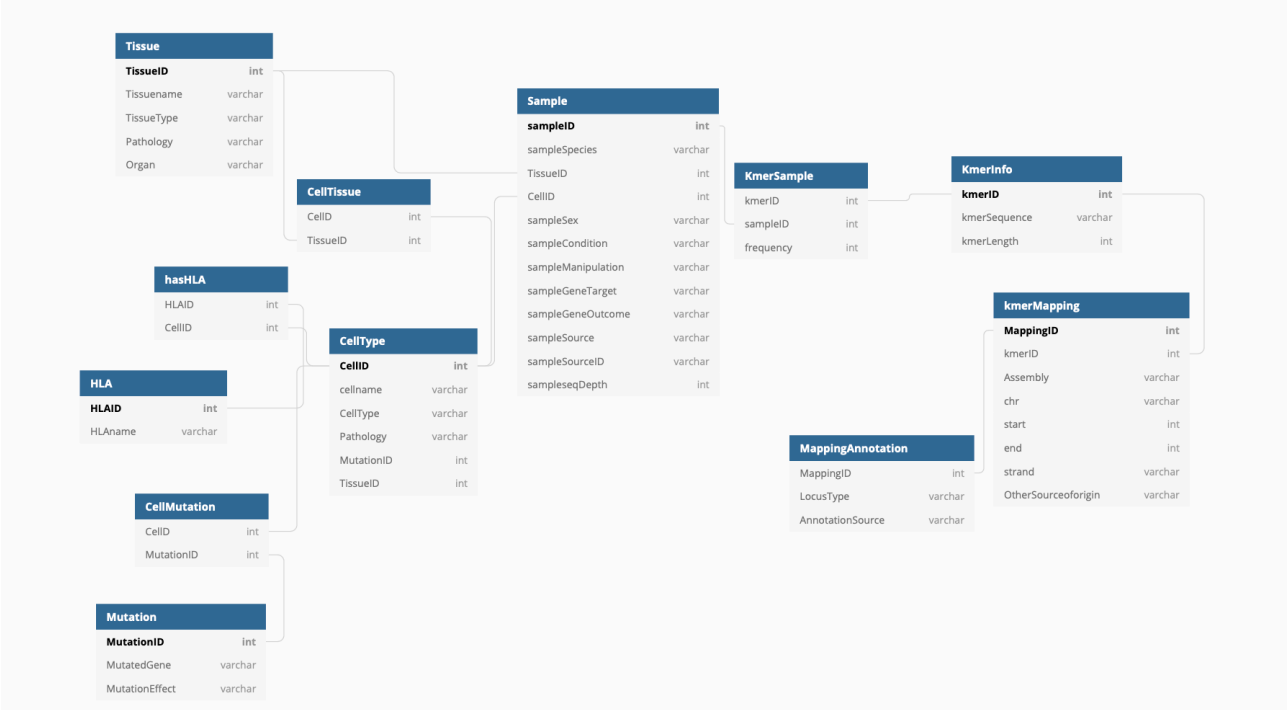

Figure B.1: Kmer database schema. This schema practicaly captures two types of data for a particular study: the sample metadata and the kmer metadata, the latter seeming from the study analysis.

Section C
Results

C.1. Capturing the Kmerome

C.1.1. The Panel of Normals

A total of 210 GTEx samples from different RNA-Seq studies were analyzed in order to capture the kmerome of healthy human cells. Samples from both sexes were included to a 1:1 ratio in order to account for all the sex-associated genetic variation that might be at play for different tissues. Sex-specific tissues were also included, maintaining the 1:1 ratio to samples from other tissues and opposite sex. The only exception was testis tissue, which was entirely excluded from the generation of the Panel of Normals, as testis transcripts have exhibited antigenicity in various cancers (Cancer Testis Antigens, CTAs).

Kmer-ization of the GTEx RNA-Seq studies led to the following results:

kmer Length	Group	Number of kmers
27	**Panel of Normals**	*72,413,616*
30		*70,033,658*
33		*67,444,853*

Table C.1: Number of kmers that comprise the Panel of Normals, for k=27, 30 and 33.

C.1.2. Manipulation Case Study 1

The first case study presented here contains the following samples:

Sample	Condition	Group
S2.1	Control	Control
S2.2	Control	Control
S2.3	Control	Control
S2.4	Treated	Treatment
S2.5	Treated	Treatment
S2.6	Treated	Treatment

Table C.2: Composition of samples comprising RNA-Seq Case Study 1.

In this case study, a certain **exon is being targeted for skipping**, as it had been associated with increased expression in cancer.

For the control group, the number of kmers that follow the criteria below is calculated, per kmer length:

- Kmers present in **all** samples that comprise the control group
- Kmers **not** present in the Panel of Normals

For the treatment group, the number of kmers that follow the criteria below is calculated, per kmer length:

- Kmers present in **all** samples that comprise the treatment group
- Kmers **not** present in the Panel of Normals

and an extra set that follows also this criterion:

- Kmers **not** present in the Control group

This kind of kmer algebra allows for the calculation of a first set of metrics, shown on the table below:

1. Number of kmers present in the control **cancer** sample, and not in healthy
2. Number of kmers present in the treated **cancer** sample, and not in healthy
3. Number of **treatment**-specific kmers

kmer Length	Kmer Set	Number of kmers
27	Control minus PoNs	17,604,370
	Treated minus PoNs	16,717,527
	Treatment-specific	536,488
30	Control minus PoNs	16,953,665
	Treated minus PoNs	15,987,519
	Treatment-specific	493,703
33	Control minus PoNs	16,248,209
	Treated minus PoNs	15,218,827
	Treatment-specific	452,363

Table C.3: Number of kmers for different sets from the control and treated groups, for k=27, 30 and 33.

Already, these numbers capture two things; first of all, the **difference in the expression of kmers between healthy cells and cancer cells**. Taking into account that the PoNs kmerome is as big as 70 million kmers, the additional >15 million observed here that are expressed exclusively in the cancer samples are an important signal for putative antigenicity. Secondly, the third metric of treatment-specific kmers manages to capture **the effect of the treatment on the cancer cells**, acting as a first layer of support for the "manipulation of the kmerome" speculation.

However, the expression of the kmers in question might be a chance occurrence or they may even be sequencing artifacts. The following comparisons act as a second, more strict layer to distinguish between kmers whose expression was a chance event and those who demonstrate a significant difference in expression between control and treatment samples.

After excluding only PoNs kmers from all samples and normalizing with regards to the number of reads sequenced -post trimming-, the Fold Change and $\log_2$(FoldChange) between the treatment group and the control group are calculated. The kmers are then filtered for a $\log_2$(FoldChange) greater than 3 -which corresponds to an 8-fold increase of expression in the treatment group versus the control group-, as well as for at least 3 counts per million reads.

It should be noted that samples that did not give any signal for a specific kmer were assigned 0.01 counts per million reads, which corresponds to 1 count (unnormalized) for the majority of studies.

The histogram below captures the distribution of the $\log_2$(FoldChange) metric, per kmer length.

