Silesian University of Technology

dr n. med. ALEXANDER CORTEZ

MOLECULAR MECHANISMS OF TUMOR CELL RESISTANCE TO THE FGFR KINASE INHIBITOR

DOCTORAL DISSERTATION

Doctoral dissertation performed under the supervision of
prof. dr hab. inż. Joanna Polańska
Department of Data Science and Engineering
Faculty of Automatic Control, Electronics and Computer Science
Silesian University of Technology
prof. dr hab. n. med. Katarzyna Lisowska
at Mieczysław Chorąży Center for Translational Research
and Molecular Biology of Cancer
Maria Sklodowska-Curie National Research Institute of Oncology
Gliwice Branch

Gliwice 2023

Pragnę serdecznie podziękować mojej Promotorce Pani prof. dr hab. Katarzynie Lisowskiej za umożliwienie mi rozpoczęcia kariery naukowej i za pomoc w utrzymaniu się na tej ścieżce, za pomoc merytoryczną i przekazaną mi wiedzę oraz za ukierunkowanie naukowe i etyczne.

Dziękuję również mojej Pani Promotor prof. dr hab. inż. Joannie Polańskiej za umożliwienie udziału w programie Applied Integrative Data Analysis i tym samym umożliwienie realizacji niniejszej pracy doktorskiej. W szczególności dziękuję za zmotywowanie i ukierunkowanie w trakcie pisania rozprawy.

Pani mgr Agacie Wilk dziękuję za wysiłek i wiedzę przekazaną mi w trakcie realizacji prac badawczych oraz serdeczność i koleżeńskość.

Jestem wdzięczny wszystkim pracownikom Działu Analiz Bioinformatyczno Biostatystycznych za niebywałą cierpliwość i wyrozumiałość okazaną mi w trakcie pisania pracy doktorskiej i realizacji prac badawczych.

Wszystkim pracownikom Centrum Badań Translacyjnych i Biologii Molekularnej Nowotworów, a zwłaszcza kolegom i koleżankom Doktorantom dziękuję za przychylność, dobrą atmosferę i wsparcie okazane mi w trakcie realizacji pracy doktorskiej.

Dziękuje moim Rodzicom i mojej Żonie Magdalenie za wsparcie i wyrozumiałość oraz ogromną cierpliwość.

Prace doktorską dedykuję

mojej Żonie i Dzieciom oraz moim Rodzicom

Content

List of abbreviations

Abbreviation/Acronym	Full name/Description
ABCG2	drug efflux transporter
ABL	Abelson murine leukemia viral oncogene homolog 1
ADC	adenocarcinoma
AFP	alpha-fetoprotein
AKT	Ak strain transforming
APOL1	Apolipoprotein L1
AUC^{ROC}	area under the receiver operating characteristic
B1	RT-112
B2	UM-UC-14
BEST	"Biomarkers, Endpoints, and other Tools"
BNP	brain natriuretic peptide
BRAF	v-raf murine sarcoma viral oncogene homolog B1
BRCA1/2	breast cancer genes 1 and 2
CBFS	correlation-based feature selection
CCF	cancer cell fraction
CCNB2	Cyclin B2
CDKN2A	Cyclin Dependent Kinase Inhibitor 2A
cDNA	complementary DNA
CDT1	Chromatin Licensing And DNA Replication Factor 1
CEA	carcinoembryonic antigen
CELONKO	"Development of modern biomarkers and development of an innovative FGFR kinases inhibitor"
CENPO	Centromere Protein O
CFTR	cystic fibrosis transmembrane conductance regulator
CI	confidence interval
CKD	chronic kidney disease
CML	chronic myeloid leukemia
COAs	clinical outcome assessments
COPD	chronic obstructive pulmonary disease
CPL304110	pan-FGFR inhibitor manufactured by Celon Pharma S.A.
CRC	colorectal cancer
CTCs	circulating tumor cells
DCA	deep count autoencoder

DEGs	differentially expressed genes
DGE	differential gene expression
DMEM/F12	Dulbecco's Modified Eagle Medium (DMEM): Nutrient Mixture F-12
DNN	deep neural network
DTC	differentiated thyroid cancer
DTIC	dacarbazine
DUSP6	Dual Specificity Phosphatase 6
EGFR	Epidermal growth factor receptor
EMEM	eagle's minimum essential medium
EMT	epithelial-mesenchymal transition
ES	enrichment score
FC	fold change
FCS	functional class scoring
FDA	Food and Drug Administration
FDR	false discovery rate (q value)
FGFR	fibroblast growth factor receptor
GA	genetic algorithms
GIST	gastrointestinal stromal tumors
GLM	Generalized Linear Model
GPU	graphics processing unit
GRB7	Growth Factor Receptor Bound Protein 7
GSEA	gene set enrichment analysis
HCC	hepatocellular carcinoma
HER2	epidermal growth factor receptor 2
hiPathia	HIgh throughput PATHway Interpretation and Analysis
HTBD	high-throughput biological data
HTS	high-throughput sequencing
ICA	independent component analysis
IHC	immunohistochemistry
IQGAP2	IQ Motif Containing GTPase Activating Protein 2
KRAS	Kirsten rat sarcoma virus
L1	NCI-H1581
L2	NCI-H1703
LCC	large cell carcinoma
LC-MS	liquid chromatography-mass spectrometry

log	logarithm
MAF	minor allele frequency
MAPK	Mitogen-Activated Protein Kinase
MATH	mutant-allele tumor heterogeneity
MIBC	muscle-invasive bladder cancer
minDiff	minimal difference
minFC	minimal fold change
MPSS	massively parallel signature sequencing
NCBiR	National Centre for Research and Development
NIH	National Institutes of Health
NMIBC	non-muscle-invasive bladder cancer
NMR	nuclear magnetic resonance
NOTCH1	Neurogenic locus notch homolog protein 1
NPV	negative predictive value
NSCLC	non-small cell lung cancer
ORA	over-representation analysis
PA	pathway analysis
PARP	Poly (ADP-ribose) polymerase
PC	principal component
PCA	principal component analysis
PDBs	pathway databases
PI3K	phosphatidylinositol 3-kinase
PPV	positive predictive value
PREDICT	Pipeline for Rapid Evaluation and Discovery of Important bi-omarker CandidaTes
PSA	prostate-specific antigen
PTB	pathway topology based
PTEN	Phosphatase and tensin homolog deleted on chromosome ten
R	resistant cell line
RAF	rapidly accelerated fibrosarcoma
RAS	Rat sarcoma
RASA1	ras p21 protein activator 1
RBM14	RNA Binding Motif Protein 14
RCC	renal cell carcinoma
RDS	respiratory distress syndrome
RFE	recursive feature elimination

RHEB	Ras Homolog, MTORC1 Binding
RNA-seq	RNA sequencing
RND1	Rho Family GTPase 1
ROC	receiver operating characteristic
ROS	reactive oxygen species
RPMI 1640	Roswell Park Memorial Institute 1640 (cell culture medium)
S	sensitive cell line
S1	SNU-16
S2	KATO III
SAGE	serial analysis of gene expression
SCC	squamous cell carcinoma
scRNA-seq	single-cell RNA sequencing
SFS	sequential forward selection
SIMLR	single-cell interpretation via multi-kernel learning
SNE	stochastic neighbor embedding
SSRP1	Structure Specific Recognition Protein 1
STAT	signal transducer and activator of transcription
STS	soft-tissue sarcomas
TCR	T cell receptor
TKIs	tyrosine kinase inhibitors
TP53	transformation-related protein 53
t-SNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
UV	ultraviolet light
WLS	Wnt Ligand Secretion Mediator
ZIFA	zero-inflated factor analysis
ZINB	zero-inflated negative binomial

Abstract

Fibroblast Growth Factor Receptor (FGFR) signaling constitutes one of the most prominent pathways involved in cell growth and development as well as cancer progression. All members of the FGFR family have oncogenic gene alterations involved in some human cancers. For instance, FGFR1 amplification is found in the bladder, gastric, breast, and lung cancers, while liver, uterine, lung, and gastric cancers may exhibit FGFR2 amplification, mutations, and fusions. Bladder and lung cancers frequently display FGFR3 mutations and fusions. This indicates that FGFR is a potential target for the new anti-cancer treatment.

This study was aimed at the identification of potential biomarkers indicating cancer cells' sensitivity/resistance toward a novel small-molecule pan-FGFR inhibitor developed by Polish pharmaceutical company Celon Pharma S.A. Within previous project (CELONKO project; STRATEGMED II program financed by NCBR) RNA sequencing (RNA-seq) experiment was conducted on cell lines that were either resistant or sensitive to that inhibitor. Using the RNA-seq data, a comprehensive analysis of gene expression in cell lines from three different cancer types (lung, stomach, and bladder) was performed to identify potential predictive biomarkers related to mechanisms of FGFR tyrosine kinase inhibitors (FGFR-TKIs) resistance.

To address the limitations of standard analytical methods in low sample size experiments, which often yield results that do not meet the requirements of clinically suitable biomarkers, the "Pipeline for Rapid Evaluation, and Discovery of Important biomarker CandidaTes" (PREDICT) was developed. Applying statistical properties implemented in the PREDICT pipeline, resulted in smaller numbers of candidate biomarkers, however, with more promising properties. Importantly, by removing numerous uncertain candidates, PREDICT pipeline application may reduce the number of entities entering the validation phase what could lead to cost- and effort reduction in biomarker discovery.

Based on signaling pathway analysis, combined with the use of PREDICT pipeline and literature search, it was possible to uncover the link with potential resistance mechanisms towards FGFR-TKIs for the majority of selected genes. These findings indicate that resistant tumor cells exhibit compensatory activation of pathways regulating cell proliferation, migration rate, survival, invasiveness, and antiapoptotic properties, in response to FGFR-TKIs treatment.

By comparing gene sets selected in three different cancer types, several potentially universal biomarkers of FGFR-TKIs resistance were identified, including *SSRP1* (Structure Specific Recognition Protein 1), *CCNB2* (Cyclin B2), *CDT1* (Chromatin Licensing And DNA Replication Factor 1), and *CENPO* (Centromere Protein O). These genes were commonly dysregulated in both stomach and bladder cancer and showed the same direction of change in expression in these two cancer types. They may serve as universal biomarkers for predicting FGFR-TKIs resistance in patients with diagnosed stomach or bladder cancer.

In conclusion, the use of the PREDICT pipeline led to the filtering out the unwanted results, and the selected biomarker candidates possess characteristics suitable for a biomarker that can be applied in clinical settings. An extensive literature search uncovered the link with potential resistance mechanisms towards FGFR-TKIs for the majority of selected genes. The next step in biomarker development would be validation/qualification phase to confirm that the differential expression observed in the discovery phase can be seen using other methods and on the different biological material.

Streszczenie

Sygnalizacja poprzez receptory czynników wzrostu fibroblastów (*ang. Fibroblast Growth Factor Receptor,* FGFR) stanowi ważny mechanizm regulujący procesy proliferacji i różnicowania komórki. W wielu nowotworach mechanizm ten jest zaburzony, a główną przyczyną są różnego typu nieprawidłowości genomowe. Na przykład amplifikacja FGFR1 występuje w raku pęcherza moczowego, żołądka, piersi i płuc, podczas gdy w raku wątroby, macicy, płuc i żołądka może wystąpić zarówno amplifikacja, jak i mutacje oraz fuzje FGFR2 z innymi genami. Mutacje i fuzje FGFR3 często występują w raku pęcherza moczowego i płuc. Dlatego hamowanie sygnalizacji FGFR jest przedmiotem badań i prób klinicznych w ramach rozwoju nowych terapii celowanych.

Celem niniejszej pracy było zidentyfikowanie potencjalnych biomarkerów związanych z wrażliwością/opornością komórek nowotworowych na nowy małocząsteczkowy inhibitor FGFR opracowany przez polską firmę farmaceutyczną Celon Pharma S.A. W ramach wcześniejszego projektu (projekt CELONKO, program STRATEGMED II sfinansowany przez NCBR) przeprowadzono eksperyment sekwencjonowania RNA (*ang. RNA sequencing,* RNA-seq) na liniach komórkowych opornych lub wrażliwych na ten inhibitor. Korzystając z danych z eksperymentu RNA-seq, przeprowadzono kompleksową analizę profilu ekspresji genów w liniach komórkowych z trzech różnych typów nowotworów (płuca, żołądka i pęcherza moczowego), aby zidentyfikować potencjalne biomarkery predykcyjne związane z mechanizmami oporności na inhibitor FGFR.

Zestaw danych z sekwencjonowania DNA charakteryzował się niską liczebnością próbek, co jest typowym ograniczeniem wielu podobnych eksperymentów. Standardowe metody analizy nie radzą sobie dobrze z tym typem danych. Ponadto, brak odpowiednich filtrów sprawia, że wyniki mogą nie spełniać wymagań stawianych biomarkerom do zastosowań klinicznych. Dlatego, w ramach niniejszej pracy opracowano schemat obliczeniowy nazwany „*Pipeline for Rapid Evaluation, and Discovery of Important biomarker CandidaTes*" (PREDICT). Zastosowanie własności statystycznych zaimplementowanych w schemacie PREDICT pozwoliło wyselekcjonować mniejsze liczby potencjalnych biomarkerów, ale o bardziej obiecujących cechach. Eliminacja niepewnych kandydatów na etapie obliczeniowym, dzięki zastosowaniu schematu PREDICT, pozwoli na redukcję kosztów i wysiłku na etapie walidacji, który jest kolejną fazą rozwoju potencjalnego biomarkera.

Na podstawie analizy szlaków sygnałowych, połączonej z użyciem schematu PREDICT oraz przeglądem literatury, odkryto związek z potencjalnymi mechanizmami oporności na inhibitor FGFR dla większości wyselekcjonowanych genów. Otrzymane wyniki wskazują, że komórki uodpornione na działanie inhibitora FGFR wykształciły kompensacyjną aktywację szlaków regulujących proliferację komórek, tempo migracji, przeżycie, inwazyjność i hamowanie apoptozy.

Porównując zestawy genów wyselekcjonowane w trzech różnych typach raka, zidentyfikowano kilka potencjalnie uniwersalnych biomarkerów oporności na inhibitory FGFR, a mianowicie *SSRP1* (*ang. Structure Specific Recognition Protein 1*), *CCNB2* (*ang. Cyclin B2*), *CDT1* (*ang. Chromatin Licensing And DNA Replication Factor 1*) i *CENPO* (*ang. Centromere Protein O*). Te geny miały zmienioną ekspresję zarówno w raku żołądka, jak i pęcherza moczowego i wykazywały ten sam kierunek zmiany ekspresji w obydwu typach raka. Dlatego mogą one służyć jako uniwersalne biomarkery do predykcji oporności na inhibitory FGFR u pacjentów ze zdiagnozowanym rakiem żołądka lub pęcherza moczowego.

Podsumowując, użycie schematu PREDICT skutkuje odfiltrowaniem niepożądanych wyników, a wyselekcjonowane geny kandydackie posiadają cechy odpowiednie dla biomarkera, który może znaleźć praktyczne zastosowanie kliniczne. Przegląd literatury pozwolił na określenie związku większości wyselekcjonowanych genów z potencjalnymi mechanizmami oporności na inhibitory FGFR. Wytypowani kandydaci w kolejnym kroku ich rozwoju jako biomarkera powinni być włączeni do fazy walidacji, z zastosowanie różnych metod i na różnym materiale biologicznym.

I. Introduction, hypothesis, aim

Cancer is an increasingly prevalent disease that affects millions of people worldwide. The detection of cancer is carried out using various techniques such as imaging, tissue biopsies, and blood tests. These methods are essential in the early diagnosis of cancer, which is critical for successful treatment and improving patient outcomes [1].

One of the elements of cancer diagnosis, monitoring, prognosis, and personalized treatment is the evaluation of different biomarkers. Biomarkers are measurable substances found in the blood, tissues, or other body fluids that indicate the presence of cancer or the risk for cancer development. Biomarkers are also useful tools in the monitoring and treatment of the disease, as they provide valuable information on the biological behavior of cancer, its progression, and its response to treatment [2].

Over 25 different tumor markers have been approved so far and are routinely used in clinical settings for both diagnosis and treatment monitoring [2, 3]. While some markers are cancer type-specific, others are linked to two or more cancer types. Even though any biological molecule has the potential to act as a tumor marker, most markers are either glycoproteins or proteins [2].

Initially, tumor markers were developed to test for cancer in asymptomatic people, but only a few markers have proven effective for this purpose. Presently, prostate-specific antigen (PSA) is the most commonly used tumor marker, although it has very low specificity. In fact, only a limited number of markers have clinically relevant predictive values for early-stage cancer diagnosis and are only effective when testing high-risk patients. Moreover, tumor markers are not the definitive method for diagnosing cancer. Tissue biopsy and histopathological evaluation is always required for definitive diagnosis. Alpha-fetoprotein (AFP) is another example of a tumor marker that can aid in the diagnosis of cancer, specifically hepatocellular carcinoma (HCC). However, AFP levels can also be increased in some liver diseases, although a certain threshold usually indicates HCC. Another example of a biomarker that has been extensively studied is the human epidermal growth factor receptor 2 (HER2). HER2 is overexpressed in some types of breast cancer and its expression indicates worse prognosis. However, HER2-targeted therapies have been developed, resulting in improved outcomes for patients with HER2-positive breast cancer [2].

In in-silico biomarker search studies that rely on data from high-throughput experiments, preselecting potential biomarkers can be accomplished using a variety of methods. Traditional

statistical techniques such as ANOVA or t-tests may be used, as well as more advanced techniques like uniform manifold approximation and projection (UMAP) and machine learning algorithms. However, due to the limitations of these methods, additional filtering methods are often necessary to identify biomarkers that will meet clinical requirements. These filters may include considerations such as the biological relevance of the biomarker, its stability and reproducibility across different sample types, and its ability to provide accurate predictions of clinical outcomes to ensure that the most relevant and reliable biomarkers are identified. Despite a decade of intense effort and substantial investments of resources and labor, the number of biomarkers that have been clinically validated and approved by the regulatory agencies (e.g. Food and Drug Administration, FDA) is disappointingly small [4].

With the increasing availability of transcriptomic data, particularly from small sample size experiments, it has become increasingly important to develop robust and reliable methods for identifying biomarker candidates for further validation. The aim of my research was to develop a new pipeline specifically suited for selecting potential biomarker candidates based on data acquired from a small sample size RNA sequencing experiment. Statistical properties implemented in the pipeline developed in my research, were aimed at selection of smaller numbers of candidate biomarkers, however, possessing better characteristics and suitable for a biomarker that can be applied in clinical settings. Moreover, removing numerous uncertain candidates, by applying this pipeline, may lead to cost- and effort-reduction at a validation phase which is further step required in biomarker discovery. Additionally, through pathway analysis there was undertaken attempt of in-silico validation of selected biomarker candidates, to ensure that they are biologically relevant and clinically useful. The ultimate goal of this research is to contribute to the ongoing efforts to improve cancer diagnosis and treatment by identifying more accurate and reliable biomarkers.