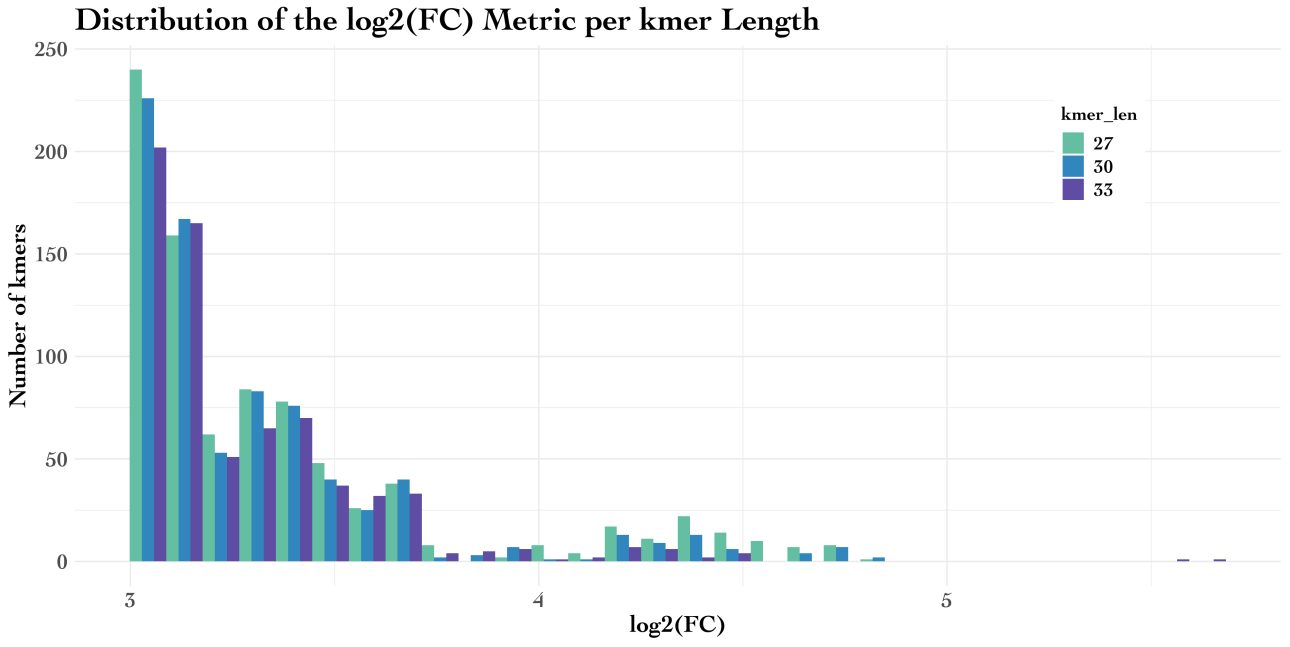

Figure C.1: Histogram of $\log_2$(FoldChange) values for the Treatment group of Case Study 1, k=27, 30, 33.

As expected, most kmers passing filters are gathered near the $\log_2$(FoldChange) > 3 threshold, and only a handful exhibit a $\log_2$(FoldChange) > 4. This filter is already rigid enough, corresponding to an 8-fold change of expression between treatment and control, yet still captures the potential of kmers to be significantly expressed versus a control group.

The following can be noted for the kmers passing the filters:

kmer Length	Group	Number of kmers Passing Filters
27	Treatment	847
30	Treatment	778
33	Treatment	694

Table C.4: Number of kmers passing filters from the Treatment group, for k=27, 30 and 33.

C.1.3. Manipulation Case Study 2

The second case study presented here contains the following samples:

Sample	Condition	Group
S1.1	Treated	Group 1
S1.2	Treated	Group 1
S1.3	Treated	Group 2
S1.4	Treated	Group 2
S1.5	Control	Control
S1.6	Control	Control

Table C.5: Composition of samples comprising RNA-Seq Case Study 2.

In this case study, **pharmacological inhibition of a gene** occurs, with the protein encoded by said gene being linked to various mechanisms of cancer progression. The difference between the two treatment groups is the **concentration of the compound** with which they were treated.

For the control group, the number of kmers that follow the criteria below is calculated, per kmer length:

- Kmers present in **all** samples that comprise the control group
- Kmers **not** present in the Panel of Normals

kmer Length	Group	Number of kmers Passing Filters
27	Control minus PoNs	11,516,285
30	Control minus PoNs	11,732,211
33	Control minus PoNs	11,940,893

Table C.6: Number of kmers minus Pons for the Control group, for k=27, 30 and 33.

The sets of kmers for the control groups and their size manage to capture the difference in kmer expression between cancer cells and healthy cells, taking into account that the PoNs contains about 70 million kmers and the control groups for this study exhibit expression of about 12 million kmers extra.

For the treatment groups, the number of kmers that follow the criteria below is calculated, per kmer length:

- Kmers present in **all** samples that comprise the treatment group
- Kmers **not** present in the Panel of Normals

and an extra set that follows also this criterion:

- Kmers **not** present in the Control group

and, owing to the sample architecture of this study, after excluding both PoNs and Control kmers, the number for the following set of kmers is calculated as well:

- Kmers common between the treatment groups

kmer Length	Group	Number of kmers minus PoNs	Number of treatment-specific kmers	Number of treatment-specific kmers (all treatment groups)
27	Group 1	9,198,792	645,038	184,728
	Group 2	9,544,186	426,260	
30	Group 1	6,069,007	426,845	122,187
	Group 2	9,755,018	437,358	
33	Group 1	9,577,684	674,167	192,454
	Group 2	9,961,257	446,483	

Table C.7: Number of kmers for different sets of the Treatment Groups, k=27, 30, 33.

The sets of kmers for the treatment groups and their size manage to capture the difference in kmer expression between cancer cells and healthy cells, as well as the effect of the treatment. Moreover, similarities between the different treatment groups are observed, which might indicate a universal effect of the treatment across the samples, irregardless of the compound concentration.

After excluding only PoNs kmers from all samples and normalizing with regards to the number of reads sequenced -post trimming-, the Fold Change and $\log_2$(FoldChange) between the treatment groups and the control group are calculated. The kmers are then filtered for a $\log_2$(FoldChange) greater than 3 -which corresponds to an 8-fold increase of expression in the treatment group versus the control group-, as well as for at least 3 counts per million reads.

It should be noted that samples that did not give any signal for a specific kmer were assigned 0.01 counts per million reads, which corresponds to 1 count (unnormalized) for the majority of studies.

The histograms below, one per group, capture the distribution of the $\log_2$(FoldChange) metric, per kmer length.