II. Fibroblast Growth Factor Receptors (FGFRs)

FGFRs are transmembrane proteins that belong to the subfamily of tyrosine kinase receptors (RTK) which consist of five members (FGFR1-5) that have amino acid sequence homology [5]. These receptors possess three distinct regions: an extracellular domain, a hydrophobic transmembrane domain, and an intracellular tyrosine kinase domain [6, 7]. Unlike the other members, FGFR5, also known as FGFRL1, doesn't has a tyrosine kinase domain, but it still participates in controlling the over-activation of the FGF/FGFR1 signaling pathway [8, 9]. FGFRs, by affecting signaling pathways like STAT, PI3K/AKT, and RAS/RAF/MAPK play a crucial role in the regulation of migration, invasion, proliferation, and cell survival [10]. The FGF/FGFR signaling pathway is crucial to various processes such as embryogenesis, angiogenesis, wound healing, and maintaining tissue homeostasis (Figure 1). It has a significant impact on differentiation and apoptosis as well [10].

The activation of FGFRs mainly occurs through the binding of fibroblast growth factors (FGFs), leading to the dimerization and intracellular kinase transautophosphorylation of the receptors [11]. This process triggers intracellular signaling pathways [12]. In addition, FGFRs can be activated through chromosomal translocation, which results in gene fusion with other constantly expressed genes, leading to receptor activation without ligand binding [13]. In the gastric tissue, FGFR2 signaling triggered by FGF10 regulates the maintenance of stomach progenitor cells, morphogenesis, and cellular differentiation during the early stages of epithelial growth, before differentiation occurs [14, 15]. In lung development, FGFR signaling is essential as it increases lung epithelial cell growth and regulates mesenchymal cell proliferation and airway bud formation, as well as their branching [16, 17]. FGFR2 plays a crucial role in shaping bladder mesenchyme by influencing the sonic hedgehog (Shh) signaling pathway [18].

II.1. FGFR aberrations in human disease

Dysregulation of FGFR signaling has been implicated in the development of several human diseases, including cancer, skeletal disorders, and developmental disorders (Figure 1). Abnormal FGFR signaling has been associated with various congenital disorders, such as craniosynostosis and Crouzon syndrome. Mutations in FGFR genes can cause various types of skeletal dysplasia, such as achondroplasia and thanatophoric dysplasia, leading to skeletal disorders [19]. FGFR mutations have also been linked to the development of atopic dermatitis and other skin diseases [20].

-1- Lung
+ Developmnet
+ Disease (COPD, RDS, pulmunary fibrosis, cancer)

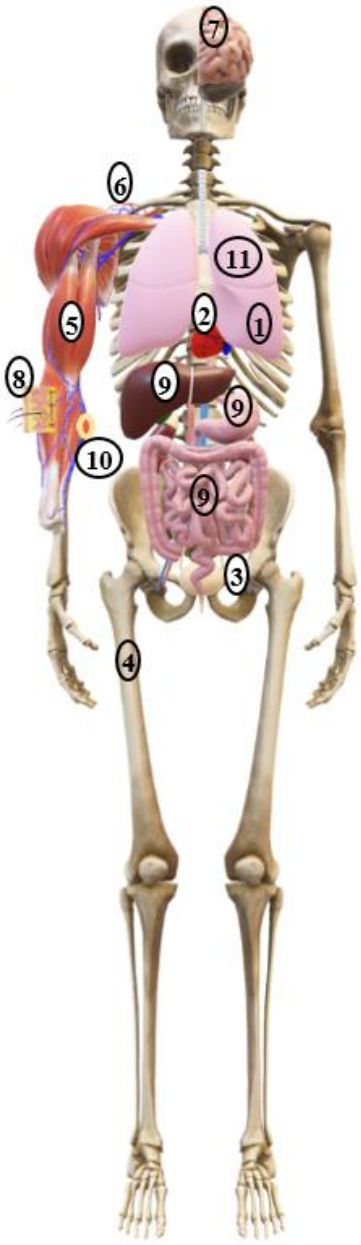
-2- Heart
+ Inducing cardiomyocyte cycle reentry
+ Cardiac development, diseases and repair

-3- Urinary system
+ Nephron development
+ Ureteric branching and induction
+ Diseases (Structural kidnay and lower urinary tract, CKD)
+ Kidney injury and repair

-4- Skeleton
+ Development and homeostasis
+ Repair/regeneration
+ Genetic diseases (ACH, PS)

-5- Skeletal muscle
+ Self-renewal of satellite cells
+ Maintenance and regeneration

-6- Blood/Lymph vessel
+ Angiogenesis
+ Lymphatic development
+ Lymphagiogenesis
+ Diseases related to EndMT, ischemic diseases, lymphedema
+ Regeneration



-7- Brain
+ Cerebral cortex development
+ Hindbrain development
+ Synaptic transmission
+ Neutral plasticity

-8- Skin/appendage
+ Skin Surface expansion
+ Skin protection from UV light damage
+ Advancement of hair follicies and hair growth
+ Sebaceous gland proliferation
+ Outgrowth of the glands (Salivary, lachrymal, mammary gland)

-3- -7- -9- Metabolism
+ Bile acids metabolism
+ Fatty acid metabolism
+ Glucose metabolism
+ Mineral metabolism (phosphate, sodium and calcium)

-10- Inflammation
+ Accentuating or suppressing
+ Inflammation in different diseases such as COPD, CKD

-11- Tumors
+ Tumor growth
+ Invasion and migration tumors
+ Tumor angiogenesis

Figure 1. Summary of the role of FGF/FGFR signaling in various aspects of health and disease. This signaling pathway is involved in the development of multiple organs including the lung, heart, urinary system, brain, skeleton, muscle, and skin. Additionally, it plays a role in tissue repair, regeneration, and inflammation. Endocrine FGFs regulate metabolism in various organs like the kidney, liver, brain, intestine, and adipose tissue. However, malfunctions of FGF/FGFR signaling can lead to diseases such as genetic conditions, cancer, chronic obstructive pulmonary disease (COPD), and chronic kidney disease (CKD) [19].

In a molecular profiling study performed by Next-Generation Sequencing on a large scale by Helsten et al. [21], it was found that 7% of cancers have aberrations of FGFRs. Genomic changes such as gain-of-function mutations, gene amplification, gene fusions, and chromosomal translocation can led to constitutive activation or increased signaling of FGF receptors [21, 22].

The most frequent FGFR aberration in lung, stomach, and bladder cancer is gene amplification [21, 23].

FGFR2 amplification is associated with increased tumor cell proliferation in 1.2-9% of gastric cancer patients [24]. All members of the FGFR family have oncogenic gene alterations involved in some human cancers (Figure 2). For instance, FGFR1 amplification is found in the bladder, gastric, breast, and lung cancers [25-28], while liver, uterine, lung, and gastric cancers may exhibit FGFR2 amplification, mutations, and fusions [23, 29, 30]. Bladder and lung cancers frequently display FGFR3 mutations and fusions [23, 31, 32], whereas FGFR4 mutations are rare in cancers [33]. FGFR4 oncogenic mutations, N535K/D, and V550E/L, have been found in rhabdomyosarcoma [16], while the G388C mutation may contribute to tumor progression, despite being a polymorphism [34].

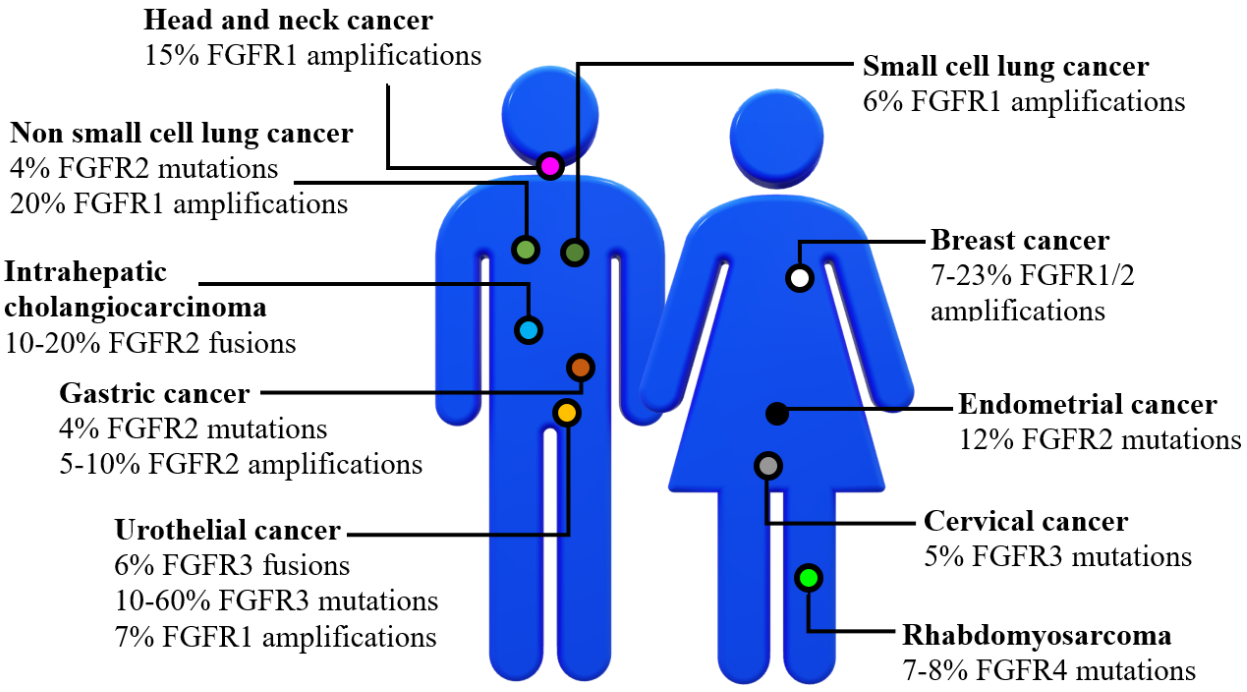


Figure 2. Types of cancer with FGFR genomic changes [35].

III. Lung, stomach, and bladder cancers characteristics

Lung, stomach, and bladder cancer display FGFR aberrations most frequently as compared to other cancers. In addition, lung cancer is a leading cancer type according to overall cancer incidence and mortality. Thus, elaboration of effective FGFR inhibitor would potentially help to save significant number of human lives.

III.1. Lung cancer

Lung cancer remains the leading type of cancer in both incidence and mortality worldwide, with 2,206,771 new cases and 1,796,144 deaths reported in 2020 (Figure 3). Approximately 18.4% of all cancer deaths are attributed to lung cancer [36].

Most new cases of lung cancer are diagnosed at an advanced stage due to the lack of symptoms in the early disease. Patients with advanced lung cancer have a poor prognosis, with a 5-year relative survival rate estimated at 5.2% [36].

Lung cancer is categorized based on its histological type. Typically, lung carcinomas are classified by their size and appearance. Non-small cell lung carcinoma (NSCLC), which makes up around 84% of all lung cancers, is further divided into adenocarcinoma (ADC, around 40-50% of cases), squamous cell carcinoma (SCC, approximately 20-30% of cases), and large cell carcinoma (LCC, 10% of cases) [37].

Various risk factors have been linked to and used as indicators of the likelihood of developing lung cancer. The most significant risk factors are smoking and age. Other factors that may contribute to the risk of lung cancer include sex, race/ethnicity, family history of lung cancer, COPD, emphysema, and exposure to asbestos and radon [38].

III.2. Stomach cancer

In 2020, stomach cancer was the sixth most common type of cancer worldwide (Figure 3), with approximately 1.1 million new cases, and the fourth leading cause of cancer-related deaths, with roughly 800,000 fatalities [36, 39]. Men have a twofold higher incidence rate than women. In some South Central Asian countries such as Iran, Afghanistan, Turkmenistan, and Kyrgyzstan, it is the most commonly diagnosed cancer and the leading cause of cancer death in men. The

highest incidence rates are observed in Eastern Asia and Eastern Europe, particularly China, had the highest number of stomach cancer cases, with nearly 820,000 new cases and 580,000 deaths, whereas Northern America and Northern Europe tend to have lower rates [36, 39-41].

Stomach cancer has a poor prognosis with a five-year survival rate estimated at less than 20% [39, 42-44] due to its asymptomatic early stage and the majority of cases being diagnosed at an advanced stage [45, 46].

Stomach cancer is rare in persons under 45 years of age, with the incidence and death toll from the disease rising with increasing age [39, 40]. The occurrence of stomach cancer is about twice as common in men as in women [36, 39, 41].

During the first half of the 20th century, stomach cancer was the top cause of death from cancer in the United States and Europe [47, 48]. However, in recent decades, the incidence and death rate from stomach cancer have significantly decreased in several countries [36, 39].

Stomach cancer is a complex disease with multiple factors contributing to its development, including lifestyle and environmental risks like obesity, low socioeconomic status, family history, smoking, inherited predisposition, low physical activity, Helicobacter pylori infection, radiation exposure, gastroesophageal reflux disease, poor diet, and alcohol use [46].

III.3. Bladder cancer

Bladder cancer is the twelfth most common malignancy globally, with a reported 573,278 new cases in 2020 (Figure 3) [36, 41]. It affects men more frequently than women, with a ratio of 3 to 4, which is thought to be due to lifestyle and exposure differences, although a higher risk has also been linked to stasis of urine-contained carcinogens in men with prostate enlargement and urinary retention [49, 50].

The most significant risk factor for bladder cancer is advanced age, with the average age of diagnosis ranging from 70 to 84 years [51]. The increased risk of bladder cancer is attributed to a combination of factors including exposure to carcinogens such as tobacco smoke, benzene chemicals, and aromatic amines, along with a reduction in the DNA repair ability as a result of aging [49].

Bladder cancer can present at a stage of non-muscle-invasive (NMIBC) disease, muscle-invasive (MIBC), or metastatic disease [52], with 75% of patients diagnosed with NMIBC and 50% of these being categorized as low-grade [53].

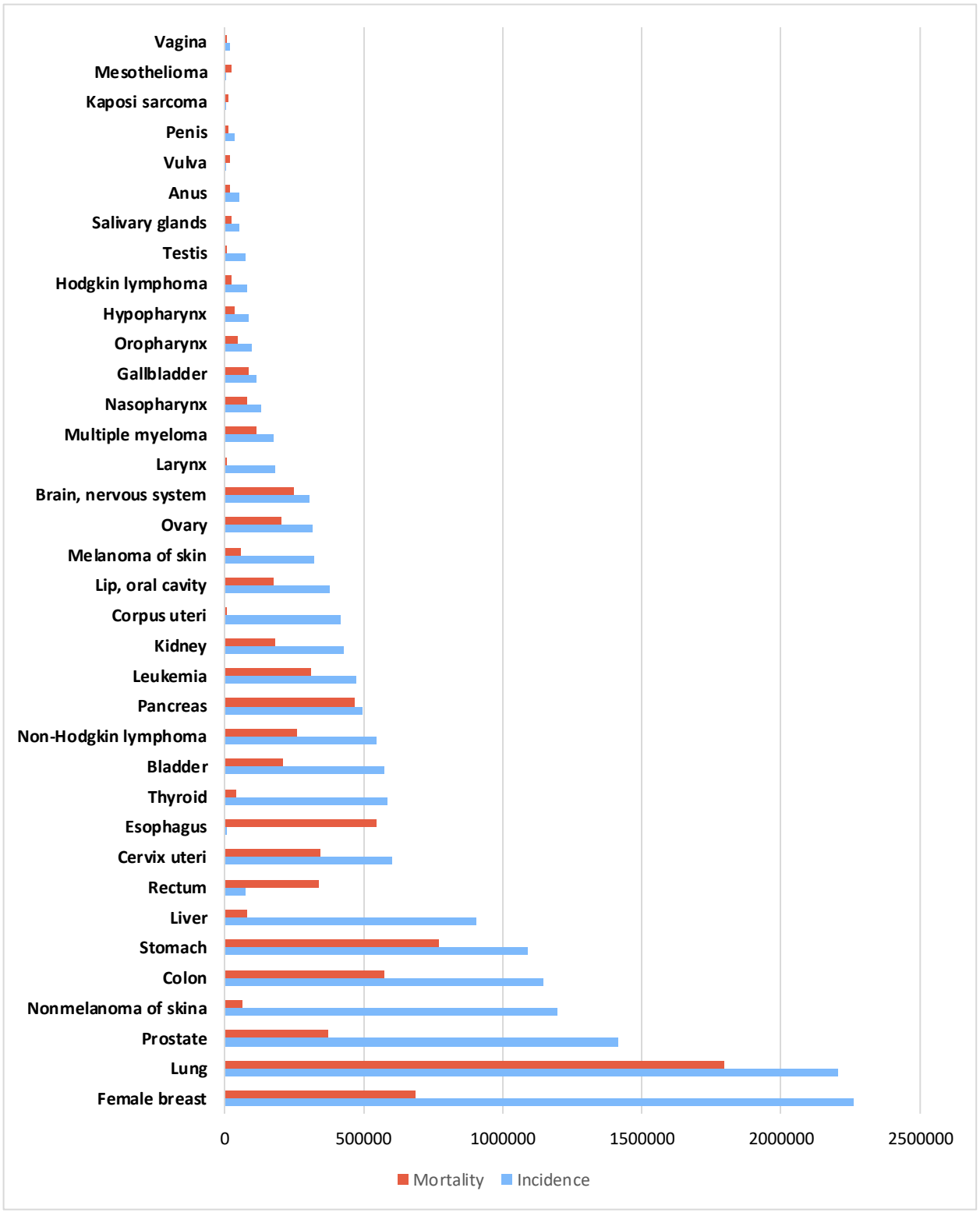


Figure 3. New cases and deaths from 36 types of cancers in 2020. Based on data published by GLOBOCAN [36].

IV. Tyrosine kinase inhibitors (TKIs)

FGFR belongs to the large family of tyrosine kinase inhibitors. In 2001, the United States Food and Drug Administration (FDA) approved the first tyrosine kinase inhibitor to be used in the clinics [54]. Imatinib targets the Abelson (ABL) tyrosine kinase expressed as a deregulated fusion protein (BCR–ABL) in chronic myeloid leukemia [55]. With the discovery of gene alterations and rearrangements in BRAF, NTRK, ROS1, ALK, and EGFR, the development of TKIs has rapidly progressed [56]. Today there are more than 20 TKIs approved by the FDA [57-63]. The information on these drugs is summarized in Table 1. These drugs have been used mainly in combination (or in the sequence) with traditional chemotherapy and radiotherapy in the treatment of advanced cancers, including lung adenocarcinoma (ADC), stomach cancer, and bladder cancer, and have significantly improved patient outcomes [37, 64].

Table 1. TKIs approved by FDA [65].

TKIs	Company	Time of release	Application of disease	Target
Imatinib	Novartis	2001	CML, GIST	Abl, PDGFR, SCFR
Gefitinib	AstraZeneca	2003	NSCLC	EGFR
Nilotinib	Novartis	2004	CML	Bcr-Abl, PDGFR
Sorafenib	Bayer	2005	Advanced RCC	Raf, VEGFR, PDGER
Sunitinib	Pfizer	2006	GIST, Advanced RCC	PDGFR, VEGFR,
Dasatinib	Bristol-Myers Squibb	2006	CML	Bcr-Abl, SRC, PDGFR
Lapatinib	GlaxoSmithKline	2007	Breast cancer	EGFR
Pazopanib	GlaxoSmithKline	2009	Advanced RCC,STS,NSCLC	VEGFR, PDGFR
Crizotinib	Pfizer	2011	NSCLC	ALK
Ruxolitinib	Novartis	2011	myelofibrosis	JAK1, JAK2
vandetanib	AstraZeneca	2011	Advanced Thyroid cancer	VEGFR, EGFR
Axitinib	Pfizer	2012	Advanced RCC	VEGFR
Bosutinib	Wyeth	2012	CML	Abl, SRC
Afatinib	Boehringer Ingelheim	2013	NSCLC	EGFR
Erlotinib	Roche	2013	NSCLC	EGFR
Ceritinib	Novartis	2014	NSCLC	ALK

Osimertinib	AstraZeneca	2015	NSCLC	EGFR
Lenvatinib	Eisai	2015	DTC	VEGFR
Alectinib	Roche	2015	NSCLC	ALK
Regorafenib	Bayer	2017	HCC, CRC,GIST	VEGFR, EGFR
Neratinib	Puma	2017	Breast cancer	HER2
Brigatinib	Ariad	2017	NSCLC	ALK

IV.1. Clinical development of FGFR-TKIs

Since the deregulation of FGFR signaling has been linked to the development and progression of cancer, FGFR has become exploited as potential therapeutic target [66]. Erdafitinib was first approved for metastatic urothelial carcinoma based on positive results from a phase II trial (Table 2, NCT02355597) [67, 68]. Currently, a phase III trial (NCT03390504) is underway to compare the efficacy of Erdafitinib versus Vinflunine or Docetaxel or Pembrolizumab in advanced urothelial cancer. Meanwhile, Pemigatinib was authorized as the first targeted therapy for advanced cholangiocarcinoma in 2020 [69, 70]. Both these drugs are undergoing multiple clinical trials for various indications, including non-small cell lung cancer, advanced solid tumors, breast cancer, liver cancer, castrated prostate cancer, and more. Many other FGFR tyrosine kinase inhibitors (FGFR-TKIs) candidates are in the early stages of clinical trials [71, 72]. More data about FGFR-TKIs that are already approved for clinical use, and others that are investigated in clinical trials for their effectiveness in treating tumors related to FGFR aberrations, is shown in Table 2.

Table 2. FGFR-TKIs FDA approved and under development [71-73].

Drug	**Company**	**Target**	**Approved/ clinical trials**
Erdafitinib (JNJ-42756493)	Janssen	Pan-FGFR	FDA approved Phase I/IIa NCT02421185 NCT03473743 Phase II NCT02365597 NCT03210714 NCT04083976

			NCT02699606 NCT03827850 NCT02952573 NCT03999515 NCT04172675 Phase III NCT03390504
Pemigatinib (INCB054828)	Incyte	Pan-FGFR	FDA approved Phase II NCT02924376 NCT02872714 NCT04003610 NCT03914794 NCT03822117 NCT03011372 NCT04256980 NCT04003623 Phase III NCT03656536
Futibatinib (TAS-120)	Taiho Pharm	Pan-FGFR	FDA approved Phase II NCT02052778 NCT04024436
CH5183284 (Debio-1347)	Debio	FGFR1/2/3	Phase II NCT03834220
ASP5878	Astellas	Pan-FGFR	Phase I NCT02038673
Dovitinib (TKI258)	Novartis	FGFR1/2/3; KIT; VEGFR	Phase II NCT01861197 NCT01379534
PRN1371	Principia	Pan-FGFR	Phase I NCT02608125
LY2874455	Eli-Lilly	Pan-FGFR; VEGFR2	Phase I NCT01212107

Infigratinib (BGJ398)	Novartis	Pan-FGFR	Phase II NCT02150967 NCT02160041 NCT04233567 Phase III NCT03773302 NCT04197986
AZD4547	AstraZeneca	Pan-FGFR	Phase II NCT02465060
Derazantinib (ARQ-087)	Basilea	Pan-FGFR; RET; DDR2; KIT;VEGFR; PDGFRβ	Phase I/II NCT01752920 NCT04045613 Phase II NCT03230318
E7090	Eisai	FGFR1/2/3	Phase II NCT04238715
HMPL-453	Chi-Med	FGFR1/2/3	Phase II NCT04353375
Rogaratinib (BAY-1163877)	Bayer	Pan-FGFR	Phase III NCT03410693
Roblitinib (FGF401)	Novartis	FGFR4	Phase I/II NCT02325739
ODM-203	Orion	FGFR; VEGFR1/2/3	Phase I/IIa NCT02264418
ICP-192	InnoCare	Pan-FGFR	Phase II NCT04492293
H3B-6527	Eisai /H3	FGFR4	Phase I NCT02834780
Fisogatinib (BLU-554)	Blueprint	FGFR4	Phase I NCT02508467

IV.2. FGFR-TKIs resistance mechanisms

The use of small-molecule inhibitors of FGFR activity as an anti-cancer strategy holds great promise. However, the development of resistance to these drugs is becoming a significant challenge. Several mechanisms of acquired resistance have been documented in the literature (Figure 4) [37].

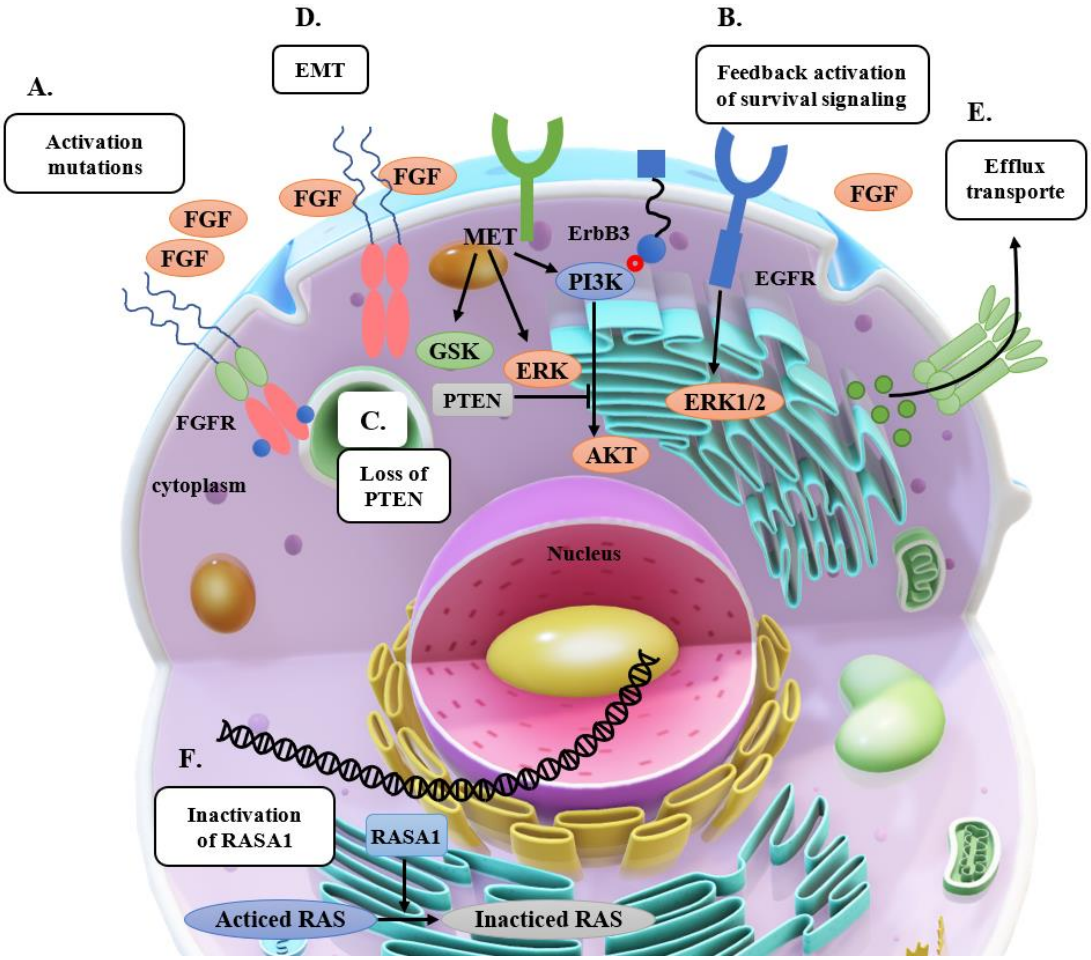


Figure 4. Mechanisms of resistance to FGFR inhibitors: (A) gatekeeper mutations in the FGFR kinase domain, (B) activation of alternate signaling pathways like EGFR, ERBB3, or MET, (C) loss of PTEN leading to increased activation of PI3K-AKT, (D) the epithelial-mesenchymal transition (EMT) may lead to resistance to FGFR inhibitors, (E) drug efflux regulation by ABCG2, and (F) the inactivation of RAS by RASA1. Resistance to FGFR inhibitors can arise when RASA1 is inactivated [74].

The first significant mechanism of resistance to FGFR inhibitors is a mutation at the gatekeeper residue of the protein. This residue is situated in the active site and regulates access to the FGFR hydrophobic pocket localized behind the ATP-binding pocket. Different mutations, such as FGFR1V561M and FGFR1N546K, have been linked to resistance to FGFR inhibitors [75, 76]. The FGFR1N546K mutation leads to increased ATP affinity and thus resistance to the drug, while FGFR1V561M reduces drug affinity [76]. The FGFR3V555M mutation is also linked to resistance to PD173074 and AZD4547 inhibitors [77]. FGFR2V564F was found to cause resistance to

BGJ398 in patients with cholangiocarcinoma as it creates a hindrance in the binding pocket of FGFR2 and BGJ398 [78]. RNA sequencing analysis and in vivo study on a mouse model revealed 7 missense mutations in the FGFR2 kinase domain in breast cancer (I567S, N568H/T, V581L, E584G, S587L, K660R, and K678M). It was confirmed that the decrease in sensitivity to AZD4547 in breast cancer was caused by the presence of these mutations in FGFR2 [79].

Another significant factor contributing to resistance to FGFR inhibitors is the feedback activation of the survival loop due to FGFR inhibition (Table 3). This was demonstrated through a high-throughput proteomic study of DMS114 (SCLC cell line) and RT112 (urothelial carcinoma cell line) cells exposed to BGJ398, which showed increased activation of AKT and its target, GSK3 [80]. In urothelial carcinoma, it was revealed that the PI3K pathway can play a role in formation of resistance to FGFR inhibitors. This pathway can be activated by EGFR or ERBB3 when FGFR is inhibited [81]. The role of EGFR in resistance to FGFR inhibition was confirmed in FGFR3 mutant cancer cells: the downregulation of MAPK signaling, may lead to the sustained activation of EGFR through a reduced ubiquitination [82]. In endometrial cancer cell lines PTEN loss has been implicated as a mechanism of resistance to FGFR inhibition [83]. In breast cancer, activation of the drug efflux protein ABCG2, inactivation of Ras p21 protein activator 1 (RASA1), and overexpression of MET have been linked to resistance to AZD4547 [79]. Additionally, epithelial-mesenchymal transition (EMT) has been implicated in the resistance of gastric cancer cell lines to FGFR inhibitors [84].

Table 3. Signaling pathways involved in FGFR-TKIs resistance [75-84].

Cancer type	Signaling pathway involved in resistance	FGFR-TKIs
Bladder cancer	EGFR signaling pathway	PD173074
Breast cancer	MET, inactivation of RASA1, drug-efflux	AZD4547
Endometrial cancer	Loss of PTEN	Ponatinib
Gastric cancer	EMT	AZD4547
Gastric cancer	EMT	BGJ398
Gastric cancer	EMT	PD173074
Lung cancer	PI3K/AKT and GSK signaling pathway	BGJ398
Urothelial cancer	PI3K/AKT and GSK signaling pathway	BGJ398
Urothelial cancer	EGFR/ERBB3-AKT signaling pathway	AZD4547

V. Biomarker discovery

The earliest recorded effort to identify markers for malignancy dates back 2000 years and is documented in an Egyptian papyrus, which described a distinction between breast cancer and mastitis [85]. In 1846, Henry Bence-Jones made the discovery of abnormal protein precipitate in urine of some patients. This was later identified as an immunoglobin light chain which may be overexpressed in multiple myeloma and Waldenström macroglobulinemia. This tumor marker is now called Bence-Jones protein and used in diagnostics and monitoring the disease [85]. In 1965, Gold and colleagues isolated a glycoprotein molecule from human colon cancer specimens and uncovered the first "tumor antigen," later known as carcinoembryonic antigen (CEA) [86].

Today there is a great effort and substantial investments of resources and labor devoted to discovery of new biomarkers, even though the number of biomarkers that have been clinically validated and approved by the FDA is rather small comparing to these efforts. Some markers are specific to a single type of cancer, while others are linked to multiple cancer types. However, majority of them have low sensitivity and specificity. Most markers are proteins or glycoproteins, but low molecular weight substances, such as vanillylmandelic acid and homovanillic acid, are used for diagnosing neuroblastoma. Additionally, DNA and RNA nucleic acids are being explored as potential tumor markers [87].

V.1. Definition of a biomarker

In 2015 the FDA and the National Institutes of Health (NIH) at a joint leadership conference created a group, with the purpose of establishing common definitions and making them easily accessible through a regularly updated online document referred to as the "Biomarkers, Endpoints, and other Tools" (BEST) resource [88].

According to BEST, the basic definition of a biomarker is: “A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention.” [88]. The definition of biomarkers encompasses a wide range of characteristics, including radiographic, histologic, physiologic, and molecular attributes that can be used to assess the effectiveness of therapeutic interventions. These biomarkers should not be confused with clinical outcome assessments (COAs), which measure how a person feels, functions, or survives and are directly important to patients. COAs can be used to establish regulatory approval standards for therapeutics, while biomarkers serve multiple purposes, one of which is to

predict COAs based on measurements [89]. Within this work I will discuss mostly cancer biomarkers, applicable in cancer diagnosis, prognostication and prediction of the treatment response.

V.2. Characteristics of an ideal biomarker

In clinical application there are five main uses of tumor markers [86, 89, 92, 93]:

- Screening for cancer in a healthy or high-risk population
- Diagnosing cancer or its specific type
- Assessing a patient's prognosis
- Predicting a patient's response to a given therapy
- Monitoring tumor response to the therapy and/or recurrence.

The ideal diagnostic marker should possess three key features: (a) specificity to a particular disease, (b) early detection before clinical diagnosis, and (c) high sensitivity to minimize false positives. Furthermore, the marker's levels should closely match the extent of the tumor, showing any changes in its progression or regression, with a short half-life that allows for frequent monitoring. The screening test should be easy, cheap and reproducible (Table 4) [90].

Table 4. Features of an ideal tumor biomarker [90].

Features	Description
Highly specific	Only detectable in a single type of tumor.
Highly sensitive	Not present in normal or non-malignant health conditions.
Long lead-time	Adequate time window to change the natural course of the disease.
Levels correlate with tumor burden	The ability of the tumor marker to predict the course of the disease and its future outcomes.
Express early	Should express early in the disease progression.
Short half-life	Relatively short half-life, reflecting temporal changes in tumor burden and response to therapy.
Simple and cheap test	The ability to use the marker as a screening test; easy to assay, less expensive.
Easily obtainable specimens	Acceptability by target population.

Features	Description
Reproducible	Give reproducible results and multiplexing is possible for screening purposes.

Having a test that is highly sensitive and highly specific with all the ideal features is desirable, but often unattainable. Typically there is a trade-off. Clinical tests typically recognize "normal" and "abnormal" specimens, as well as those that fall into a gray area. Therefore, establishing the criteria for positive and negative results requires making choices [91].

For example, in situations where treatment is difficult, expensive, and harmful to the patient, but the condition itself has low transmissibility a diagnostic biomarker may need high specificity (low false positives), whereas highly infectious diseases may require diagnostic biomarkers with high sensitivity to prevent further spread of the disease and avoid false negatives during diagnosis [92]. In the case of monitoring biomarkers, it may be necessary for them to have higher sensitivity at the expense of specificity. This is because when used in patients with a diagnosed medical condition, there is no need to detect the disease, but rather to monitor its course [91, 92].

Greater sensitivity leads to a lower likelihood of having the disease if the test result is negative, resulting in a higher negative predictive value, whereas greater specificity leads to a lower probability of being disease-free if the test result is positive, resulting in a higher positive predictive value [91].

V.3. Clinically used biomarkers

According to Wishart et al. [93] an author of the MarkerDB online database of molecular biomarkers there are 26,493 clinically approved biomarkers (by 2020), despite that only several dozen are used in a daily clinical practice [87]. Currently, the most commonly utilized tumor marker is Prostate-Specific Antigen (PSA). PSA is a protein that is overexpressed in prostate cancer patients. The biomarker is used for the early detection of prostate cancer, although it has low specificity, and to monitor disease progression (in this application it performs much better). Another example of a tumor marker used for diagnosing cancer, particularly hepatocellular carcinoma (HCC), is Alpha-fetoprotein (AFP). Although, elevated levels of AFP can also indicate some liver diseases, generally a threshold level is often considered a sign of HCC [87]. Some tumor markers commonly used in the clinical setting are presented in Table 5.

Table 5. Examples of commonly used tumor markers in the clinic [87].

Marker	Application
Prostatic-specific antigen (PSA)	Prostate carcinoma
Cancer antigen 125 (CA 125)	Ovarian and fallopian carcinoma
Cancer antigen 15-3 (CA 15-3)	Breast cancer
Cancer antigen 19-9 (CA 19-9)	Pancreatic and ovarian cancer
CA-72-4	Colorectal cancer
Alpha-fetoprotein	Hepatoblastoma, hepatocellular carcinoma, and germ cell tumors
Carcinoembryonic antigen (CEA)	Colorectal, gastric, pancreatic, lung, and breast carcinomas
Beta-2-microglobulin (B2M)	Multiple myeloma and lymphoma
Human chorionic gonadotropin (hCG)	Choriocarcinoma and testicular carcinoma
Thyroglobulin	Thyroid cancer

Other tumor markers include e.g. CYFRA 21-1 (a marker for lung cancer), HE4 (a marker for ovarian cancer), Squamous cell carcinoma antigen (a marker for squamous cell lung cancer), Neuron-specific enolase (a marker for lung cancer), Chromogranin A (a marker for neuroendocrine tumor), and Thymidine kinase (a marker for multiple myeloma and chronic lymphocytic leukemia) [87].