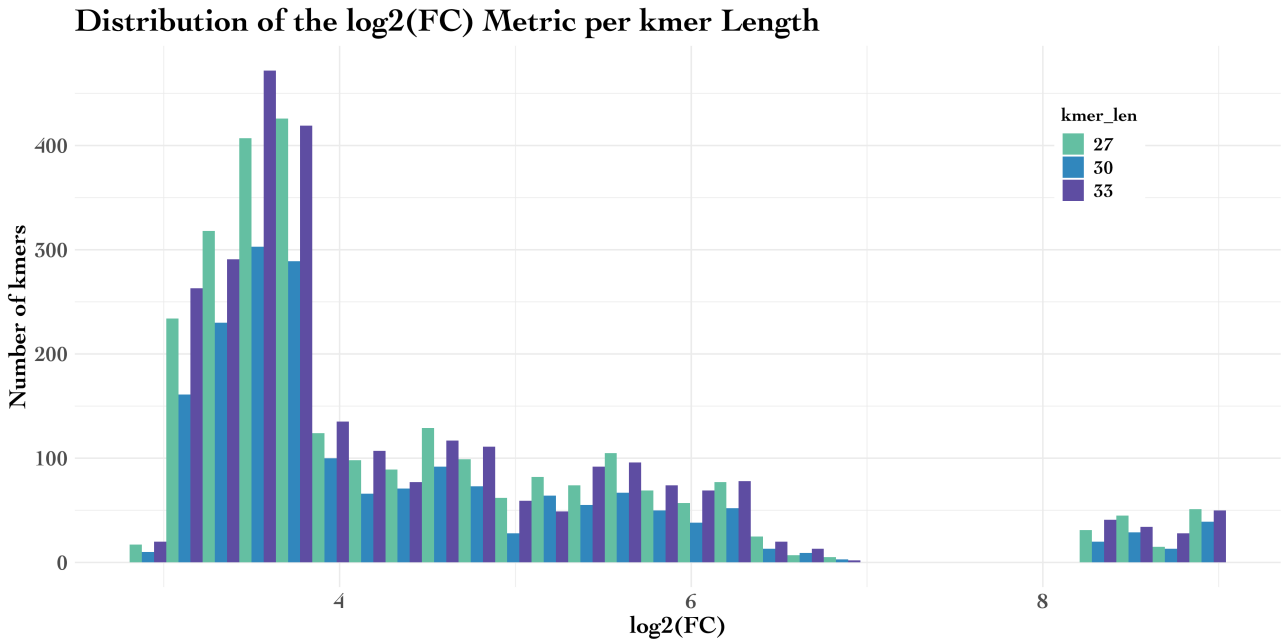

Figure C.2: Histogram of $\log_2$(FoldChange) values for Treatment Group 1 of Case Study 2, k=27, 30, 33.

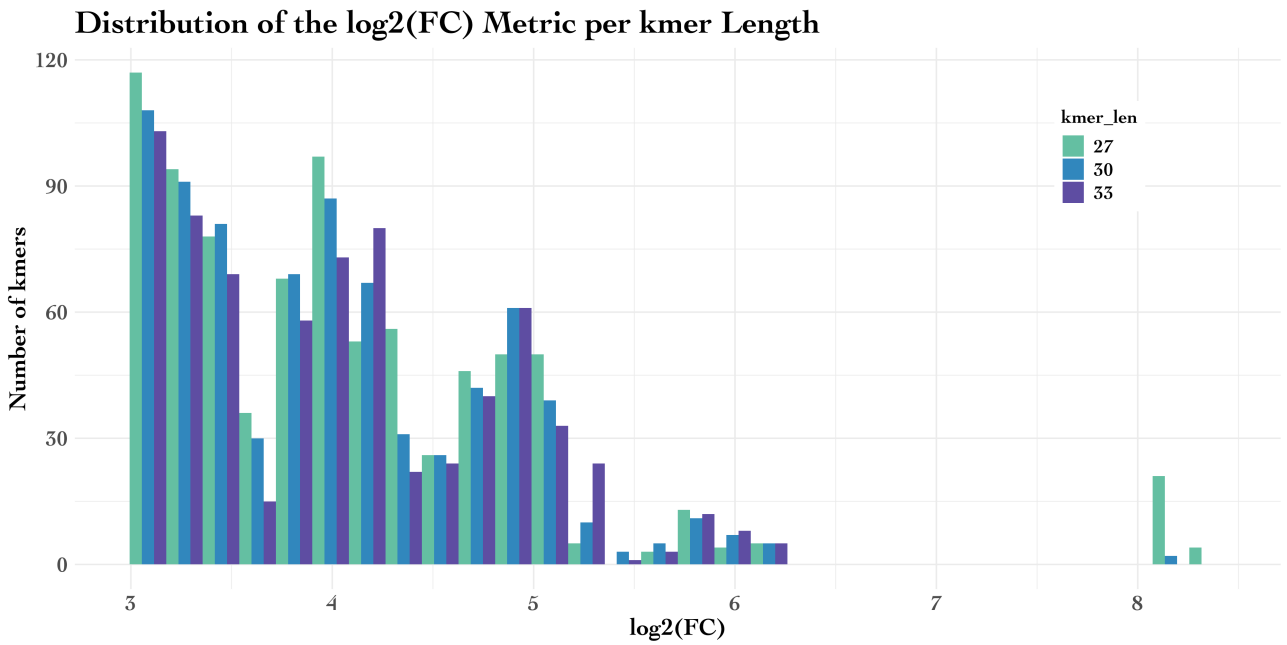

Figure C.3: Histogram of $\log_2$(FoldChange) values for Treatment Group 2 of Case Study 2, k=27, 30, 33.

Both histograms exhibit similar characteristics: the majority of the kmers have a $\log_2$(FoldChange) metric between 3 and 6.5-7, while there are a several outliers with a $\log_2$(FoldChange) metric greater than 8. The latter might be indicative of kmers that were highly expressed in the treatment groups versus non-existent in the control group.

The following can be noted for the kmers passing the filters:

kmer Length	Group	Number of kmers passing filters
27	Group 1	2,646
	Group 2	826
30	Group 1	1,875
	Group 2	775
33	Group 1	2,717
	Group 2	714

Table C.8: Number of kmers passing filters from the Treatment groups, for k=27, 30 and 33.

C.1.4. Manipulation Case Study 3

The third case study presented here contains the following samples:

Sample	Condition	Group
S3.1	Control	Control 1
S3.2	Control	Control 1
S3.3	Treated	Group 1
S3.4	Treated	Group 1
S3.5	Control	Control 2
S3.6	Control	Control 2
S3.7	Treated	Group 2
S3.8	Treated	Group 2
S3.9	Control	Control 3
S3.10	Control	Control 3
S3.11	Treated	Group 3
S3.12	Treated	Group 3
S3.13	Control	Control 4
S3.14	Control	Control 4
S3.15	Treated	Group 4
S3.16	Treated	Group 4

Table C.9: Composition of samples comprising RNA-Seq Case Study 3.

In this case study, samples are treated with **a drug with potential antitumor effects**. **Two different cell types** were utilized in this study, with both cell types originating from the **same type of tumor, thyroid gland medullary carcinoma**; samples S3.1 to S3.8 are of one particular cancer cell type, while S3.9 to S3.16 are of a different one, and all treated samples were treated with the same drug. The differences between treated groups Group 1 against Group 2, and again Group 3 against Group 4 is the **treatment time**.

For the control groups, the number of kmers that follow the criteria below is calculated, per kmer length:

- Kmers present in **all** samples that comprise the control group

- Kmers **not** present in the Panel of Normals

and, owing to the sample architecture of this study, the number for the following sets of kmers are calculated as well:

- Kmers common between control groups of the same cell type
- Kmers common between all control groups

kmer Length	Group	Number of kmers minus PoNs	Number of kmers minus PoNs (same cell type)	Number of kmers minus PoNs (diff cell type)
27	Control 1	3,287,993	2,668,824	1,353,890
	Control 2	4,474,445		
	Control 3	5,434,409	2,517,404	
	Control 4	2,828,316		
30	Control 1	2,918,981	2,346,644	1,167,378
	Control 2	4,003,878		
	Control 3	4,815,948	2,182,813	
	Control 4	2,466,761		
33	Control 1	2,524,309	2,007,132	975,659
	Control 2	3,498,412		
	Control 3	4,157,522	1,836,811	
	Control 4	2,089,916		

Table C.10: Number of kmers for different sets of kmers from the Control groups, k=27, 30, 33.

The sets of kmers for the control groups and their size manage to capture the difference in kmer expression between cancer cells and healthy cells, taking into account that the PoNs contains about 70 million kmers and the control groups for this study exhibit expression of 3 to 5 million kmers extra. Moreover, similarities are observed between kmers for the same cell type, which might be an indication of capturing cell type-specific kmers.

Last but not least, similarities between the different cell type control groups are also observed, with quite the high number of 1.3 million kmers, which might indicate two different things: first, **kmers shared for a specific type of tumor**, as in this case study both cell types come from the same tumor type, and, second, that there might be sequencing or preparation artifacts.