V.4. Predictive biomarker

A biomarker that can predict the likelihood of experiencing a favorable or unfavorable effect from exposure to a medical product or environmental agent is called a predictive biomarker [88]. To prove the usefulness of a biomarker for this purpose, a rigorous approach to clinical studies is required. In an ideal study, patients with or without the biomarker, are randomly assigned to one of two or more treatments (or a placebo), and differences in outcome should be significantly related to the presence or absence (or the level) of the biomarker. Randomization to treatment

versus control groups is important because simply showing that positive biomarker patients receiving an investigational therapy fare better than negative biomarker patients, does not establish the biomarker's predictive value [88, 89].

Clinical studies aiming to assess the predictive ability of a biomarker should typically enroll patients with a range of biomarker values or include those who are either positive or negative for binary biomarkers. However, in certain cases where there is compelling evidence that an investigational therapy will not be effective or could even be harmful in a certain biomarker-defined subgroups, the exclusion of biomarker-negative patients from the trial may be necessary. On the other hand, when a biomarker identifies a subset of patients who are most likely to benefit from the therapy, enriching the trial with those patients can increase the statistical power and help detect a larger effect of the therapy. Additionally, the use of an enrichment strategy can affect the intended population to receive the therapy after regulatory approval [88].

In the design and implementation of clinical trials, predictive biomarkers are crucial for enrichment strategies. By enrolling participants with high levels of a predictive biomarker, the treatment's actual effect can be more clearly demonstrated, especially during the pre-registration stage of drug development. Using predictive biomarkers for enrichment is a more focused approach than using prognostic biomarkers, which increase event rates but cannot select particular patients who are more likely to respond (or not) to therapy [89].

Much of the current consensus about treatment choice in clinical practice relies on the same principle. Elevated blood pressure is treated with antihypertensive medications, low Hb levels are treated with blood transfusions, and acute reperfusion is indicated for patients with ST-segment elevation on an electrocardiogram - all of which are examples of predictive biomarkers used to select patients who are likely to respond to therapy. In population health strategies, populations with high levels of predictive biomarkers are identified as needing additional intervention. For instance, patients with high HbA1C (hemoglobin A1C) levels benefit the most from aggressive diabetes treatment. The development of genetic and genomic markers for precision medicine is also a major growth area for predictive biomarkers, such as HER2 receptor-positive assays in cancer patients who are more likely to respond to treatment with Herceptin [89]. Some examples of prognostic biomarkers (currently used and potential) are shown in Table 6.

Table 6. Examples of prognostic tumor markers.

Marker	Application
Thiopurine methyltransferase (TPMT) genotype or activity	When evaluating patients who may be treated with 6-mercaptopurine or azathioprine, Thiopurine methyltransferase (TPMT) genotype or activity can be used as a predictive biomarker to identify individuals at risk for severe toxicity due to high drug concentrations [88].
BRCA1/2 gene mutations	BRCA1/2 gene mutations may serve as predictive biomarkers for sensitivity to ionizing radiation, as they can hinder the ability of the genes' protein products to repair double stranded DNA breaks, a form of damage induced by ionizing radiation [88].
Human leukocyte antigen allele (HLA)–B*5701 genotype	Abacavir treatment in HIV patients can be assessed with HLA-B*5701 genotype as a predictive biomarker to identify individuals with a high risk for severe skin reactions [88].
HER2 gene amplification	The analysis of the HER2 gene amplification is the test used in cancer diagnostics for the evaluation of the eligibility of breast cancer patients for treatment with trastuzumab or lapatinib [94].
Somatic mutations in codon 600 of the BRAF gene	Assessment of somatic mutations in codon 600 of the BRAF gene in patients with advanced melanoma in order to administer treatment based on dacarbazine (DTIC), and vemurafenib [94].
Assessment of the fusion gene EML4-ALK	Assessment of the EML4-ALK fusion gene is critical in predicting a colorectal cancer patients's eligibility for treatment with crizotinib [94].
Squamous differentiation in non-small cell lung cancer	Scagliotti et al. have suggested that the presence of squamous differentiation in non-small cell lung cancer can serve as a predictive biomarker, indicating that patients who receive

	pemetrexed are likely to experience worse survival or progression-free survival outcomes than those who receive other standard chemotherapies such as docetaxel or cisplatin in combination with gemcitabine [88].
Certain cystic fibrosis transmembrane conductance regulator (CFTR) mutations	In clinical trials assessing cystic fibrosis treatment, specific mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) gene can serve as predictive biomarkers to identify patients who are more likely to benefit from particular treatments [88].
BReast CAncer genes 1 and 2 (BRCA1/2) mutations	The presence of BRCA1/2 mutations can serve as predictive biomarkers for identifying women with platinum-sensitive ovarian cancer who are likely to benefit from treatment with Poly (ADP-ribose) polymerase (PARP) inhibitors [88].
Mutation status in codons 12 and 13 of the KRAS gene	To determine eligibility of advanced colorectal cancer patients for targeted therapy using monoclonal antibodies such as cetuximab or panitumumab, the mutation status in codons 12 and 13 of the KRAS gene is a common predictive biomarker.

V.5. Biomarker performance indices

The performance of a biomarker is evaluated through its sensitivity and specificity. Sensitivity refers to the ability to correctly detect disease in patients who have given condition (true positive), while specificity refers to the ability to correctly identify patients without given condition (true negative) [95]. In terms of predictive biomarker sensitivity refers to the likelihood of a positive biomarker test result for patients who will benefit from treatment compared to control, while specificity refers to the likelihood of a negative biomarker test result for patients who will not benefit from treatment compared to those who will [96].

KRAS is a gene biomarker that is mutated in approximately 35–45% of metastatic colorectal cancers [97]. KRAS mutations may lead to resistance to anti-EGFR antibodies, thereby negating any potential benefits of antibody therapy and preventing its effectiveness. Thus, the presence of KRAS mutations could indicate a lack of response to anti-EGFR antibodies. KRAS gene is an example of a predictive biomarker with a specificity of 0.93 (CI, 0.87 to 0.97) and the sensitivity of KRAS mutations for predicting lack of response is 0.49 (CI, 0.43 to 0.55), indicating that this biomarker is a highly accurate predictor of non-response to treatment [98].

To calculate sensitivity and specificity measures, a dichotomous prediction based on the biomarker and the patient's true disease status is used to create a 2x2 contingency table [95]. Table 7 demonstrates how the frequency of predictions from a patient sample can be used to compute sensitivity and specificity.

Table 7. Diagnostic matrix and their main parameters [95].

Disease / **Biomarker**	**Present**	**Absent**	**Total**
Positive	a (true positive)	b (false positive)	a + b
Negative	c (false negative)	d (true negative)	c + d
Total	a + c	b + d	a + b + c + d
prevalence = (a + c)/a + b + c + d)			
sensitivity = a/(a + c)			
specificity = d/(b + d)			
positive predictive value = a/(a + b)			
Negative predictive value = (d/(c + d)			
Accuracy = (a + d)/(a + b + c + d)			
Youden index = sensitivity + specificity – 1			

While sensitivity and specificity are often reported in biomarker studies, they may not always be directly applicable to clinical practice. Instead, clinicians are often more interested in the probability of disease presence or absence given a positive or negative test result, which is reflected by the positive predictive value (PPV) and negative predictive value (NPV) (Table 7) [95].

A good example of a biomarker in terms of assessment of PPV and NPV performance indices is prostate-specific antigen (PSA). The PSA level at which there are justified indications for performing a prostate biopsy is still debatable. It is commonly practiced to perform the biopsy at a concentration higher than 4 ng/ml. For this value, the positive predictive value (PPV) of detecting prostate cancer (PCa) is only 30%, but the negative predictive value (NPV) of the test is 81%, indicating that this biomarker has very low performance as a diagnostic biomarker but is a highly accurate in the manner of disease monitoring [99, 100].

The proportion of patients correctly classified by a test is referred to as predictive accuracy, which includes the sum of true positive and true negative tests. Although accuracy is occasionally reported as a global evaluation of the test, it is suggested that authors provide more than just an estimate of accuracy [95].

The Youden index, Y, is a measure of how well a test performs compared to the optimal performance. Y is defined as the sum of sensitivity and specificity minus one, and it is sometimes referred to as "regret," representing the utility loss due to uncertainty about the true state. Accuracy, on the other hand, is a weighted average of sensitivity and specificity, using disease prevalence as the weight. Since sensitivity, specificity, negative and positive predictive values, and accuracy are interrelated, knowledge of any three of these measures is sufficient to calculate the remaining two [95].

Basics of ROC Curve

The ROC (receiver operating characteristic) curve consists of a set of pairs of proportions, which represent true positive and false-positive results, or sensitivity and (1 - specificity). These pairs are obtained for various cutoff points and can be used to create an empirical ROC curve, or a smoothed curve can be generated through the fitting, typically using the binomial distribution (Figure 5) [95].

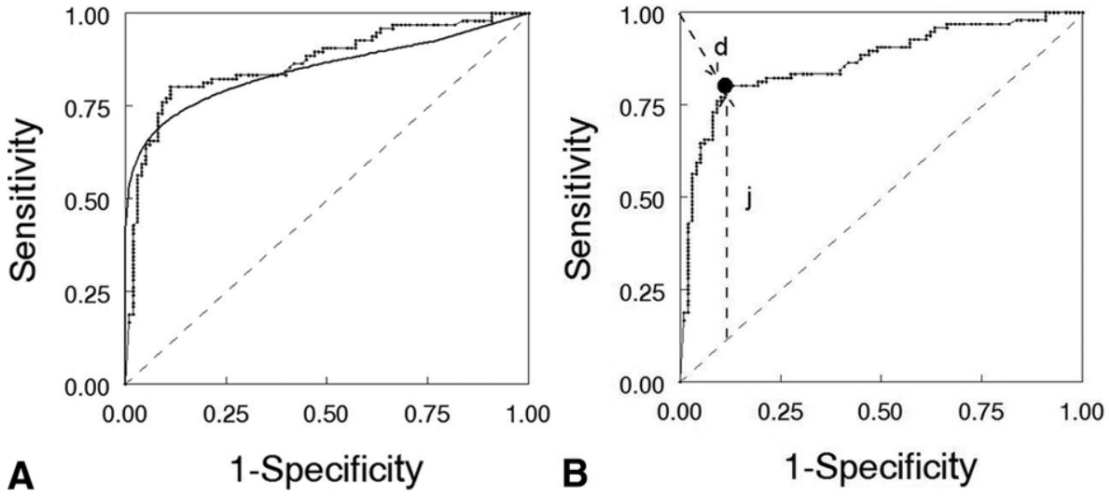


Figure 5. This graph demonstrates how the predictive value of Brain Natriuretic Peptide (BNP) for cardiogenic pulmonary edema in elderly patients (>65 yr) admitted to the emergency department for acute dyspnea is determined by the relationship between sensitivity (true positive) and 1 - specificity (true negative). (A) The ROC curve displays the empirical and continuous line, where the area under the empirical ROC curve was 0.870 (95% confidence interval 0.800–0.910). (B) The optimal threshold was selected based on the mathematical distance (d) from the point where both sensitivity and specificity are equal to 1, which corresponds to a BNP concentration of 250 pg/ml, with a sensitivity = 0.780 and a specificity = 0.900. The best cutoff should be preferably chosen based on the Youden index (sensitivity + specificity − 1), which, in this case, provided the same cutoff value. However, the optimal cutoff should consider the prevalence and cost-benefit analysis. Data adapted from Ray et al. [95, 101].

The AUC^{ROC} (area under the ROC curve), measures discrimination and is equivalent to the probability that a biomarker is higher for a diseased patient than a control. In other words in terms of predictive biomarker the area under the receiver operating characteristic (AUC^{ROC}) curve, which is a measure of how well the biomarker can distinguish between patients who will not respond to treatment and those who will. A biomarker is considered to have good discriminative properties when AUC is greater than 0.75 and excellent when it exceeds 0.90. The ROC curve, which provides a global assessment of test accuracy without any a priori hypothesis about the cutoff chosen, is relatively independent of prevalence and is a simple plot that can be easily understood visually [95].

ROC curves have three summary measures for accuracy: (1) sensitivity and specificity at a chosen cutoff point, (2) AUC^{ROC}, and (3) partial area under a portion of the curve for a prespecified range of values. However, interpreting AUC^{ROC} can be problematic due to the influence of biomarker values of no clinical relevance, and comparing two ROC curves based on the entire area may lead to different conclusions. Therefore, it is recommended to examine ROC curves in the context of the partial area or average sensitivity over a range of clinically relevant false positive rates, in addition to AUC^{ROC} [95].

Epidermal growth factor receptor (EGFR) is an example of a predictive biomarker where AUC^{ROC} is used in performance assessment. EGFR is a protein biomarker that is overexpressed or mutated in a subset of non-small cell lung cancer patients. The biomarker predicts the response to treatment with EGFR tyrosine kinase inhibitors (TKIs) such as gefitinib and erlotinib. The AUC^{ROC} for EGFR is approximately 0.7-0.8, indicating a moderately accurate predictor [102].

V.6. Validation methods for biomarkers

Sophisticated methods are commonly used to test and validate the efficacy of biomarkers. These methods can include techniques such as liquid chromatography-mass spectrometry (LC-MS), nuclear magnetic resonance (NMR) spectroscopy, and immunoassays, among others. These methods enable researchers to identify and quantify biomolecules with high specificity and sensitivity, making them valuable tools in the development and validation of biomarkers [103].

A wide range of assays can be employed in the process of validating biomarker candidate, spanning from simple methods such as immunohistochemistry (IHC) and immunoassays to high technology platforms like genomics, proteomics, and multiplex ligand-binding assays [103].

A genomics approach involves methods that measure global gene expression, such as microarrays which are commonly used for target identification and validation. Reverse transcription-polymerase chain reaction is a highly sensitive, reproducible technology that is often used to validate microarray-generated data. Comparative genomic hybridization can detect chromosomal alterations associated with certain diseases. Proteomics, on the other hand, involves global protein profiling to provide information about protein abundance, location, modification, and interactions. While proteomics is primarily a discovery technology, immunoassays are routinely used for protein biomarker assessments due to their straightforward clinical application and potential diagnostic assay translation. Multiplexing protein assays can increase throughput for the simultaneous analysis of several proteins but has limitations, such as the need to standardize assay conditions, loss of sensitivity over single assays, and quality control of each analyte in the complete multiplex panel [103].

Metabolomics is the analysis of native metabolites present in biological fluids or tissues to characterize the metabolic phenotype. This is achieved through analytical platforms such as nuclear magnetic resonance spectroscopy and the combination of liquid chromatography with mass spectrometry. While primarily utilized for biomarker discovery, it represents the ultimate end-point measurement of biological events. Nevertheless, the technology's lack of comprehensive metabolite databases and throughput limits data analysis and interpretation. The integration of these technologies in bioinformatics allows for linking expression data derived from genomics/proteomics to targeted biological pathways for a comprehensive understanding of disease biology, further validating the biomarker's application [103].

Biomarker development is significantly impacted by advancements in modern imaging techniques, which include molecular and functional imaging technologies. These approaches allow for the evaluation of cellular metabolism, cell proliferation and apoptosis, and angiogenesis and vascular dynamics using techniques such as 18F-fluorodeoxyglucose positron emission tomography, 18F-fluoro-L-thymidine and 99mTcannexin imaging, and dynamic contrast-enhanced computed tomography and magnetic resonance imaging [103].

Usually, these sophisticated methods are costly, requiring specialized equipment, advanced bioinformatics methods like the use of deep neural network (DNN) modeling, and highly trained personnel to operate and maintain them [104]. The cost of these methods can limit their accessibility, particularly in resource-limited settings, and can be a significant barrier to the widespread

adoption of biomarker testing. To address this challenge, researchers are exploring the use of alternative methods, such as point-of-care testing and wearable devices, that can provide rapid and low-cost biomarker testing [105]. As long as advanced biomarker testing methods are not widely used, many studies are being conducted to develop biomarkers for use in clinical daily practice using low-cost technologies such as immunohistochemistry (IHC) and immunoassays [92, 103].

VI. The biomarker discovery pipeline

The process of biomarker discovery can involve the use of model systems, like mouse models or cell lines, or a range of human biological samples. It typically involves a basic comparison between healthy and diseased tissues (in case of diagnostic biomarkers) to eliminate contamination by other diseases or factors. This leads to a list of potential biomarkers, referred to as "candidate biomarkers" (Figure 6), which have been found to be differently expressed between the normal and diseased states [106]. However, many of these candidates may not be differentially expressed upon further testing. To increase the accuracy of the list, it can be supplemented by information from other sources such as literature, alternative discovery methods, or expert knowledge [107].

The next stages in the biomarker development process, following the discovery (Figure 6), shift from a broad and unbiased approach to a more focused and quantitative one. This change allows for the use of more advanced analytical methods [108]. The qualification phase (Figure 6) is a crucial step in the process, which confirms that the differential expression observed in the discovery phase can be seen using other methods. The primary focus of discovery and qualification is to ensure consistency between the marker and disease, with a focus on marker sensitivity rather than specificity.

In the verification phase (Figure 6), the analysis is expanded to a larger number of human samples, incorporating a wider range of cases and controls, which takes into account the variation in the population caused by environmental, biological, stochastic, and genetic factors. This phase confirms the sensitivity of the biomarker candidate and begins to evaluate its specificity [107].

The next stage of biomarker development is "assay optimization," in which the few candidate biomarkers that have shown strong performance in verification are further refined and tested. Finally, in the validation stage, a research-grade version of the final assay is tested on a large number of samples that represent the full range of variation within the target population [107].

The development of biomarkers goes through several stages, including verification and validation, before they are considered ready for commercialization. During this process, the research-grade immunoassay is refined to meet the high standards required for clinical tests. All aspects of this pipeline must be carefully considered for a successful outcome [107].