For the treatment groups, the number of kmers that follow the criteria below is calculated, per kmer length:

- Kmers present in **all** samples that comprise the treatment group
- Kmers **not** present in the Panel of Normals

and an extra set that follows also this criterion:

- Kmers **not** present in the Control group

and, owing to the sample architecture of this study, after excluding both PoNs and Control kmers, the number for the following sets of kmers are calculated as well:

- Kmers common between treatment groups of the same cell type
- Kmers common between all treatment groups

kmer Length	Group	Number of kmers minus PoNs	Number of treatment-specific kmers	Number of treatment-specific kmers (same cell type)	Number of treatment-specific kmers (diff cell type)
27	Group 1	3,249,759	572,337	228,646	74,490
	Group 2	2,895,296	731,466		
	Group 3	4,872,708	1,165,253	533,300	
	Group 4	4,499,933	2,436,016		
30	Group 1	2,934,548	525,146	211,127	67,753
	Group 2	2,615,985	675,452		
	Group 3	4,278,361	1,026,668	475,594	
	Group 4	3,976,631	2,199,283		
33	Group 1	2,590,925	469,827	188,769	59,500
	Group 2	2,311,473	608,804		
	Group 3	3,662,071	882,102	413,113	
	Group 4	3,427,422	1,940,992		

Table C.11: Number of kmers for different sets of kmers from the treatment groups, k=27, 30, 33.

The sets of kmers for the treatment groups and their size manage to capture the difference in kmer expression between cancer cells and healthy cells, as well as the effect of the treatment. Moreover, similarities between the different cell type treatment groups are observed, which might indicate the effect of the treatment across cell types or for the specific type of tumor, since both cell types originate from the same tumor type.

After excluding only PoNs kmers from all samples and normalizing with regards to the number of reads sequenced -post trimming-, the Fold Change and $\log_2$(FoldChange)

between the treatment group and the control group are calculated. The kmers are then filtered for a $\log_2$(FoldChange) greater than 3 -which corresponds to an 8-fold increase of expression in the treatment group versus the control group-, as well as for at least 3 counts per million reads.

It should be noted that samples that did not give any signal for a specific kmer were assigned 0.01 counts per million reads, which corresponds to 1 count (unnormalized) for the majority of studies.

The histograms below, one per group, capture the distribution of the $\log_2$(FoldChange) metric, per kmer length.

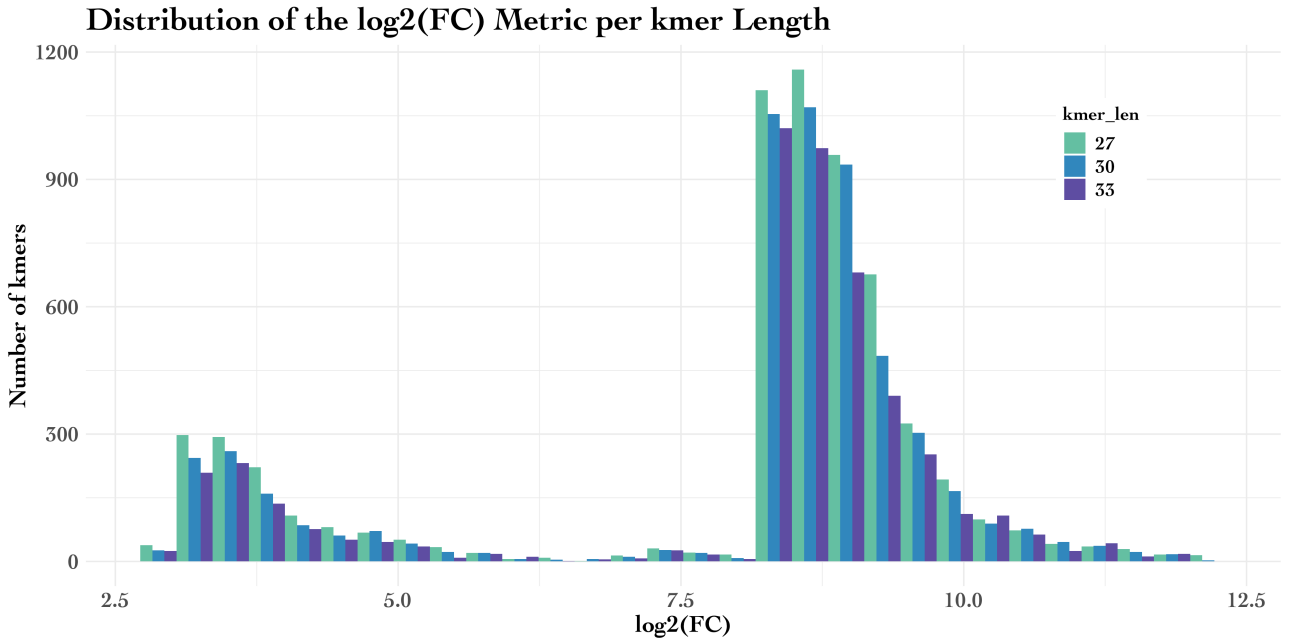

Figure C.4: Histogram of $\log_2$(FoldChange) values for Treatment Group 1 of Case Study 3, k=27, 30, 33.

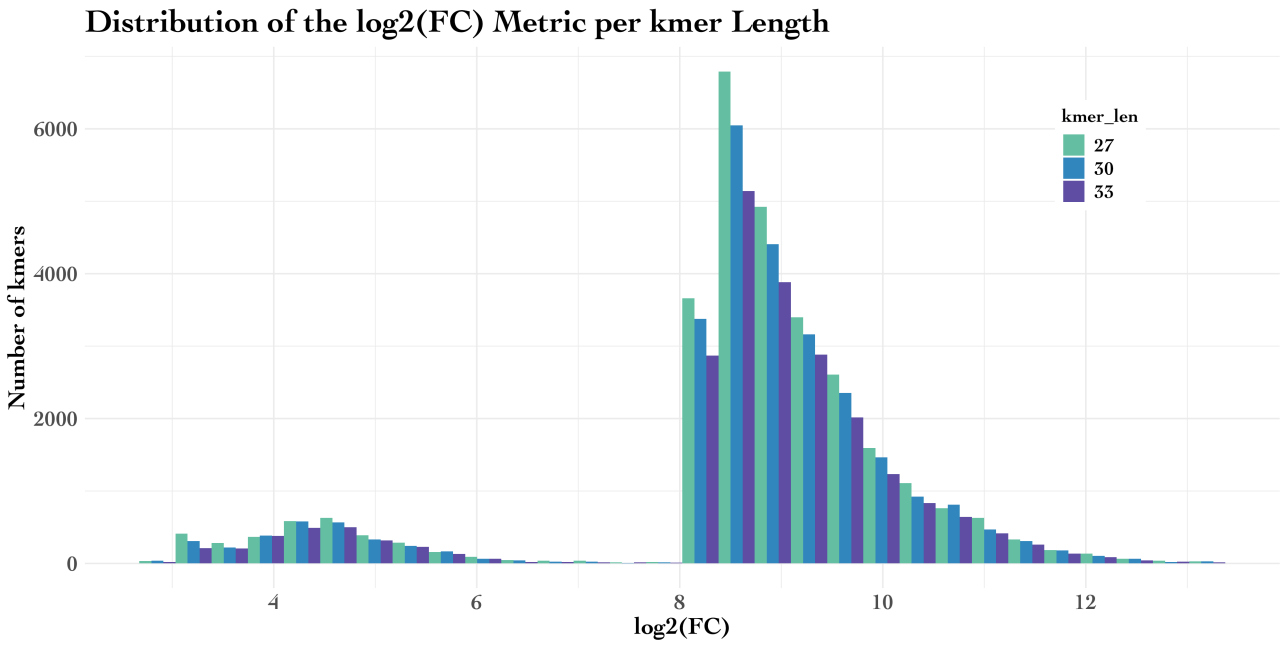

Figure C.5: Histogram of $\log_2$(FoldChange) values for Treatment Group 2 of Case Study 3, k=27, 30, 33.

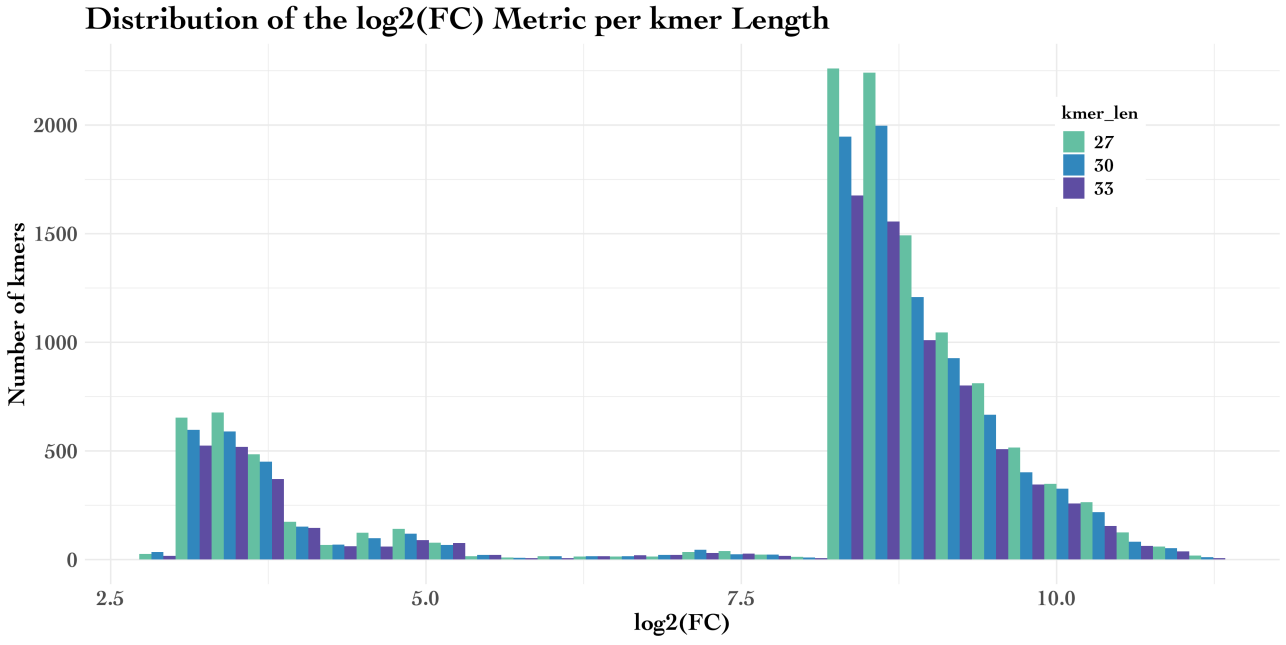

Figure C.6: Histogram of $\log_2$(FoldChange) values for Treatment Group 3 of Case Study 3, k=27, 30, 33.

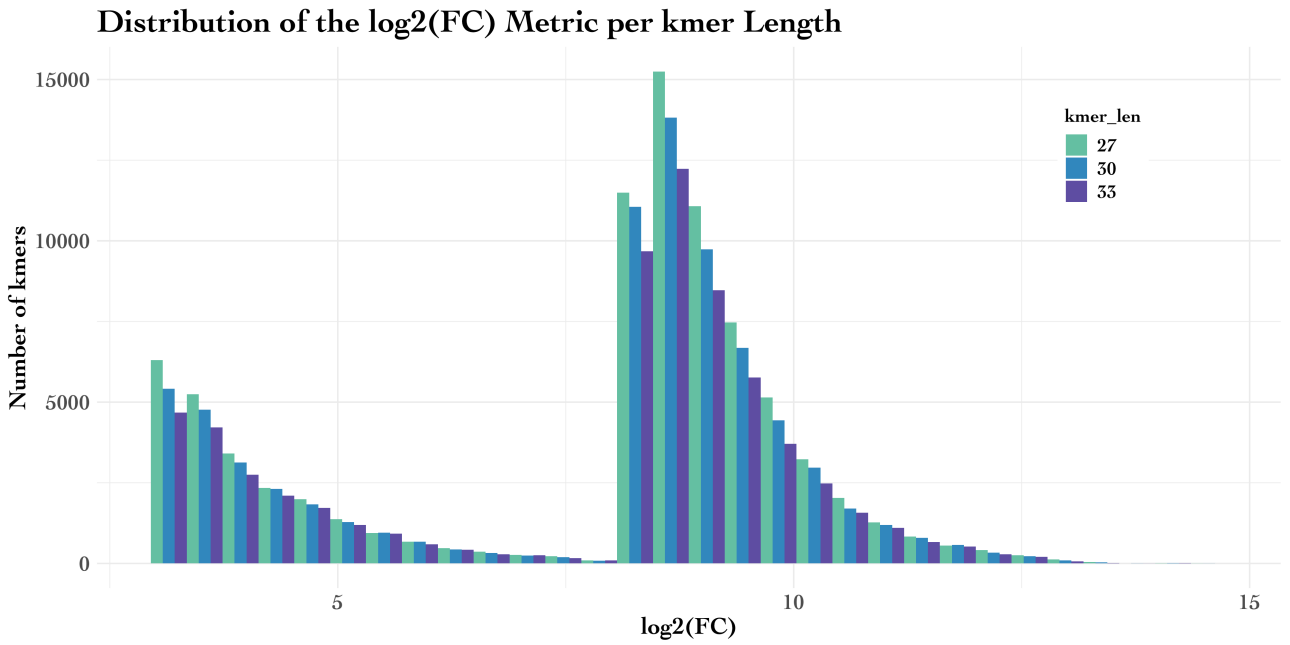

Figure C.7: Histogram of $\log_2$(FoldChange) values for Treatment Group 1 of Case Study 3, k=27, 30, 33.

All four histograms exhibit similar characteristics: a group of kmers with $\log_2$(FoldChange) closest to the threshold, and then a sudden peak in the number of kmers with $\log_2$(FoldChange) a very high $\log_2$(FoldChange) metric greater than 8, which corresponds to 256-fold change in expression between the treatment groups and the control groups. The latter indicates high abundance of kmers in the treatment groups that were non-existent in the control groups. This observation might either indicate a treatment-specific effect of extreme amplitude or an artificial signal that needs to be discarded. Further analysis needs to be carried out in order to distinguish which of the two scenarios are true.

The following can be noted for the kmers passing the filters and which are represented on the histograms above:

kmer Length	Group	Number of kmers passing filters
27	Group 1	6,040
	Group 2	29,638
	Group 3	11,798
	Group 4	82,898

30	Group 1	5,377
	Group 2	26,738
	Group 3	10,212
	Group 4	75,335
33	Group 1	4,608
	Group 2	23,128
	Group 3	8,451
	Group 4	66,157

Table C.12: Number of kmers passing filters from the Treatment groups, for k=27, 30 and 33.