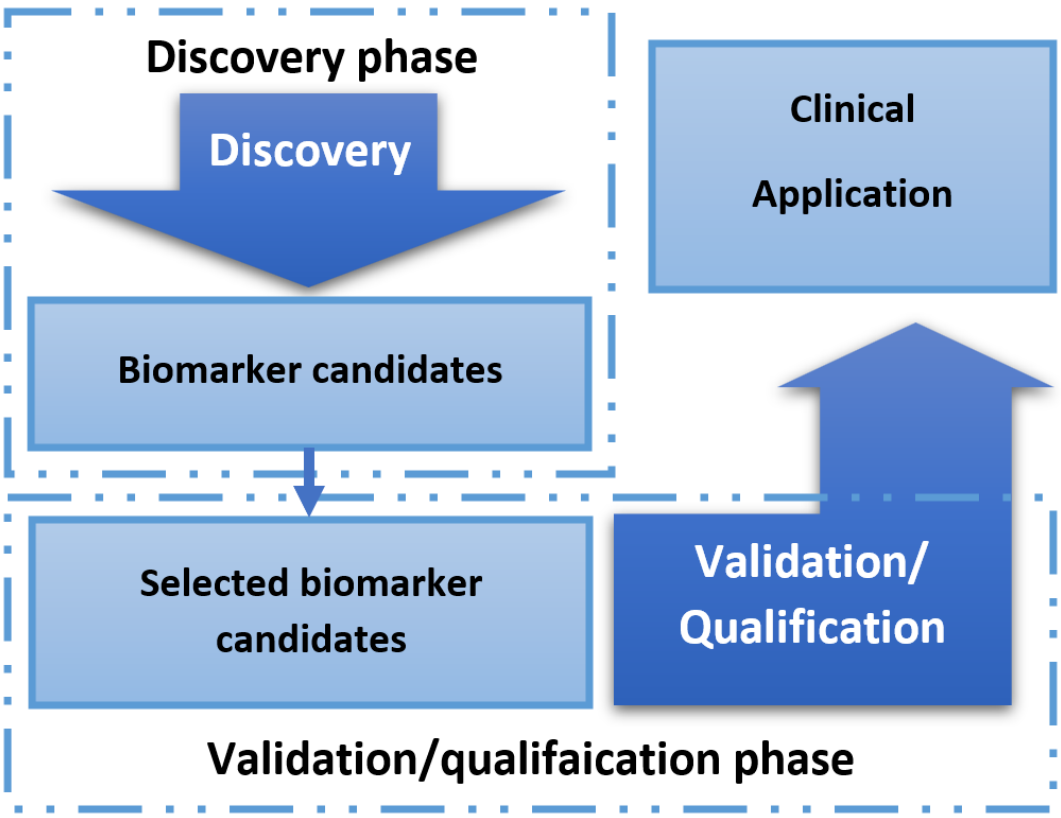


Figure 6. The presented biomarker identification pipeline consists of two main stages: discovery and validation/qualification. During the initial discovery phase, a limited number of samples are analyzed, while a more substantial number of samples are used in the later validation phase to confirm potential biomarkers before they are put into clinical use [109].

VI.1. Clinical trial design

Pre-clinical study

In this stage (Figure 7), laboratory experiments on animals or cell models are conducted to determine the potential usefulness of a particular drug, biomarker, or procedure for a specific condition. The new treatment's dosage and toxicity levels are also analyzed. It is crucial to obtain solid evidence of the safety and efficacy of the intervention before moving forward to human trials [110].

Phase 1 clinical trial designs

The results from preclinical trials often do not accurately predict real-world outcomes. If the preclinical results are promising, the next step is to submit an investigational new drug or biomarker to the regulatory agency responsible for drug approval. The first stage of human testing, known as a phase 1 trial (Figure 7), is focused on evaluating the toxicity, and pharmacokinetics, and determining an appropriate dose for further testing. Efficacy is not the primary focus but is

monitored and reported. The maximum tolerated dose is usually used in later phases, but with new and diverse drugs, the dose-response curve may plateau and a minimally effective dose may be a better target [110].

Phase 2 clinical trial designs

The purpose of a phase 2 study is to evaluate the efficacy and continued safety of a drug once its dose and safety have been established (Figure 7). This study is crucial in deciding whether the drug has enough clinical benefits to undergo a phase 3 study on a larger scale. The effectiveness of a drug is usually measured by its ability to decrease the cancer burden and is quantified by the response rate. In exceptional cases, if the drug shows significant efficacy in phase 2, the need for phase 3 testing may not be necessary. Phase 2 studies come in a variety of designs, including single-arm or randomized multiple-arm [110].

Single-arm phase 2 studies, which evaluate efficacy using historical controls, are the most common type [111]. A popular approach is Simon's two-stage design, where the enrollment process is divided into two phases. The second stage only begins if a set response criterion is met during the first stage. This design minimizes participants' exposure to an ineffective treatment [112]. On the other hand, randomized phase 2 trials offer objective comparisons but require a larger sample size. These trials usually evaluate a high probability of effectiveness in phase 3 trials, rather than measuring definitive clinical benefit. After phases 1 and 2, the sponsor and investigators may meet with the FDA to review the IND (Investigational New Drug) and determine if it is viable to proceed to a phase 3 trial. In some cases, FDA-accelerated approval may occur based on phase 2 data. Erdafitinib is an example of FGFR-TKIs that was approved for metastatic urothelial carcinoma based on outstanding results from a phase II trial (NCT02355597) [72].

Phase 3 clinical trial designs

Phase 3 clinical trials are the standard for determining the superiority of a new drug or combination, compared to the current standard of care (Figure 7). These trials compare the efficacy of the new treatment with the existing one, with improved overall survival often being the primary endpoint. In case a new drug is not worse than the standard of care in terms of efficacy, a phase 3 noninferiority trial is conducted. These trials are used for drugs that may offer advantages such as reduced toxicity or cost. However, they are complex and have limitations [113]. Phase III cancer trial design in oncology drug development may integrate biomarker-based objectives. Such design helps improve development of more effective anticancer therapies. The results of phase 3 trials

serve as the basis for FDA approval, and if the new drug is found to be safe and effective, a new drug application is submitted to the FDA. Further safety and efficacy studies for the intended population are conducted in phase 4 trials after the final approval.

Phase 4 clinical trial designs

Once a drug has received regulatory approval, Phase 4 trials (Figure 7), also referred to as post-approval or post-marketing studies, are conducted to gather additional information about the longer-term effects, both positive and negative, as well as the optimal usage of the drug. These trials are necessary as even the most thorough Phase 3 trials may not uncover issues that become evident once the drug is widely used [110].

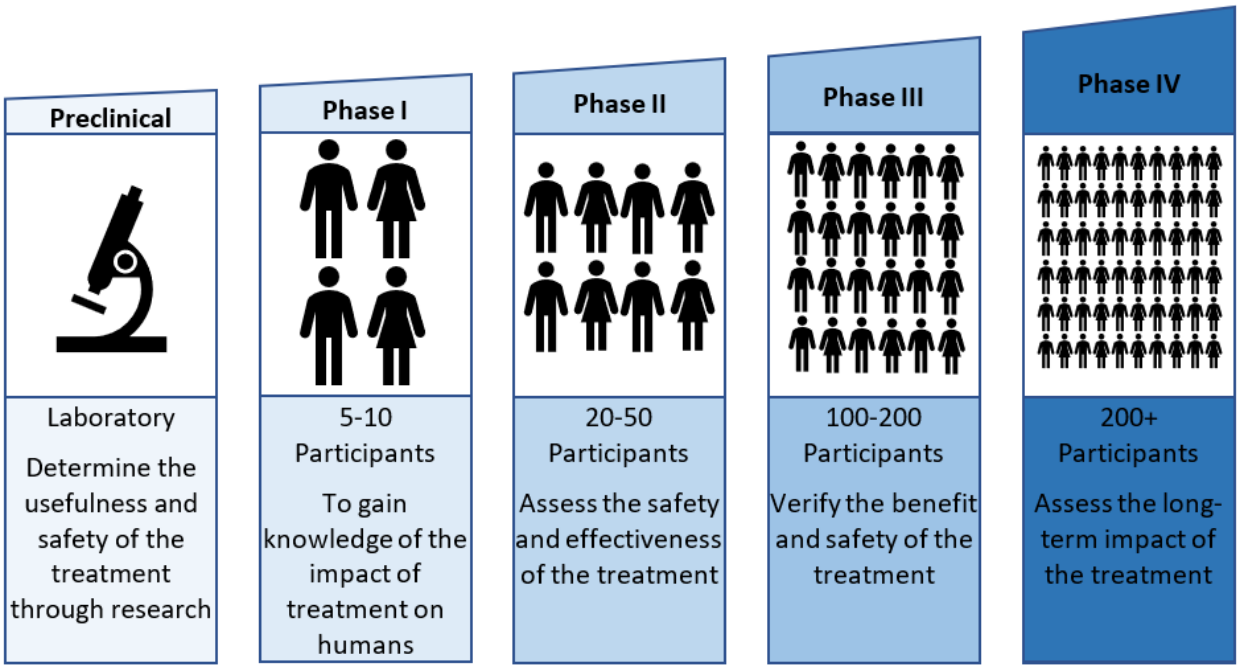


Figure 7. Phases of clinical trials.

VI.2. Biomarker-based clinical trial design

The optimal strategy for the simultaneous development of a drug and its accompanying diagnostic tool involves [114]:

(I) identifying a predictive biomarker, which is a biological measurement that can indicate a patient's likelihood of responding to the drug. This understanding is based on the mechanism of action of the drug and the role of its target in the disease's

pathophysiology. This biological knowledge is validated and improved through pre-clinical and early phase clinical trials. Most successful predictive biomarkers for cancer drugs have involved a single gene or protein, rather than a set of features. The latter type of biomarker is more commonly used as a prognostic indicator [5], reflecting the progression of the disease and the effect of standard therapy, but not as a predictive biomarker for response to specific drugs.

(II) the development of an accurate and validated test to measure the biomarker is crucial in the co-development process.

(III) employ the validated test in designing a clinical trial that explores the efficacy of the experimental drug and the correlation between its efficacy and the biomarker measurement.

Phase II trials

The evaluation of potential predictive biomarkers is often carried out during phase II trials that include patients with tumors from a single primary source. The two-stage single arm phase II design, proposed by Simon, has been expanded by Pusztai and Hess [115] and Jones and Holmgren [116], to account for a single binary candidate marker. This design aims to ensure that the potential benefits of the drug are not missed if its effects are restricted to test-positive patients, and to avoid testing too many patients if its effects are broad enough to not require a marker. Freidlin et al. [117] have proposed a design for a randomized phase II trial, which uses a single binary biomarker to determine whether the drug should advance to a phase III enrichment trial, an all-comers trial, or if it should be discontinued.

The evaluation of predictive biomarkers becomes more complex in certain phase II trials, where there isn't a known cut-point for the biomarker or multiple candidate biomarkers exist. A notable example of this is the BATTLE I trial for non-small cell lung cancer (NSCLC), which evaluated four different tests in the context of four drug regimens [118]. The trial assigned treatment to the regimens randomly, however, the randomization weights were adjusted as the trial progressed based on the treatment that performed best within each biomarker strata. The two primary goals of the adaptive randomization were to efficiently screen the four treatments across four predetermined NSCLC patient strata and to provide patients with a trial that could adapt to assign the best drug regimen for their form of the disease.

Phase IIa basket discovery trials

"Umbrella" discovery trials study advanced cancer patients with various primary disease sites that are unresponsive to typical treatments [119]. The patients undergo tumor DNA sequencing and, using a pre-determined algorithm, it is determined if there is a present "actionable" mutation - meaning a drug is available with a range of molecular targets that align with the tumor's genomic alterations, suggesting the potential for therapeutic benefit. The level of evidence that a drug is actionable for a specific mutation is often based on pre-clinical or biological data or data from a different type of tumor. Basket trials have a single drug and aim to determine the types of patients for whom it should be further developed (Figure 8). In some instances, multiple drugs are available and the trial may randomly compare outcomes for drugs selected based on the actionability rules to those chosen by physicians without access to genomic data.

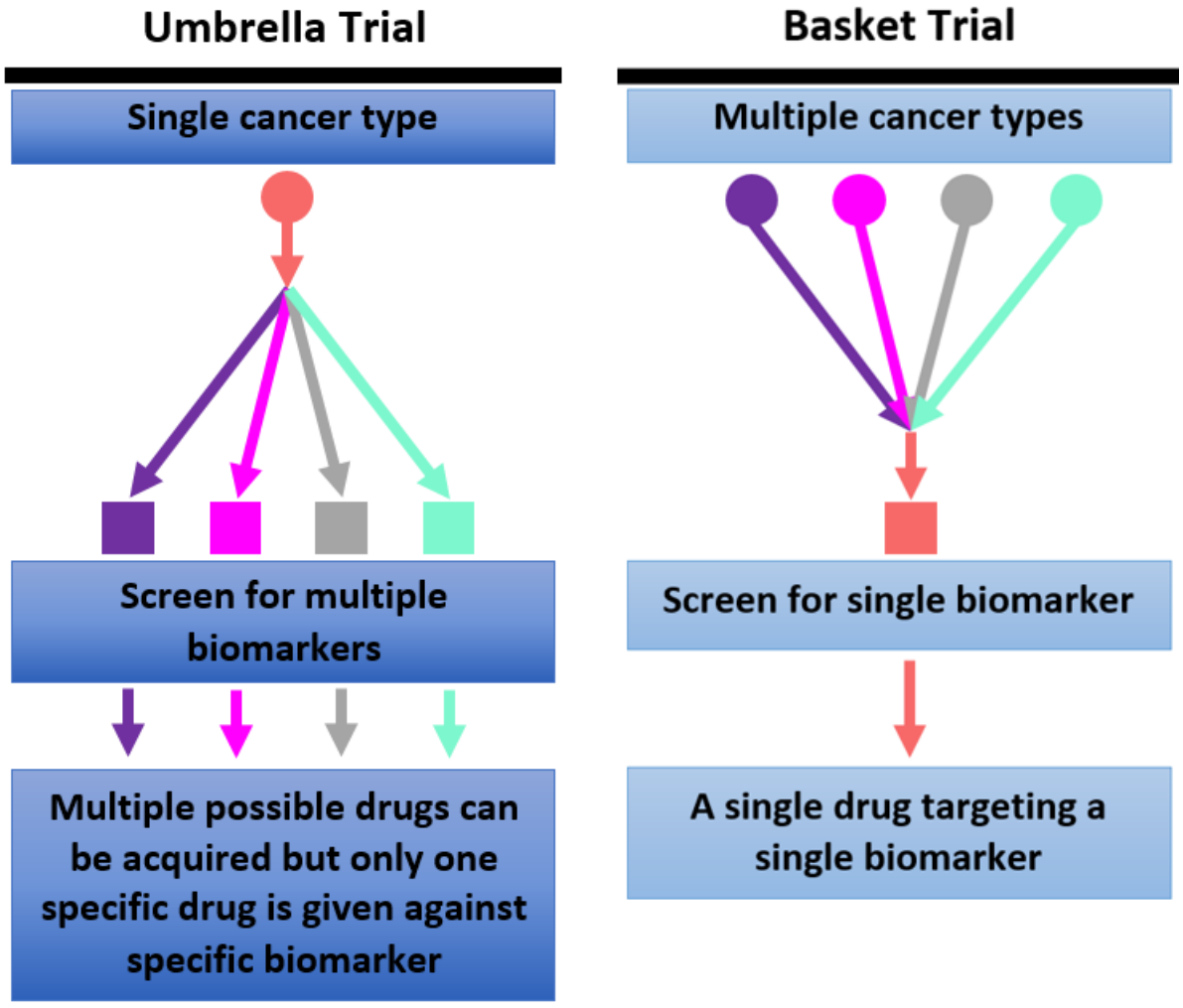


Figure 8. Schematic diagram of umbrella and basket trials.

Phase III targeted (enrichment) designs

Targeted or enrichment designs limit trial participation to patients believed to be most likely to respond to the experimental drug. This design employs a validated diagnostic test to determine eligibility for a randomized clinical trial that compares the new drug regimen with a control. Such trials have been instrumental in the approval of several drugs that have a well-understood molecular target for the disease, including trastuzumab [120], vemerafinib [121], and crizotinib [122].

The use of an enrichment design is suitable when there is solid biological evidence indicating that individuals who test negative are unlikely to see any improvement from the new medication. In these situations, including test negative patients, raises ethical questions and can lead to misinterpretation of the trial results [114].

Phase III biomarker stratified design

When a predictive biomarker has been identified, but there is not enough evidence to suggest that patients with negative test results won't benefit from the experimental treatment, it is advisable to enroll both positive and negative test result patients in the phase III trial comparing the new treatment to the control. In this scenario, it's crucial to have a pre-defined analysis plan in the trial protocol, outlining how the predictive biomarker will be used in the analysis. The plan will typically outline the testing strategy for assessment of the treatment in the positive and negative test result patients and overall. This strategy must maintain the overall type I error rate of the trial and the trial must be designed to have enough statistical power to support these tests. By balancing the randomization, the trial assures that only patients with valid test results will be included in the trial [114].

Karuri and Simon [123] proposed a phase III design for a single binary biomarker trial in which futility monitoring is done based on a joint prior distribution for the treatment effect in positive and negative patients. This design takes into account prior evidence of a reduced effect in negative patients and uses it in monitoring the trial. The design is Bayesian, but the rejection region based on posterior probability is calibrated to meet frequentist type I error requirements. This approach allows for earlier termination of accrual in negative patients compared to traditional futility analysis methods.

The authors' Hong and Simon created a run-in design that utilizes a response endpoint, such as a pharmacodynamic or immunologic effect, measured after a brief run-in period on the experimental treatment, as the predictive biomarker [124]. Another approach, known as the prospective-retrospective approach, was outlined by Simon et al. [125]. This method allows for a focused re-analysis of an already completed phase III trial using archived tumor specimens to determine the predictive value of a biomarker. The method necessitates that the majority of the specimens be stored and a single marker-focused analysis plan be formulated before the blinded assays take place. This method was applied to determine that K-RAS mutation was a negative predictive biomarker for the response of colorectal cancer patients to anti-EGFR antibodies.

Phase III adaptive

Jiang and colleagues [126] proposed a design known as the "Biomarker Adaptive Threshold Design" for trials where a biomarker exists at the start, but a clear dividing line between positive and negative patients is not set. In this design, all patients have their tumor specimens collected upon enrollment, but the biomarker value is not used to determine eligibility. The analysis plan does not require the index measurement to be conducted in real-time. At the end of the trial, the optimal threshold for the biomarker is established through the use of a pre-determined metric. The optimal threshold can also be evaluated using bootstrap resampling, which provides confidence intervals. The confidence associated with a given biomarker value x can be interpreted as the probability that a patient with this value of x will benefit from the new treatment since the treatment is believed to only be effective for patients with a biomarker value above the threshold.