C.2. Capturing the ORFome

Three different ORF panels were created from healthy, cancer and cancer-treated samples, upon analysis of 45 Ribo-Seq studies. Each panel is comprised of the following numbers of ORFs:

Panel	Number of ORFs	Number of Panel-specific ORFs
Healthy	333,461	-
Cancer	151,222	8,969
Cancer Plus	171,013	0

Table C.13: Number of ORFs and panel-specific ORFs per ORF Panel.

The Healthy Panel ORF is considered the Panel of Normals analogue for the Ribo-Seq studies, capturing all ORFs translated in healthy cells, and is thus used as a baseline. Panel-specific ORFs were isolated after panel comparisons. Cancer Panel-specific ORFs do not overlap with any of the Healthy Panel ORFs, while Cancer Plus Panel-specific ORFs do not overlap with any of the Healthy Panel ORFs or the Cancer Panel ORFs. The latter was calculated this way in order to capture potential manipulation-specific translation of ORFs. However, all Cancer Plus Panel ORFs seem to overlap with the other two panels.

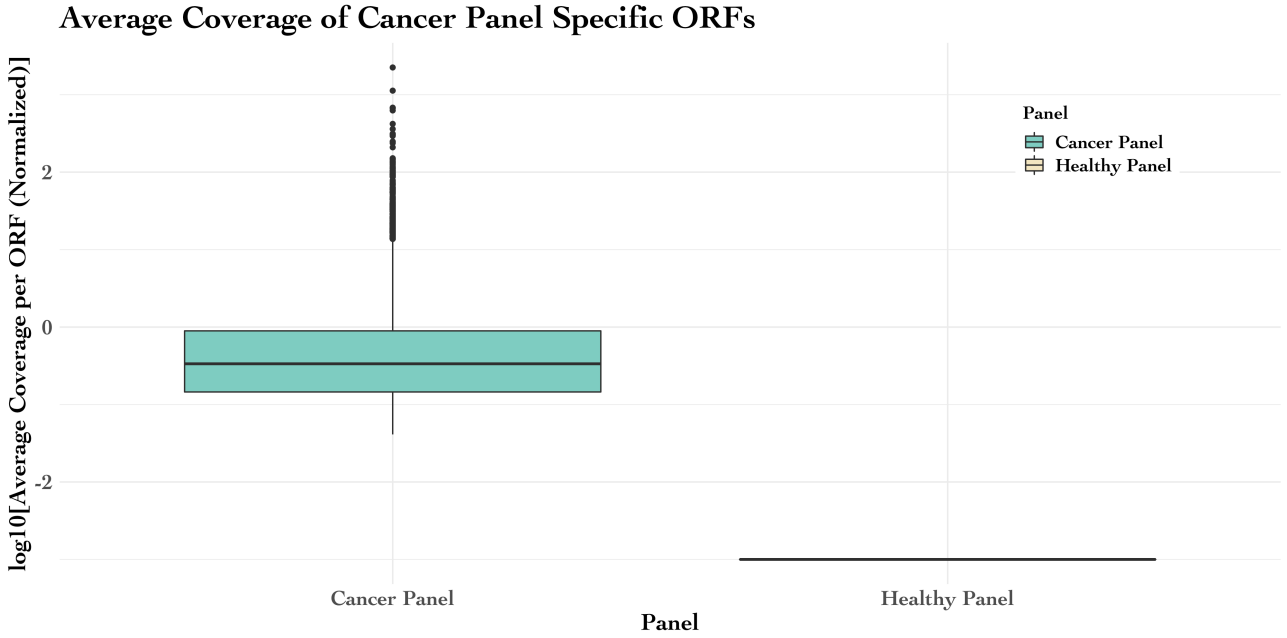

Figure C.8: Boxplot of the $\log_{10}$(Average Coverage per ORF) of Cancer-Specific Panels, capturing that, indeed, said ORFs are cancer-specific.

The plot above captures the average coverage per ORF of the Cancer panel Specific ORFs -log10 representation for clearer depiction. Even though the majority of the Cancer Specific ORFs exhibit a $\log_{10}$(Average Coverage) between -1 and 0, which corresponds to an average coverage of 0.1 to 1, the outliers are strong indicators of cancer-specific translational events.

However, the basic purpose of the ORF Panels, apart from accurately capturing the Translation Space, is to act as an extra layer of filtering for the Kmerome, with the end goal of characterizing the Antigen Space with precision.

C.3. Kmer Algebra with the Kmer DB

The creation of the Kmer Database was based on two fundamental goals:

1. The storing of a large amount of analyzed information in a robust environment
2. The need for efficient and fast querying and filtering.

To this end, the database was created and tested with a small set of analyzed RNA-Seq studies, in order to ensure it has been properly constructed, as well as to examine its user-friendly nature and effectiveness in querying.

kmerinfo@public (local)

kmerid	kmersequence	kmerlength
1	AAAAAAAAAAAAAAAAAAAAAAAAAAA	27
2	GTGTGAAACCTCGAGCCGATGTGGCCT	27
3	TTTGGTGATACTGTCAAGCCCTATCCT	27
4	TGGGATCGAAAGCAGAAAAGCATGTCC	27
5	GGGCTGCACGCAAGAGACAGGATTCTG	27
6	AAACCTAGGTTTAGAGTGTGCTAGGAT	27
7	GGGCCCTTTCGTCCACAGCAAAAATGC	27
8	AGCTTGCCCCTCTCTGCTCCCATGGGG	27
9	AGGGAATCGACACTTCTCAGAAACTTG	27
10	TGGAAGGTCCTGCGTTCAATTTCCTGG	27
11	CCAGTTGGTGCGACTCATAAACAACCT	27
12	TAAAATCAGCAATGAAGAAAGGCACAT	27
13	GATGCTCAGAGCAGCTAATGAAGGGAG	27
14	CAGCCTCTCAAAGTGCTGGGACTACAG	27
15	GCTCACCGCAAGCTCCCCCTCCCGGGT	27
16	AAAACTAAATATCAATAATTCTACAAC	27
17	CATAGAAGGGGACGGAAAAGACGACCC	27
18	TCTTTTGAAATTCGATCTTTTTCCAGA	27
19	GTCCATGGGGCTGGTGGCTGTCATGCT	27
20	CTGGTTGGAGCTCGGGCTGACATAATC	27

kmersample@public (local)

kmerid	sampleid	frequency
1	1	43104
2	1	5
3	1	6
4	1	15
5	1	14
6	1	15
7	1	11
8	1	12
9	1	9
10	1	62
11	1	7
12	1	9
13	1	14
14	1	14
15	1	9
16	1	39
17	1	5
18	1	13
19	1	7
20	1	9

Figure C.9: Two of the tables that comprise the kmer databases, containing kmer information on the sequence, and length (table kmerinfo on the left) and kmer frequency (table kmersample on the right). Table kmersample also has the role of matching the kmer-id and sample-id keys.

The database has exhibited extraordinary capabilities, making kmer comparison across a variety of different fields and parameters easier, faster, and more effective than manually doing so. It is, thus, ready to host the results of the Kmerome analyses.

Section D

Discussion

D.1. Capturing the Expression Space of Antigens

State of the art approaches for capturing the antigen space through the analysis of RNA-Seq samples primarily rely on the annotation of the human genome. They usually focus on certain types of genomic regions that might act as sources for neoantigens -e.g. endogenous retroviruses, non-coding regions, etc.-, as well as on phenomena occurring due to the deregulation of normal procedures -e.g. intron retention, exon skipping or inclusion, non-synonymous mutations, etc.- that lead to aberrant products during expression or translation. Few exceptions of an annotation-free approach can be found in the literature.

The current study opts for the latter, an annotation-free, hypothesis-free approach for capturing the expression space, which in turn shapes a putative antigen space on the expression level. The approach relies on the kmer-ization of the RNA-Seq reads to sequences of lengths 27, 30, and 33 nucleotides long, corresponding to the 9-11 amino acid long antigens. This annotation-free generation of the Kmerome of different samples allows for direct comparisons between normally expressed kmers, as in the case of the Panel of Normals, and kmers expressed in cancer samples.

Kmer differences between cancer and healthy samples can be significant in size, as shown in the three case studies presented in the previous section. The 210 samples that comprise the Panel of Normals have been shown to express about 70 million kmers. In the case studies, results support that not only the cancer samples express kmers different to what healthy cells produce, but they do so in large numbers, reaching even 17 million antigens in one case.

Therefore, it would not be an overstatement to support that the method proposed in the frame of this thesis accurately captures the expression space, without forcing any biases, while constituting an effective method for separating cancer-specific expression events from healthy ones. Further downstream analysis will allow for accurate characterization of the biological nature of these cancer-specific events, and for exclusion of potential artifacts.

D.2. Manipulation of the Antigen Space

In the frame of this thesis, a new approach for sensitizing patients to immunotherapy was proposed: the manipulation of the antigenic and immunogenic profile of cancer patients, via drug administration or gene targeting. The method built with that end goal enables re-analysis of a vast collection of publicly available datasets to harvest of existing yet, up until now, unnoticed transcriptomic information. Moreover, it is, to a large degree, automated and can be further adjusted to allow for massive, parallel, and fast analysis of

RNA-Seq datasets. Furthermore, different layers of metrics have been introduced and tested on their ability to characterize the effect of a certain treatment on the expression space.

Testing of the latter is demonstrated in the present study on three different case studies. The three studies included three different manipulation approaches; targeted exon skipping, targeted inhibition of gene with a pharmacological compound, and administration of drug with previous indications of antitumor effects. Moreover, the study architecture -number of samples, number of control and treatment groups-, treatment time, and treatment concentration were all additional parameters for the case studies.

In effect, none of these parameters were taken into account when analyzing the RNA-Seq samples. On the contrary, all samples were processed in the exact same way, yielding results in a hypothesis-free manner, with the use of universal metrics that allow for cross-sample, cross-study evaluation of treatment effect. Analysis returned a high yield in treatment-specific kmers, with numbers to the hundred thousands for each treatment group. Further testing between comparable treatment groups, as in Case Studies 2 and 3, yielded even shared kmers across the treatment groups. The latter is a strong indicator of treatment-specific effects on the expression space, regardless of variables like treatment time or concentration. Of course, downstream analysis should be carried out to exclude sequencing or preparation artifacts and determine biologically significant outcomes.

An additional layer of metrics was incorporated in order to rigidly filter for important treatment effects. After exclusion of PoNs kmers, kmers were filtered on two bases: first, passing a signal threshold of 3 counts per million reads -treatment group average, and second, passing a $\log_2$(FoldChange) threshold of 3 -which corresponds to an 8-fold increase in expression in the treatment group versus the control group. All case studies resulted in several hundred to thousands of kmers passing aforementioned filters, thus presenting a more adamant argument in favor of the "manipulation of the kmerome" speculation.

D.3. Capturing the Translation Space of Antigens

Undoubtedly, not everything that gets transcribed gets to be translated, and although every kmer is a putative antigen, the vast majority will not reach the immunogenic, or even the antigenic, status. This calls for an extra layer of kmer filtering, this time with input from the translation space.

An exhaustive analysis of publicly available Ribo-Seq datasets was performed, with 45 studies already incorporated into the Open Reading Frame Panels, and with only a dozen

left. Upon analysis of the remaining Ribo-Seq studies, the ORF Panels will be complete, as all publicly available information on Ribo-Seq data will have been integrated.

Generated ORF Panels already capture a significant portion of the effect of cancer on the translation space, with thousands of cancer-specific ORFs identified. Those cancer-specific ORFs show no overlap whatsoever with healthy ORFs, indicating strong effects of tumor on the translation space. This fact alone points to the Cancer Panel being able to act as an additional filter for kmers; in an analogous manner, the Healthy Panel can act as a negative filter for kmers.

D.4. Ongoing Work and Future Prospects

Although this diploma thesis has built sturdy foundations to serve its purpose, it would not be a hyperbole to declare that there are still margins for improvement. To state this, one should recall the final goals of the present study:

1. Creation of universal metrics that capture change in the antigen space upon drug administration and/or gene expression perturbation.
2. Use of aforementioned metrics to prioritize for manipulations that will sensitize patients to immunotherapy.

As far as goal #1 goes, the broader one and the basis of this whole research structure, the following are already being tested or considered for downstream implementation into the pipeline:

1. Filtering of kmers with the ORF Panels.
2. Alignment of kmers to the genome. Characterization of the origin of reads.
3. Transition to a fully automated pipeline which will allow for parallel analysis of massive sets of data.
4. Automated loading of analyzed results to the Kmer Database. Handling of kmer comparisons inside the Kmer Database, utilizing the effectiveness, robustness and speed of DB querying.
5. Incorporation of the MHC-antigen binding score into the universal metrics.
6. Incorporation of a presentation score into the universal metrics, based on Mass Spectrometry Immunopeptidomic data.

Bibliography

1. Early History of Cancer. *American Cancer Society.* https://www.cancer.org/cancer/cancer-basics/history-of-cancer/what-is-cancer.html (2018).

2. Perets E. Cancer Metastasis Put Under the Microscope. *Advanced Science News.* https://www.advancedsciencenews.com/cancer-metastasis-put-microscope/ (2018).

3. NCI Dictionary of Cancer Terms: Cancer. *National Cancer Institute.* https://www.cancer.gov/publications/dictionaries/cancer-terms/def/cancer (2011).

4. Metastatic Cancer. *National Cancer Institute.* https://www.cancer.gov/types/metastatic-cancer (2015).

5. Risk Factors for Cancer. *National Cancer Institute.* https://www.cancer.gov/about-cancer/causes-prevention/risk (2015).

6. Gisselsson, D. Intratumor Diversity and Clonal Evolution in Cancer—A Skeptical Standpoint. in *Advances in Cancer Research* vol. 112 1–9 (Elsevier, 2011).

7. Ritchie, H. & Roser, M. Causes of Death. *Our World in Data.* https://ourworldindata.org/causes-of-death (2018).

8. Roser, M. & Ritchie, H. Cancer. *Our World in Data.* https://ourworldindata.org/cancer (2015).

9. The Human Genome Project. *National Human Genome Research Institute.* https://www.genome.gov/human-genome-project (2020).

10. Behjati, S. & Tarpey, P. S. What is next generation sequencing? *Arch Dis Child Educ Pract Ed* **98**, 236–238 (2013).

11. Slatko, B. E., Gardner, A. F. & Ausubel, F. M. Overview of Next Generation Sequencing Technologies. *Curr Protoc Mol Biol* **122**, e59 (2018).

12. The Irvine Lab. *Koch Institute for Integrative Cancer Research at MIT*. https://irvine-lab.mit.edu/ (2020).

13. Keshavarz-Fathi, M. & Rezaei, N. Chapter 1 - Cancer Immunology. in *Vaccines for Cancer Immunotherapy* (eds. Rezaei, N. & Keshavarz-Fathi, M.) 1–17 (Academic Press, 2019).

14. Rich, R. R. *Clinical immunology: principles and practice.* (Elsevier, Mosby, 2008).

15. McCullough, K. C. & Summerfield, A. Basic Concepts of Immune Response and Defense Development. *ILAR J* **46**, 230–240 (2005).