The adaptive signature design approach is flexible when it comes to selecting the method for identifying a single candidate subset in which the treatment effect will be tested on the validation set. Many prediction methods can be applied using the training set, but it's important to note that the goal is not to develop a prognostic classifier but rather to classify patients based on their likelihood of benefiting from the new treatment. Matsui et al. [127] created a model that predicts a continuous score indicating the expected benefit of the new treatment compared to the control, rather than just dividing patients into two subsets. Gu et al. [128] created a two-step strategy for creating a model that predicts the outcome based on treatment and selected biomarkers. The biomarkers are chosen through a group lasso approach, which groups the main effects of a biomarker with its interactions with treatments and can be applied to two or more treatments.

VII. Concept of Diversity

The concept of diversity is widely used in various scientific fields, including ecology [129], biology [130], sociology [131], linguistics [132], and investment and portfolio theory [133]. Diversity refers to the range and distribution of specific characteristics in a given population, which can change due to intra-population interactions and environmental factors. This concept, variety, or heterogeneity can be applied to any population, including those that evolve based on their level of diversity. While diversity may seem straightforward, quantifying it can be complex, as it often requires a full distribution function, making it difficult to measure using a single metric [134, 135]. Figure 9 illustrates examples of biological population dynamics at different scales that are influenced by diversity.

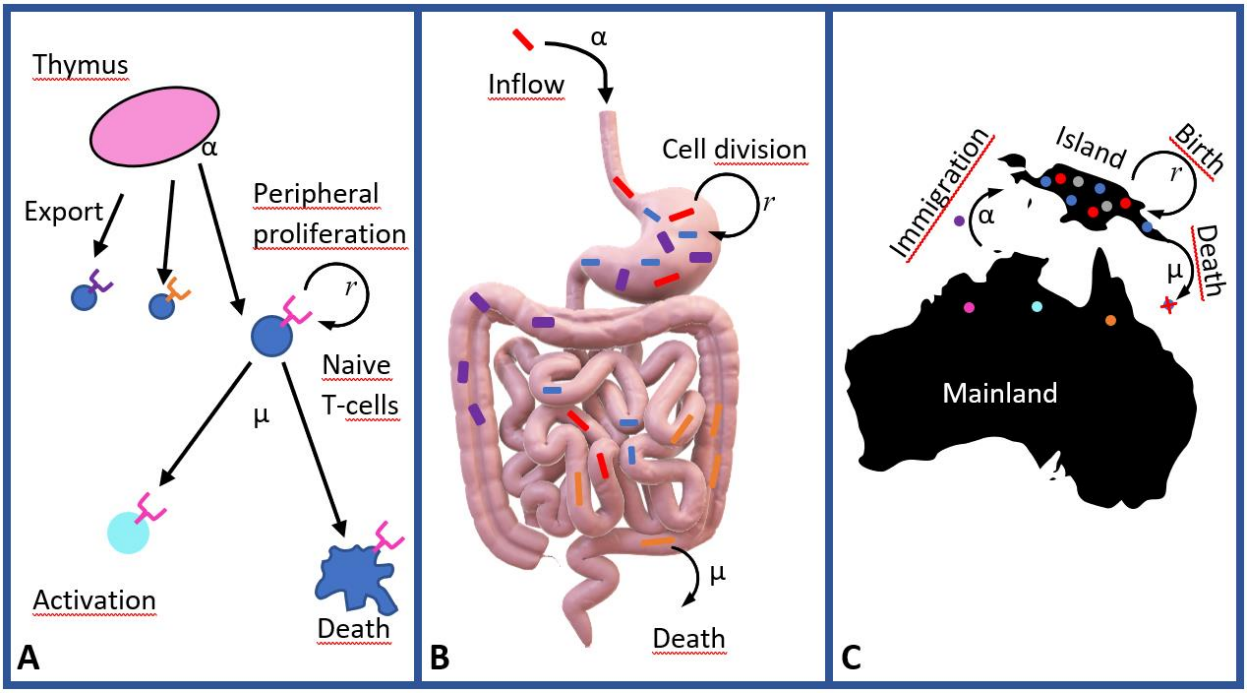


Figure 9. The concept of diversity may be significant in various complex and multicomponent populations. For instance, (A) in vertebrates, naive T cell generation takes place in the thymus, and each T cell has only one T cell receptor (TCR). Naive T cells can multiply and die in the peripheral blood, and though the number of possible T cell receptors that can be expressed is enormous (> 1015), only a few different TCRs (perhaps 106−108) exist in an organism. The diversity of the T cell receptor repertoire plays a crucial role in the organism's response to antigens. (B) In the gut, microbes are ingested and create a form of the community by competing, proliferating, and dying. (C) Island ecology showcases a time-dependent pattern of species diversity when a large number of species migrate onto an island and organisms proliferate and die [136].

Different diversity indices and related concepts are widely used in various applications, including ecology, social sciences, economics, and medicine [130, 134, 135, 137-139]. From a broader perspective, diversity indices could be beneficial in creating stable energy distribution

systems [140]. Despite the ambiguity of diversity's definition, the concept holds relevance across various disciplines and areas of application.

VII.1. Inter-tumor diversity

The motivation for large-scale genomic analysis of tumors was to identify the "driver" mutations that could serve as therapeutic targets. These mutations activate oncogenic signaling or deactivate tumor suppressor pathways, with few drivers present in any one tumor [141]. The success of targeted therapies such as imatinib in treating CML and the promising results of early trials in other treatments [142, 143] suggested that discovering new driver genes could lead to the development of therapies targeting these genes or the pathways they affect, potentially making the same kind of treatment useful in different types of tumors [144, 145].

In the early examples like ABL kinase activated by genomic translocation in CML, mutated BRAF in melanoma, EGFR mutation, or amplification in lung cancer, the driver mutation itself was targeted by the therapy [142]. The driver mutation may also make the tumor particularly sensitive to inhibition of a specific cellular function, providing a window of opportunity for a targeted attack. For instance, BRCA-mutant tumors with DNA repair deficiencies may be highly sensitive to PARP inhibitors [146, 147]. By combining genomic analysis of thousands of tumors and cell lines with comprehensive cell-line testing, it is possible to create a broad catalog linking known mutations with drug sensitivity [148]. If the mutations driving a tumor and their impact on cellular pathways were known, then a more informed choice of patient-specific or combination therapy could be made [145].

However, the identification of driver mutations through genomic analysis of solid tumors has proven to be challenging. Despite the discovery of numerous mutations, few are common enough to be considered major targets, appearing in more than 10% of cases for a single type of cancer, and even fewer are present in multiple cancer types. Some of the most frequent alterations involve losses of tumor suppressor genes, such as TP53, PTEN, and CDKN2A, which cannot be effectively targeted with current drugs [149].

The difficulty in crafting personalized cancer therapy is largely due to the diversity of driver mutations among tumors and the scarcity of targetable mutations in solid tumors. This can be seen in the NCI-MATCH trial, where alteration was found in 37.6% of the 5,954 patients and after applying clinical and molecular exclusion criteria only 17.8% of patients were assigned to

the treatment arm [150]. Thus, the tumor heterogeneity is the major obstacle to achieve the widespread availability of personalized cancer therapy.

The utilization of immunotherapy offers a solution to the difficulties posed by the genetic variability among patients' tumors. By harnessing a patient's immune system to fight cancer using tumor-specific neoantigens, the differences in tumors from other patients become irrelevant. However, even immunotherapy can be hindered by the diversity that occurs within a single patient's tumor [149].

VII.2. Inter-patient diversity

The concept of diversity is a fundamental consideration in the assessment of inter-patient variability. Inter-patient variability refers to the range of different responses that individuals can exhibit in response to a given intervention, treatment, or exposure. This variability can be attributed to a variety of factors, including genetic variation, age, sex, environmental exposure, and overall health status. For example, individuals with different genetic backgrounds may exhibit differential responses to a particular medication due to differences in drug metabolism or sensitivity. Similarly, age-related changes in physiology can influence how patients respond to treatments, with older individuals often exhibiting slower or less robust responses compared to younger individuals [151, 152].

To address this challenge, healthcare providers and researchers must take a multifaceted approach when designing and implementing treatment strategies. This approach requires a comprehensive understanding of the underlying mechanisms of disease, patient characteristics, and potential risk factors that may influence treatment outcomes. By adopting a more holistic perspective on patient care, clinicians can optimize treatment outcomes and improve overall health and well-being. Moreover, a more comprehensive understanding of inter-patient diversity has the potential to drive advancements in personalized medicine. By tailoring treatments to the specific needs of individual patients, healthcare providers can improve treatment efficacy while minimizing the risk of adverse events [152-154].

VII.3. Intra-tumor diversity

The obstacles presented by genetic diversity within an individual patient's tumor can be greater than those caused by diversity among different tumors. Intra-tumor genetic heterogeneity,

which refers to heritable differences in DNA, is the quantifiable form of diversity within a tumor that presents the biggest challenge to therapy [155, 156].

Over forty years ago, it was noted by Nowell that after tumor initiation, its evolution continues through mutation and selection [157]. This is because cancer cells have higher mutation rates compared to normal cells due to deficiencies in DNA repair. As a result, by the time a tumor becomes clinically detectable, it can become a genetically diverse group of subclones.

Figure 10 depicts the process of intra-tumor evolution. The initial mutations in the cell that starts the tumor, are referred to as "truncal" [158] and are carried by all its descendants unless some cells lose the genomic locus (or reverse this mutation). Truncal mutations encompass “drivers” that initiate the tumor as well as “passengers” that are mutated in the starting cell but don't play a role in tumor initiation [141]. Subsequent mutations that give a selective advantage for a clone, are also considered driver mutations. At the point of the presentation, the tumor may contain multiple subclones with different mutations [149].

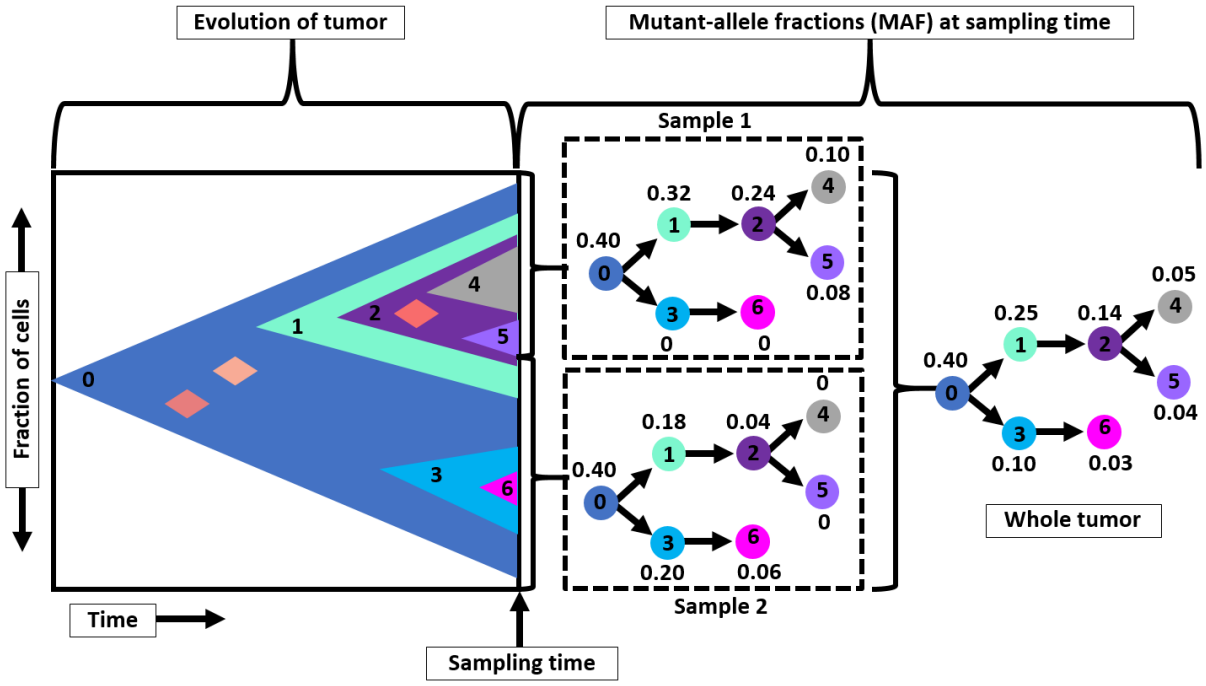


Figure 10. The evolution of subclones in a tumor. On the left, the colored regions represent cancer cells and the white background represents cells with normal DNA. The tumor-initiating clone (clone 0) gives rise to new subclones (represented by numbered triangles) that expand or die (diamonds) over time, each subclone containing all the mutations from its progenitors, as well as its own subclone-specific mutations. The mutant-allele fractions (MAF) on the right side of the diagram show the proportion of DNA in a sample with a tumor-specific mutation. These examples are based on heterozygous mutations in cancer cells and a tumor "purity" of 80%, meaning 20% of cells in the sample are normal DNA. If the mutation is heterozygous and the purity is 80%, the fraction of cells with the mutation (cancer cell fraction, CCF) is 2.5 times its MAF. The originating clone has higher MAF/CCF values compared to subclones, and mutations in subclones may go undetected either due to sampling (e.g. subclones 3 and 6 in Sample 1; subclones 4 and 5 in Sample 2) or due to having MAF values too low to be detected (e.g. subclone 2 in Sample 2) [149].

Under Nowell's model of subclonal evolution, context can transform a previously considered passenger mutation into a driver. A mutation that was once considered a passenger can become a driver in the context of tumor progression or treatment, providing resistance to a drug, independence from a primary driver mutation, or a higher likelihood of metastasis, which benefits certain subclones. Early studies of mouse cancer models support this theory [159-161].

VII.4. Challenges to conventional and targeted therapy

Two separate studies of nearly 400 head and neck squamous cell carcinoma patients [162, 163] showed that a higher degree of mutant-allele tumor heterogeneity (MATH) was linked to increased mortality. Among 2,433 breast cancer patients, those whose tumors were in the highest quartile of MATH values had a shorter survival time specifically related to breast cancer compared to those in the lowest quartile [164].

To effectively cure cancer through targeted therapy in the presence of subclonal evolution (Figure 10), all subclones must have the target, and none can have any mutations or mechanisms that nullify the therapy. The presence of intra-tumor heterogeneity can also create additional challenges. If the target is present in a minor cancer cell fraction, the effectiveness of targeted therapy may be limited. In cases where the target is not a truncal mutation, the therapy will have no direct impact on cancer cells lacking it [149].

Having a target with a complete cancer cell fraction of 1 is not enough to guarantee success with targeted therapy. When a tumor reaches a size of 10^9 cells and becomes clinically detectable, it is highly likely that at least one subclone has already developed a resistance mutation. Studies by Bozic and Nowak [165] suggest that by this point, a radiographically detectable tumor may already have up to 10 resistant subclones, each with a unique mechanism of resistance. This means that even if targeted therapy leads to temporary remission, the growth of resistant subclones will eventually cause a relapse. This pattern of initial response followed by regrowth is a common outcome with targeted therapy and has been observed in various studies [144, 166, 167]. The earlier progression of melanoma after targeted therapy is often seen in patients whose resistance mutations have a high minor allele frequency and can be detected in pretreatment tumors [168].

The use of combination therapies to treat cancer is also hindered by genetic diversity. Diversity among tumors, with respect to driver mutations, is a major challenge, as less than 10% of patients can be treated with a single targeted therapy based on the existing list of driver mutations.

Additionally, the presence of heterogeneity within a tumor may hinder the efficacy of combination therapy, as certain subclones may not possess the required targets or harbor resistance mutations. Moreover, the timing of administration can also play a role in the failure of combination therapies, as the progeny of a subclone that is resistant to the first agent may develop resistance to the second agent before it is used [169].

Given the clear difficulties posed by genetic diversity in tumors, it is essential that this diversity be considered in oncology research and clinical practice. This requires a comprehensive approach that incorporates diverse elements such as clinical trials, genomic studies, and integration of clinical and genomic information. Further research aimed at uncovering the underlying mechanisms of intra-tumor diversity may lead to innovative treatment options that make use of, rather than being limited by, this diversity [149].

VIII. Research project CELONKO

The doctoral project was carried out as part of a study entitled "Development of novel biomarkes and innovative FGFR kinases inhibitor as an anti-cancer therapy" (CELONKO) funded by the National Centre for Research and Development (NCBR), under the STRATEGMED II program. The project was carried out by Celon Pharma S.A., the inventor of the novel FGFR inhibitor [170], in a scientific-business consortium with the Institute of Tuberculosis and Lung Diseases in Warsaw, the Military Institute of Aviation Medicine in Warsaw, the Maria Sklodowska-Curie National Research Institute of Oncology in Warsaw and Gliwice, and the Medical University of Gdańsk.

The drug is intended to be used for treatment of stomach, bladder, and lung cancer. As part of the project, a diagnostic test was developed to identify patients with the known FGFR receptor aberrations. This will enable the selection of patients who will benefit the most from personalized therapy based on a novel FGFR inhibitor.

Another goal of the CELONKO project was to identify potential new candidates for biomarkers predictive of resistance to FGFR inhibitor-based therapy. Due to technical constraints it was rather impossible to search for indicators of tumor sensitivity to FGFR inhibitor. That's why we were focused on the search for potential resistance mechanisms and biomarkers, exploring cell lines with acquired resistance to FGFR inhibitor.

Initially, potential candidate selection was attempted by analyzing the signaling pathways involving FGFR receptors using the western blot technique. The results were not sufficient for selecting a biomarker, so it was decided to use an RNA sequencing (RNA-seq) experiment and subsequent data analysis to identify in-silico candidates for a predictive biomarker associated with a potential mechanism of resistance to FGFR inhibitors.

IX. Experimental design

Considering the challenges presented by genetic heterogeneity in tumors, it is imperative to take into account this diversity in both oncology research and clinical practice [149]. Thus in the CELONKO project besides a typical clinical trial on the assessment of safety and effectiveness of pan-FGFR inhibitor CPL304110 (WO/2014/141015) [170] there was also a task devoted to finding potential predictive biomarkers related to resistance to FGFR inhibitors.

In the experimental design covered in my doctoral dissertation, several types of cancer were selected, specifically lung, stomach, and bladder cancer (Figure 11). These three types of cancer were chosen because FGFR aberrations are most commonly observed in them [21, 23]. The selection of the type of cancer was also determined by the high incidence and death rate of these cancers, thus requiring new therapeutic solutions (more information is described in chapter III.1.-III.3.).

In order to address the diversity concept related to inter-tumor diversity in the experimental design covered in this doctoral dissertation, two different cell lines were chosen for each cancer type (Table 8, Figure 11). The selection was based on the presence of a molecular background that favors sensitivity to FGFR inhibitors, specifically amplification of one of the FGFR1-4 genes (Table 8). Additionally, cell lines with the highest sensitivity to the tested inhibitor were selected, as well as those for which a resistant cell line could be derived (Figure 11). To mimic intra-patient diversity, two biological replicates were used for each cell line (Figure 11).