16. Thomas, L. On immunosurveillance in human cancer. *Yale J Biol Med* **55**, 329–333 (1982).

17. Burnet, F. M. The Concept of Immunological Surveillance. *Immunological Aspects of Neoplasia* **13**, 1–27 (1970).

18. Beatty, G. L. & Gladney, W. L. Immune Escape Mechanisms as a Guide for Cancer Immunotherapy. *Clinical Cancer Research* **21**, 687–692 (2015).

19. Dunn, G. P., Old, L. J. & Schreiber, R. D. The Immunobiology of Cancer Immunosurveillance and Immunoediting. *Immunity* **21**, 137–148 (2004).

20. Schumacher, T. N., Scheper, W. & Kvistborg, P. Cancer Neoantigens. *Annu. Rev. Immunol.* **37**, 173–200 (2019).

21. Brennick, C. A., George, M. M., Corwin, W. L., Srivastava, P. K. & Ebrahimi-Nik, H. Neoepitopes as cancer immunotherapy targets: key challenges and opportunities. *Immunotherapy* **9**, 361–371 (2017).

22. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. Chapter 8. in *Molecular biology of the cell* (5th ed.). New York: Garland Science. p. 550 (2008).

23. Schopf, F. H., Biebl, M. M. & Buchner, J. The HSP90 chaperone machinery. *Nature Reviews Molecular Cell Biology* **18**, 345–360 (2017).

24. Kovacsovics-Bankowski, M. & Rock, K. A phagosome-to-cytosol pathway for exogenous antigens presented on MHC class I molecules. *Science* **267**, 243–246 (1995).

25. Pettit, S., Seymour, K., O'Flaherty, E. & Kirby, J. Immune selection in neoplasia: towards a microevolutionary model of cancer development. *British Journal of Cancer 7* (2000).

26. Sharma, P., Hu-Lieskovan, S., Wargo, J. A. & Ribas, A. Primary, Adaptive and Acquired Resistance to Cancer Immunotherapy. *Cell* **168**, 707–723 (2017).

27. NCI Dictionary of Cancer Terms: Tumor Microenvironment. *National Cancer Institute.* https://www.cancer.gov/publications/dictionaries/cancer-terms/def/tumor-microenvironment (2011).

28. The Tumour Microenvironment. Nature Reviews Cancer. *Nature*. https://www.nature.com/collections/khylqkxqbr.

29. Tormoen, G. W., Crittenden, M. R. & Gough, M. J. Role of the immunosuppressive microenvironment in immunotherapy. *Adv Radiat Oncol* **3**, 520–526 (2018).

30. The Nobel Prize in Physiology or Medicine 2018. *NobelPrize.org* https://www.nobelprize.org/prizes/medicine/2018/summary/.

31. Guo, Z. S. The 2018 Nobel Prize in medicine goes to cancer immunotherapy (editorial for BMC cancer). *BMC Cancer* **18**, (2018).

32. Immunotherapy for Cancer. *National Cancer Institute*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy (2015).

33. Types of immunotherapy. *Cancer Research UK*. https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/immunotherapy/types.

34. Monoclonal Antibodies. *National Cancer Institute*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/monoclonal-antibodies (2019).

35. Monoclonal antibodies (MABs). *Cancer Research UK*. https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/immunotherapy/types/monoclonal-antibodies.

36. T-cell Transfer Therapy. *National Cancer Institute*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/t-cell-transfer-therapy (2019).

37. CAR T-cell therapy. *Cancer Research UK*. https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/immunotherapy/types/CAR-T-cell-therapy.

38. Immune System Modulators for Cancer Therapy. *National Cancer Institute*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/immune-system-modulators.

39. Cytokines. *Cancer Research UK*. https://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/immunotherapy/types/cytokines.

40. Immune Checkpoint Inhibitors. *National Cancer Institute*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/checkpoint-inhibitors (2019).

41. Cancer Treatment Vaccines. *National Cancer Institute*. https://www.cancer.gov/about-cancer/treatment/types/immunotherapy/cancer-treatment-vaccines (2019).

42. Why Cancer Immunotherapy? *Cancer Research Institute (CRI)*. https://www.cancerresearch.org/immunotherapy/why-immunotherapy.

43. Galluzzi, L., Chan, T. A., Kroemer, G., Wolchok, J. D. & López-Soto, A. The hallmarks of successful anticancer immunotherapy. *Sci. Transl. Med.* **10**, eaat7807 (2018).

44. Hegde, P. S. & Chen, D. S. Top 10 Challenges in Cancer Immunotherapy. *Immunity* **52**, 17–35 (2020).

45. Ventola, C. L. Cancer Immunotherapy, Part 3: Challenges and Future Trends. *P T* **42**, 514–521 (2017).

46. Smart, A. C. *et al.* Intron retention is a source of neoepitopes in cancer. *Nat Biotechnol* **36**, 1056–1058 (2018).

47. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, eaau5516 (2018).

48. Barbieri, I. & Kouzarides, T. Role of RNA modifications in cancer. *Nature Reviews Cancer* 1–20 (2020).

49. Kim, H. *et al.* PRMT5 control of cGAS/STING and NLRC5 pathways defines melanoma response to antitumor immunity. *Sci. Transl. Med.* **12**, eaaz5683 (2020).

50. Germano, G. *et al.* Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116–120 (2017).

51. GTEx Portal. https://gtexportal.org/home/.

52. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol* **17**, 13 (2016).

53. Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* (2011) 27(6): 764-770

54. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).

55. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

56. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011).

57. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/. (2020)

58. erhard-lab/gedi. *GitHub.* https://github.com/erhard-lab/gedi.

59. dbdiagram.io - Database Relationship Diagrams Design Tool. https://dbdiagram.io/home.

60. PostgreSQL. https://www.postgresql.org/.

61. Navicat GUI. https://www.navicat.com/en/.

Appendix I: List of Abbreviations

APCs	Antigen Presenting Cells
CAR	Chimeric Antigen Receptor
CTAs	Cancer Testis Antigens
CTLs	Cytotoxic T Cells
CTLA-4	Cytotoxic T-Lymphocyte-Associated protein 4
DAMPs	Damage-Associated Molecular Patterns
DB	Database
DCs	Dendritic Cells
DNA	Deoxyribonucleic Acid
ER	Endoplasmic Reticulum
ERAP	Endoplasmic Reticulum Aminopeptidase
FTO	Fat Mass and Obesity-associated Protein
GTEx	Genotype-Tissue Expression
HSP90	Heat Shock Protein 90
ICI	Immune Checkpoint Inhibition
IFNs	Interferons
MABs	Monoclonal Antibodies
MHC	Major Histocompatibility Complex
MMR	DNA Mismatch Repair
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PAMPs	Pathogen-Associated Molecular Patterns
PD-1	Programmed Cell Death Protein 1
PD-L1	Programmed Death-Ligand 1
PoNs	Panel of Normals
PRMT5	Protein Arginine N-Methyltransferase 5
PRR	Patter Recognition Receptor
Ribo-Seq	Ribosome Profiling
RNA	Ribonucleic Acid
RNA-Seq	RNA Sequencing
SNV	Single-Nucleotide Variant

TAP Transporter Associated with Antigen Processing

TCR T Cell Receptor

TILs Tumor Infiltrating Lymphocytes

TMB Tumor Mutation Burden

TME Tumor Microenvironment

TSAs Tumor Specific Antigens

WES Whole Exome Sequencing

WGS Whole Genome Sequencing