In the scope of this experimental design, I was unable to address better the intra-tumor diversity problem. In order to do this, I would have to use samples collected from different areas of the tumor taken from one patient. Because the research work for this doctoral thesis preceded the clinical trial phases of the CELONKO project, I did not have direct access to patient-derived material. However, in an ongoing clinical trial led by Celon Pharma S.A. company various biological material is being collected which will allow us to continue research and take that aspect into consideration.

My research aimed to develop a pipeline for selecting potential biomarker candidates based on data acquired from a small sample size RNA sequencing experiment conducted on genetic material collected from human cell lines.

For my study, it was crucial that the developed pipeline was based on statistical properties, and uncomplicated, while enabling fast analysis without heavy computational burdens, resulting in cost-effective implementation. It was also essential to select biomarkers that possess the necessary characteristics of a good biomarker (described in chapters V.2. and XI.3.) and that can be used clinically in resource-limited settings, facilitating widespread adoption of proposed biomarker/s testing, mainly using immunohistochemical (IHC) staining.

The RNA-seq data obtained from the experiment (Figure 11) I have used as sample data to develop my pipeline (described in chapter XI.3. below). The inclusion of three cancer types in the experimental design (Figure 11) was done to develop a pipeline that could have a broad application for selecting biomarkers for different types of cancers, serving as a pan-cancer biomarker selection solution. Therefore, this pipeline has the potential to be applied not only for the selection of predictive biomarkers but also for other types of biomarkers and can be implemented on data from various types of cancer.

IX.1. Cell Lines and Cell Culture Reagents

SNU-16, KATO III, NCI-H1581, NCI-H1703, RT-112, and UM-UC-14 cell lines were obtained from ATCC (Table 8). NCI-H1581 cells were routinely maintained in DMEM/F12; UM-UC-14 and RT-112 in Eagle's Minimum Essential Medium (EMEM); whereas NCI-H1703, SNU-16 and KATO III were maintained in RPMI 1640 medium. All culture media contained 10% of FBS and penicillin/streptomycin (100 U/mL/100 µg/mL). Cells were grown at 37°C in a humidified atmosphere of 5% CO2. All culture media and corresponding supplements were purchased from Merck KGaA (Darmstadt, Germany) or Biowest (Riverside, MO, USA). CPL304110 (WO/2014/141015) inhibitor was provided by Celon Pharma S.A., Poland [170].

IX.2. Generation of CPL304110-Resistant Cell Lines

To develop resistance to the FGFR inhibitor (CPL304110) cell lines (Table 8) were exposed to increasing concentrations of CPL304110 (starting from 50 nM). Cells were maintained in a medium containing the inhibitor, which was replaced every three days. When the growth kinetics of treated cells were similar to wild-type cells, the concentration of CPL304110 was increased until a final concentration of 0.7 µM for SNU-16, 0.35 µM for KATO III, 2.5 µM for NCIH1581, 5 µM for NCI-H1703, 1 µM for RT-112, and 0.1 µM for UM-UC-14 was achieved

(Table 8). After 4-6 months of such culture, resistant cells were established and termed L1R, L2R, S1R, S2R, B1R, and B2R (Table 8, Figure 11).

Table 8. Cell lines used in the study.

The organ origin of the cell line	Cell line symbol	Disease	Symbol of cell line variant		Amplification	Studied inhibitor 304-110-01 (IC50 [µM])	Max concentration 304-110-01 tolerated by derived cell lines [µM]
			sensitive	resistant			
Lung	NCI-H1581	Non-small cell lung cancer. Cell type: large cell	**L1**	**L1R**	FGFR 1	0.074	2.500
	NCI-H1703	Non-small cell lung cancer. Cell type: squamous cell	**L2**	**L2R**	FGFR 1	1.300	5
Stomach	SNU 16	Gastric adenocarcinoma Derived from metastatic site: ascites.	**S1**	**S1R**	FGFR 2	0.005	0.700
	KATOIII	Gastric signet ring cell adenocarcinoma. Derived from metastatic site: pleural effusion	**S2**	**S2R**	FGFR 2	0.040	0.350
Bludder	RT112/84	Bladder carcinoma	**B1**	**B1R**	FGFR 3	0.239	1
	UM-UC 14	Renal pelvis carcinoma	**B2**	**B2R**	FGFR 3	0.031	0.100

IX.3. Cell line variants used in experimental design

For the experimental design, each mentioned cell line I had in two variants (Figure 11):

- a derived cell line resistant to the investigated FGFR inhibitor, which is CPL304110 (WO/2014/141015) [170],
- an unmodified wild-type cell line, which is also sensitive to the FGFR inhibitor.

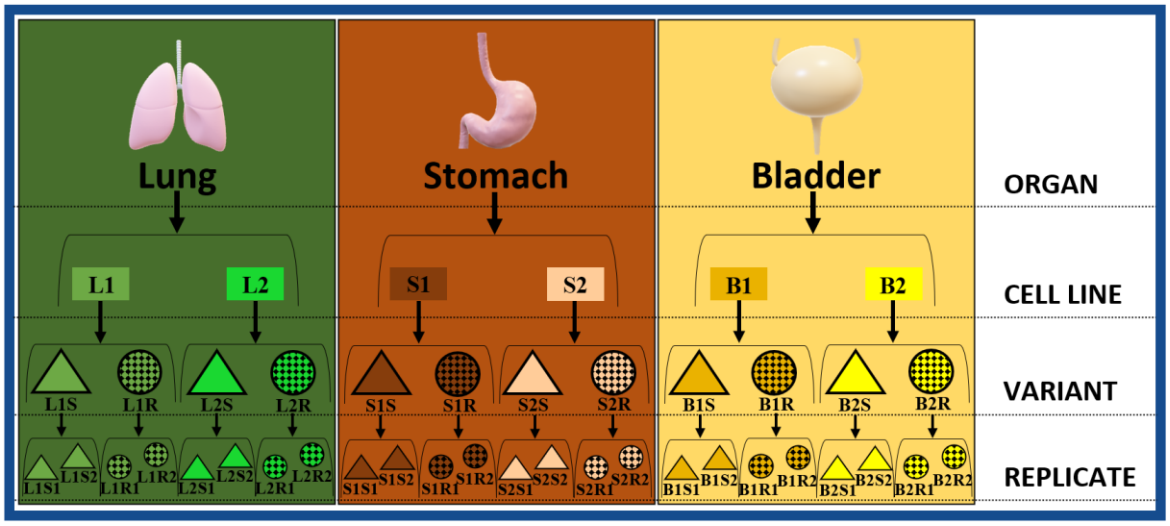


Figure 11. The RNA-seq experimental setup utilized 6 cell lines (L1: NCI-H1581, L2: NCI-H1703, S1: SNU-16, S2: KATO III, B1: RT-112, and B2: UM-UC-14), each in two variants (wild type sensitive (S: L1S, L2S, S1S, S2S, B1S, B2S) to the FGFR inhibitor and resistant cell line (R: L1R, L2R, S1R, S2R, B1R, B2R)), and two biological replicates per variant, resulting in a total of 24 experimental samples.

IX.4. Transcriptome Sequencing (RNA-seq)

For the RNA-seq experiment, SNU-16, KATO III, NCI-H1581, NCI-H1703, RT-112, and UM-UC-14 cells (sensitive and resistant cell line variants) were seeded onto 6 cm-diameter dishes. The next day, the medium was replaced with a fresh medium, and after 24 h, cells were harvested. This procedure was repeated to obtain a second biological replicate (Figure 11). Total RNA was extracted from the cells using RNeasy Plus Mini Kit, Qiagen (Hilden, Germany) with simultaneous DNase I digestion, according to the manufacturer's instructions. RNA purity and concentration were estimated with a Nanodrop ND-2000 spectrophotometer, Thermo Fisher Scientific (Waltham, MA, USA). RNA quality was assessed using the 2100 Bioanalyzer with the RNA 6000 Nano Kit, Agilent Technologies (Santa Clara, CA, USA). All the samples had an RNA integrity number (RIN) above 7.0. cDNA library preparation and transcriptome sequencing were completed by Genomed S.A., Warsaw, Poland. RNA-seq was performed using the Illumina HiSeq4000 Platform with the standard paired-end protocol (58 mln paired reads, 100 bp read length) (Figure 11).

X. Transcriptomics

In 1986, Dr. Thomas H. Roderick used for the first time the term "genomics" [171] and now the term "omics" informally encompasses various fields of biological research such as transcriptomics, proteomics, metabolomics, etc. The aim of "omics" is to collectively measure and describe the pools of biomolecules that contribute to the structures, functions, and dynamics of one or more organisms [172].

The completion of the human genome project [173] has enabled significant advances in our understanding and control of our genetic information. Additionally, high-throughput sequencing (HTS) technologies have dramatically improved transcript detection. Transcriptomics, which focuses on the heterogeneity of cell transcriptomes in specific time and space, utilizes HTS platforms such as gene expression microarrays [174], serial analysis of gene expression (SAGE) [175], massively parallel signature sequencing (MPSS) [176], RNA sequencing (RNA-seq) [177], single-cell RNA sequencing (scRNA-seq) [178], and third-generation sequencing approaches like PacBio SMRT [179] and Oxford Nanopore MinION [180]. These technologies provide insight into gene transcription levels, regulation characteristics, and molecular mechanisms involved in disease processes and regulated pathways affected by drug interventions, making them valuable tools for biomarker discovery studies [172].

X.1. RNA-sequencing (RNA-seq)

The advancements in next-generation sequencing technology have made it possible for researchers to gather much larger amounts of data than before [181-185]. RNA-sequencing (RNA-seq) in particular provides extensive insights into the gene expression levels of different organisms under different conditions with high resolution [177, 186, 187].

RNA-seq has established itself as a crucial assay for determining the relative abundance and diversity of transcripts and is widely utilized by the life sciences research community [188]. The process of RNA-seq involves extracting RNA from cells, converting it into complementary DNA (cDNA), sequencing it to produce millions of reads, and then aligning these reads to a reference genome. The number of reads mapped to a particular gene is then used as an indicator of that gene's expression level [189].

Before analyzing the differences in gene expression, the collected feature counts must undergo pre-processing, which involves trimming, filtering, and normalizing. RNA-seq gene differential expression analysis can be performed using either parametric or nonparametric methods. Parametric methods that are available in open-source packages are NBPSeq [190, 191], Voom/vst [191, 192], baySeq [193], DESeq [194], DESeq2 [195], EBSeq [196], TSPM [197], edgeR [198], ShrinkSeq [199], which can be found in R or Bioconductor. Cuffdiff2, which is part of the Cufflinks package [200], is also a parametric method. On the other hand, nonparametric methods for RNA-seq gene differential expression analysis include NOIseq [201] and SAMseq [202], both of which can be found in R or Bioconductor.

X.2. Differential analysis methods

RNA-seq procedures gather extensive information on the gene expression levels of various organisms across different conditions [177, 186, 187]. From this information, the concept of differentially expressed genes (DEGs) emerges, which are genes whose expression levels have been determined to be significantly different between two or more conditions [203, 204]. Tools have been designed to identify which genes are differentially expressed (Table 9). This differential gene expression (DGE) tools perform statistical tests on the quantification of expressed genes that are computed from the raw RNA-seq reads through mapping [204-207] and assembly [204, 208-210]. This determines which genes exhibit a statistically significant difference and provides information on the expression level and the magnitude of the difference between each gene pair. DGE analysis provides valuable insights into the genetic processes contributing to phenotypic differences, including plant growth patterns [211-213], tumor origin detection [214], and microbiome studies [215].

Table 9. Most commonly used DGE tools with their citation counts and year of release: Cuffdiff/Cuffdiff2 [200, 216], SAMseq [202], sleuth [217], baySeq [193], limma [218], NOIseq [219], DEGseq [220], edgeR [198] and DESeq2 [195]. All citations count were acquired from Google Scholar as of 9 February 2023.

DGE Tool	Citation Count	Publish Year
DESeq2 [195]	46,357	2014
edgeR [198]	29,216	2010
limma [218]	21,128	2015
Cuffdiff/Cuffdiff2 [200, 216]	3,507/11,735	2013/2012

DGE Tool	Citation Count	Publish Year
Voom [192]	4,440	2014
DEGseq [220]	3,371	2010
EBSeq [196]	1,196	2013
sleuth [217]	1,148	2017
baySeq [193]	941	2010
SAMseq [202]	512	2013
NOIseq [219]	145	2012

edgeR

The edgeR [198] method for differential analysis between two groups tests the hypothesis H0: pg1 = pg2 for each gene. This process involves using an empirical Bayes approach to regulate the degree of overdispersion among genes. The gene-wise dispersion is estimated using the conditional maximum likelihood method, which takes into account the total count for each gene. Through the empirical Bayes procedure, the dispersions are consolidated towards a common value by borrowing information from other genes. The differential expression between groups is then evaluated through an exact test that takes into account the overdispersion and resembles Fisher's exact test. Additionally, edgeR can fit a negative binomial generalized log-linear model to the read counts for each gene and perform statistical tests using likelihood ratio tests.

DESeq and DESeq2

DESeq [194] was built upon the negative binomial model introduced by edgeR and establishes a more flexible and data-driven relationship between the variance and mean. The test for differential expression between two groups (H0: pg1 = pg2) for each gene is performed using an exact test similar to Fisher's exact test. The test statistic is based on the total count of each group and the combined total count of both groups. The serial analysis of gene expression (SAGE) value is obtained by summing the probabilities of observing a value for the total count in the treatment group as extreme or more extreme, given the fixed total count across groups.

The DESeq2 method [195], an improvement on the DESeq approach, utilizes a Generalized Linear Model (GLM) for modeling the correlation between relative gene abundance and group

differences in a more complex manner. It employs a logarithmic connection between the relative gene abundance and a design matrix. The DESeq2 incorporates the fold change estimate and the dispersion estimate from an empirical Bayes approach and performs differential expression analysis using a Wald test.

baySeq

The baySeq method [193] determines differentially expressed genes by combining the empirical Bayes approach with the observed data to calculate the posterior probability of a model. The method assumes a negative binomial distribution for the data and uses a prior distribution, determined from the whole dataset, to estimate the dispersion of the data using the maximum likelihood method. baySeq also generates a posterior probability of non-differential expression and a Bayesian FDR estimate to identify the significantly differentially expressed genes.

EBSeq

EBSeq [196], initially created for finding differentially expressed isoforms, has been demonstrated to be a strong method for identifying genes with different expression levels. For a comparison between two groups, EBSeq performs tests using the negative binomial-beta empirical Bayes model and calculates the posterior probability of differential expression through Bayes' rule using the EM algorithm. In addition, EBSeq provides a Bayesian FDR estimate to aid in the identification of significantly differentially expressed genes.

Voom

The voom method [192] is a linear modeling approach that models count data, differing from the negative binomial model approach. It transforms the count data into log-counts per million (log-cpm) and uses moderated t-statistics for gene differential expression analysis.

SAMseq

The SAMseq method [202] is a nonparametric approach for differential analysis of RNA-seq count data, which doesn't rely on Poisson or negative binomial models. SAMseq employs the two-sample Wilcoxon rank statistic for comparisons between two groups. According to simulations, a sample size of 20 is considered ample for SAMseq. It adjusts for different sequencing depths by resampling from a Poisson distribution and uses permutation to generate the null distribution of the Wilcoxon rank statistic and calculate the FDR.

NOIseq

NOIseq [201] analyzes sequencing-depth corrected and normalized RNA-seq count data by comparing the logarithm of the fold change and absolute expression differences between groups to the noise distribution. If the logarithm of fold change and absolute expression differences have a probability higher than 0.8 of exceeding the noise values, the gene is considered differentially expressed.

In a comparison study by Seyednasrollah et al. [221], eight software packages for RNA-seq differential analysis were evaluated, including edgeR, DESeq2, baySeq, NOIseq, SAMseq, limma, and Cuffdiff2. The researchers used two public RNA-seq datasets to compare the number of rejections, and estimated proportion of false discoveries across the eight methods. The results of the comparison showed substantial differences between the methods, with limma, and DESeq2 being recommended as the safest choice for small sample sizes (with fewer than 5 samples in each group).

XI. Pipeline

XI.1. Dimensionality reduction techniques

Despite avoiding the bias of using a predefined gene set, whole-transcriptome analyses are typically too complex for most modeling algorithms to process directly due to their high dimensionality [222]. Furthermore, biological systems have lower intrinsic dimensionality. As an example we have differentiating hematopoietic cells that can be represented by two or more dimensions, indicating its differentiation progress towards a particular cell type, and current cell-cycle stage [223].

Dimensionality reduction methods have a long history, for example, widely used principal component analysis (PCA) has been used since 1901. With the advent of RNA-seq technology, researchers favor this linear dimension-reduction method. Non-linear methods such as UMAP (uniform manifold approximation and projection), and t-SNE (t-distributed stochastic neighbor embedding) have also been used to reduce dimension. With the rise of neural networks, there are many dimensionality reduction methods based on neural networks, such as VAE (variational autoencoder). New theoretical frameworks such as SIMLR (single-cell interpretation via multi-kernel learning) based on the above methods are being developed to handle increasingly diverse RNA-seq data [223, 224].

Dimensionality reduction techniques encompass two main approaches: Feature Selection and Feature Extraction. Feature selection involves selecting the most relevant features from a high-dimensional dataset using objective measures, to remove irrelevant, redundant, and noisy data to reduce the number of features [225]. Feature extraction, on the other hand, is used to extract the most relevant information from the original data and represent it in a lower dimensional space. This approach selects a new set of features and transforms them into a linear or nonlinear combination of the original features [226]. These techniques are crucial for reducing the dimensionality of high-dimensional data while retaining the relevant information and can be used independently or in combination to enhance performance metrics like accuracy.

Dimensionality reduction is particularly important when analyzing high-dimensional data, such as RNA-seq data [227]. Properly selected dimensionality reduction algorithms can help improve the evaluation and classification performance of different approaches in terms of metrics

like accuracy, sensitivity, specificity, recall, computational scalability, computational cost, and more [228].

XI.1.1.Feature selection dimensionality reduction approach

The approach of feature selection in dimensionality reduction is focused on reducing data by eliminating irrelevant and redundant features [229]. This technique can enhance the accuracy of predictions, improve the clarity of information, and facilitate the visualization of data. The feature selection technique comprises three variable selection categories: filter, wrapper, and embedded methods [230]. With the abundance of irrelevant and redundant features in datasets, it is essential to apply a proficient feature selection technique for the extraction of relevant features. These techniques are particularly important in selecting informative genes for RNA-seq data classification in the prediction and diagnosis of diseases, thus improving the accuracy of classification. In recent years, various feature selection approaches have been applied to medical datasets, particularly in the prediction of chronic diseases such as diabetes, hypertension, and heart diseases, among others. By applying efficient feature selection techniques, significant and non-redundant attributes can be extracted from large medical datasets, resulting in more accurate results and facilitating efficient learning algorithms [223].

The filter-based feature selection method is a non-dependent approach that efficiently reduces the number of features in a dataset. This approach uses statistical procedures to score the features and is robust against over-fitting. It is computationally less expensive and provides fast and high-quality results, making it ideal for big databases. However, this approach has limitations as it does not consider feature dependencies and classifier interactions. Moreover, it may fail to select the most useful features. The filter algorithms are evaluated based on distance, information, dependency, and consistency. Some of the popular filter-based feature selection algorithms for RNA-seq include ANOVA, T-test, Information gain, Fisher score, Chi-squared test, and Correlation-based Feature Selection (CBFS) [223].

Wrapper-based feature selection is a method of selecting relevant features by considering the performance of a learning algorithm. It searches for the optimal subset of features by using a specific classifier to evaluate the quality of the selected features. The classifier is run multiple times to assess the quality of each feature, and a score is assigned based on the accuracy of the model. This approach considers feature dependencies and has been shown to result in improved predictive metrics and classifier performance compared to filter-based methods. However, wrapper-based methods can be computationally expensive, requiring the use of an additional learning

algorithm, and may result in overfitting on small training datasets. Methods that can be applied to RNA-seq include Sequential Forward Selection (SFS), Genetic Algorithms (GA), Recursive Feature Elimination (RFE), and Backward Elimination Method, among others [223].

Embedded feature selection methods use the learning process to guide the feature selection process, often referred to as the nested subset method. They measure the usefulness of feature subsets during the training process, specifically to optimize the performance of the learning algorithm. This approach is computationally inexpensive, less prone to over-fitting, and better captures dependencies between features, resulting in faster solutions and better classifiers. However, these methods have limitations such as taking dependent classification decisions, affecting the selected features, and being computationally costly. Embedded feature selection methods include Decision Trees, Support Vector Machines, Least Absolute Shrinkage and Selection Operator, Elastic Net, Ridge Regression, Artificial Neural Networks, and Sequential Forward Selection [223].

XI.1.2.Feature extraction dimensionality reduction approach

The dimensionality reduction approach called feature extraction involves transforming a dataset into a simpler representation of features by creating new optimal component features. This approach is a general method that uses techniques such as Principal Component Analysis (PCA), Non-Linear PCA, Kernel-PCA, and Independent Component Analysis [230].

Principal component analysis (PCA) [231], is a widely used statistical technique for analyzing large, and complex datasets. The goal of PCA is to identify patterns and structures within the data that help explain the variability among the samples. PCA works by transforming the original data into a new set of variables named principal components (PC), which are linear combinations of the original features. These principal components are ordered in decreasing order of explained variance, meaning that the first principal component accounts for the largest amount of variability in the data, followed by the second, third, and so on. This iterative process is continued until the new component is almost ineffective or reaches the threshold set by users. By examining the loadings of each variable on each principal component, one can identify which variables contribute the most to the variability observed in the data.

Independent component analysis (ICA) [232], also known as blind source separation (BSS), is a statistical technique used to uncover the underlying factors of random variables, meas-

ured signals, and values. It linearly transforms variables into independent components with minimal statistical dependencies. ICA is different from PCA in that it requires source signals to meet two conditions: independence, and non-Gaussian distribution.

RNA-seq data often experience dropout events, especially in single-cell RNA-seq (scRNA-seq), which can render classic dimensionality reduction algorithms inappropriate. Pierson and Yau [233] proposed a modification of the factor analysis framework to address this issue, resulting in a method called zero-inflated factor analysis (ZIFA) that utilizes an additional zero-inflation modulation layer to reduce the dimension of scRNA-seq data. While ZIFA has more powerful projection capabilities than the above-mentioned linear methods, its use of the zero-inflation model comes at the cost of increased computational complexity.

By utilizing the variational sparse approximation of the Bayesian Gaussian process latent variable model, GrandPrix [234] projects data into lower dimensional spaces with high efficiency using a small number of inducing points to produce a full posterior distribution. The algorithm optimizes the coordinate position in the latent space by maximizing the joint density of the observation data and subsequently establishes a mapping from the low-dimensional space to the high-dimensional space.

t-distributed stochastic neighbor embedding (t-SNE) is a cutting-edge technique for reducing the dimensionality of data with non-linear structure, creating a low-dimensional representation that retains local structure [235]. The algorithm is built on the foundation of SNE (Stochastic Neighbor Embedding by Hinton and Roweis [236]), which converts high-dimensional distances between points into conditional probabilities of similarities. The modifications to SNE in t-SNE include a symmetric version, and the use of a Student's t distribution to measure similarities in the low-dimensional space.

Uniform manifold approximation and projection (UMAP) is a non-linear dimension reduction technique that outperforms t-SNE in both global structure preservation, and computational efficiency [237]. UMAP assumes that data are uniformly distributed on a locally connected Riemannian manifold, and models the manifold's fuzzy topology to find a low-dimensional embedding. The algorithm involves building a weighted k-neighbor graph using the nearest-neighbor descent algorithm [238] and computing a low-dimensional representation that preserves desired characteristics of this graph.

Deep count autoencoder (DCA), proposed by Eraslan et al. [239], is a deep learning method that performs denoising, and imputation of RNA-seq data in a single step. DCA uses an autoencoder with three hidden layers of 64, 32, and 64 neurons, respectively, and zero-inflated negative binomial (ZINB) loss functions [240] that learn three parameters (mean, dispersion, and dropout) of the negative binomial distribution. The primary output of DCA is the denoised reconstruction, which is represented by the mean parameter of the distribution. DCA is highly parallelizable using a graphics processing unit (GPU) for increased speed and can capture the complexity, and non-linearity of RNA-seq data.

Xiang et al. [224], to develop a strategy to evaluate the stability, accuracy, and computing cost of dimensionality reduction methods performed a comparison study of ten dimensionality reduction methods used for high-dimensional RNA-seq data. This study included among others comparison of methods like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), Single-cell interpretation via multi-kernel learning (SIMLR), Independent component analysis (ICA), and deep count autoencoder (DCA). Their results overall showed that t-SNE had the highest accuracy, and computing cost, while UMAP was the most stable with moderate accuracy, and the second highest computing cost. However, the performance of each method varied across evaluation criteria. For example, SIMLR, and PCA outperformed UMAP in terms of accuracy, but SIMLR had weaker computing cost, and PCA had weaker stability. Despite being time-efficient, linear techniques like PCA, and ICA performed poorly in highly heterogeneous data.

XI.2. RNA-seq data analysis using standard pipeline

After standard quality control, the raw sequencing data I analyzed using the standard pipeline described in chapter X.1. This pipeline included quantifying using the Salmon tool against the reference genome GRCh38 (hg38) [241]. Quantified transcripts I have imported into the R environment with a tximport v. 1.22.0 [242]. Low-abundance genes I have prefiltered, keeping only rows with at least 10 reads total. Gene counts I have normalized using the median-of-ratios method [195].

As described in chapter X.2. from a comparison study by Seyednasrollah et al. [221] on different software packages for RNA-seq differential analysis we know that DESeq2 is recommended as the safest choice for small sample sizes. Thus considering the main limitation of the experimental design covered in my dissertation related to the small sample size (Figure 11) the

differentially expressed genes I have identified using the DESeq2 package [195] version 1.34.0, with FDR (false discovery rate) adjusted (Benjamini–Hochberg correction) P value (short: q value) cutoff 0.050 (cutoff was selected based on this literature position [243] and my own experience), and log2 fold change (log2FC) cutoff 0.500 (the choice of this cutoff threshold was arbitrary, as it is the most commonly used for log2FC and, at the same time, not the most rigorous). 192 differentially expressed genes have been identified for lung cancer, while 1,109, and 552 differentially expressed genes were detected for stomach, and bladder cancer, respectively (Figure 16).

Upon closer examination of the obtained results, I noticed that the standard method of selecting DEGs produces many results that do not meet the requirements set for biomarkers (described in chapter V.2.). For example, many results did not have a consistent direction of change but rather were random as shown example in Figure 12.A, which would indicate the potential low sensitivity of such a biomarker candidate, and low reproducibility of the testing result. A predictive biomarker at the time of testing in a patient with cancer that is not sensitive to e.g. FGFR inhibitor-based therapies cannot at one time indicate the presence of a particular resistance mechanism, and at another time, its absence. Other results, despite a large fold change difference between the compared R, and S variants (Figure 11), had a difference too small to reach the detectable threshold technics used in daily diagnostic clinical practice (Figure 12.B). Furthermore, some results, despite having a sufficiently large difference, had a small minimal fold change (*minFC* – explained more below in chapter XI.3.), and as a consequence, the biological effect could be undetectable for such a detected difference [195, 244] (Figure 12.C). Such a potential biomarker candidate would exhibit very low specificity and sensitivity if it were not possible to detect a difference in its level between healthy tissue and diseased tissue, and if its level were not disease-specific but varied depending on factors other than the presence or absence of the disease.

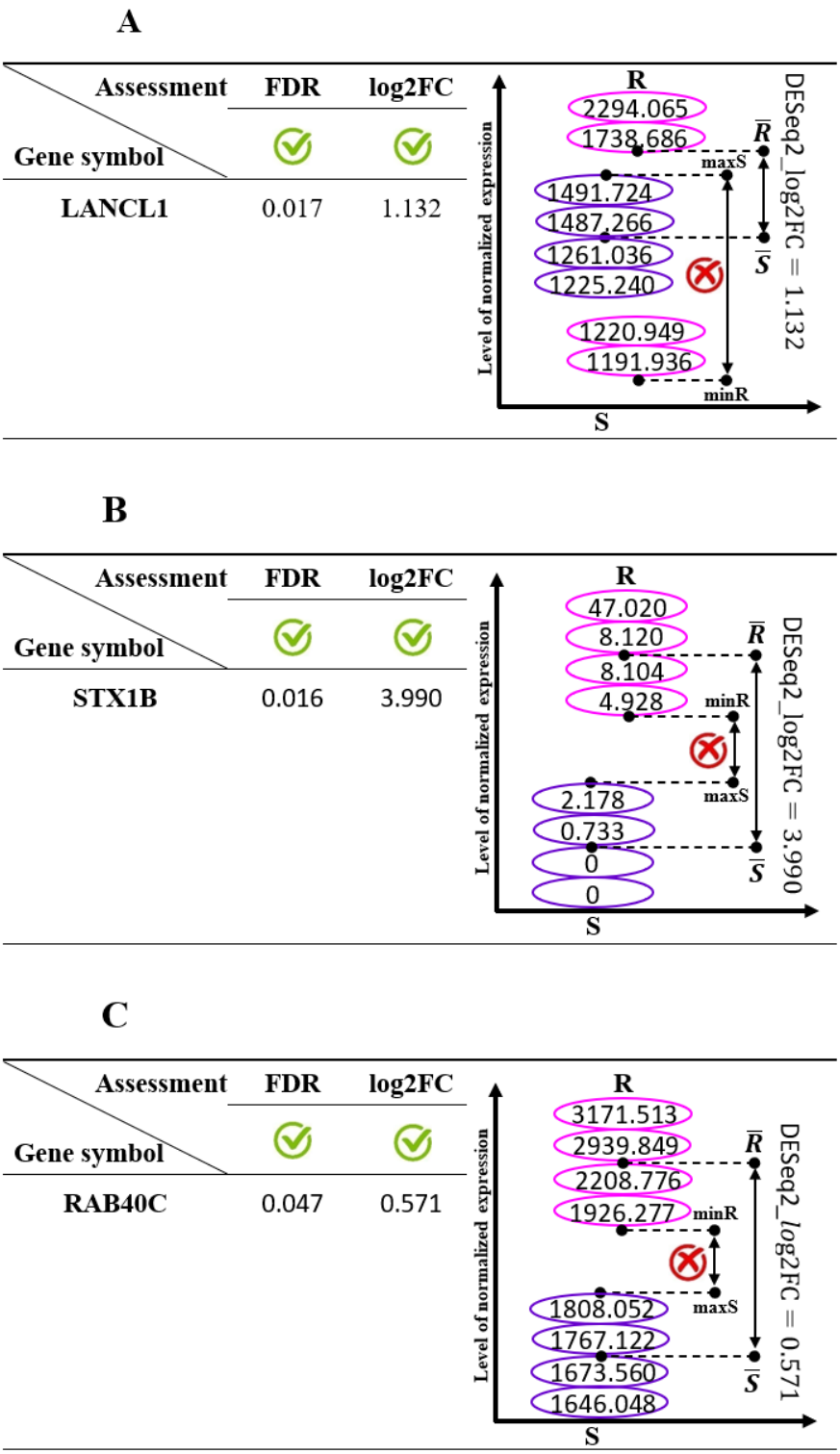


Figure 12. Examples of candidate biomarkers identified by the standard RNA-seq data analysis pipeline with significant q value (FDR) < 0.050, and proper fold change (log2FC) value > 0.500, but lacking the characteristics that a biomarker should possess: (A) LANCL1 is an example with no consistent direction of change, (B) STX1B is an example that, despite a large fold change difference between the compared R, and S variants, has a very small difference between the extreme internal values of the R, and S sets (minR & maxS), (C) RAB40C is an example that, although it exhibits the appropriate minimum difference between the extreme internal values of the R, and S sets, the fold change between these values is very small, making it unlikely to detect such a difference using methods typically used in clinical practice. The designation ✓ indicates a desired result, while ⊗ indicates an undesired result.

Theoretically, one could try to remove such results by reducing the number of features through the adoption of a specific cutoff point for some significance parameter or effect size, such as q value or fold change, respectively. However, even after applying such a cutoff or by tightening the criteria for the threshold, not all results that do not meet the criteria of a valid biomarker are removed.

The occurrence of such unwanted results is largely related to the limitation associated with a small number of samples covered in the experimental design. However, due to cost reasons, many studies are conducted with a small number of replicates, so solutions for this type of research are much needed.

Another possibility to obtain discriminative features, one might suggest, is to apply other dimensionality reduction methods, among others, UMAP (Uniform Manifold Approximation and Projection) and PCA (Principal Component Analysis) (described more in chapter XI.1.). However, my research aims to develop a pipeline for selecting potential biomarker candidates possessing characteristics suitable for a clinical biomarker that can be applied in resource-limited settings, facilitating widespread adoption of proposed biomarker/s testing, mainly using immunohisto-chemical (IHC) labeling. Therefore as those methods would indicate detecting biomarkers by some HTS technique, thus those methods were not in the line with my research.

XI.3. Pipeline development - PREDICT

In order to address the objectives of this dissertation, in particular, to account for the characteristics of a clinical biomarker, I developed a "**P**ipeline for **R**apid **E**valuation and **D**iscovery of **I**mportant biomarker **C**andida**T**es" (**PREDICT**) that allows selecting candidates with desired characteristics of a proper biomarker.

XI.3.1. *minFC & minDiff* definitions

The PREDICT pipeline includes two measures, namely the minimal Fold Change (*minFC*), and the minimal Difference (*minDiff*) (Figure 13. A, and B, respectively).

minFC (minimal Fold Change) (Figure 13. A) – let $X = \{x_1, x_2, ..., x_n\}$ be a set of expression levels measurements of a particular gene for samples belonging to one group, and let $Y = \{y_1, y_2, ..., y_m\}$ be a set of expression levels measurements of a particular gene for samples belonging to the other group. We define *minFC* as:

$$minFC = \begin{cases} \frac{minX}{maxY} & if\ \bar{X} > \bar{Y} \\ 1 & if\ \bar{X} = \bar{Y} \\ \frac{minY}{maxX} & if\ \bar{X} < \bar{Y} \end{cases}$$

where *minX* and *minY* denote the lowest value in set *X*, and *Y* respectively, and *maxX*, and *maxY* denotes the highest value in set *X*, and *Y* respectively. $\bar{X}$, and $\bar{Y}$ denotes the mean value for set *X*, and *Y* respectively. *minFC* > 1 (Log2*minFC* value > 0) shows that expression value intervals for the groups do not overlap, and *minFC* ≤ 1 (log2*minFC* value ≤ 0) shows that expression value intervals for the groups do overlap.

I adopted a threshold of log2*minFC* = 0.100. This threshold was based on my expertise, and literature reports on the potential level of FC above which biologically meaningful results are considered to occur [195, 244]. Genes with the log2*minFC* value below the threshold are filtered out.

minDiff (minimal Difference) (Figure 13. B) – let $X = \{x_1, x_2, ..., x_n\}$ be a set of expression levels measurements of a particular gene for samples belonging to one group, and let $Y = \{y_1, y_2, ..., y_m\}$ be a set of expression levels measurements of a particular gene for samples belonging to the other group. We define *minDiff* as:

$$minDiff = \begin{cases} minX - maxY & if\ \bar{X} \geq \bar{Y} \\ minY - maxX & if\ \bar{X} < \bar{Y} \end{cases}$$

where *minX* and *minY* denote the lowest value in set *X*, and *Y* respectively, and *maxX*, and *maxY* denotes the highest value in set *X*, and *Y* respectively. $\bar{X}$, and $\bar{Y}$ denotes the mean value for set *X*, and *Y* respectively. *minDiff* value > 0 shows that expression value intervals for the groups do not overlap, and *minDiff* value ≤ 0 shows that expression value intervals for the groups do overlap.

I adopted a threshold of *minDiff* = 100. When establishing this threshold, the possibility of detecting such a difference using methods used to test biomarkers in clinical practice, such as IHC, ELISA, and qPCR, as well as whether such a difference would be biologically meaningful, was evaluated. To do so, the level of normalized readings (mean readings for the Sensitive samples: LS, SS, and BS (Figure 11)) for proteins that are already recognized as specific to a given organ, in our case, the lung, stomach, and bladder (Table 10), was evaluated. Information about such proteins was obtained from the Human Protein Atlas [245-247], and in most cases additionally evaluated in GeneCards®: The Human Gene Database [248]. Since these are proteins with a characterized biological function in a given organ, their expression levels must be detectable by the standard techniques mentioned above. Therefore, the presence of such a minimal difference between the test, and control samples will be detectable using the above-mentioned techniques. Thus, candidates for biomarkers meeting such criteria will also meet the condition of their applicability while maintaining low detection costs. Since 75% of the results were within the range of up to 86 (Q3 = 86.107), a value of 100 I adopted as the minimal difference threshold (Table 10). Genes with the *minDiff* value below the threshold are filtered out.

Table 10. Normalized counts for genes/proteins specific to the lung or stomach or bladder.

Categorie	Gene	Description	Mean expression for sensitive cells
stomach	*GAST*	gastrin	0.506
lung	*SFTPA2*	surfactant protein A2	0.589
stomach	*PGA3*	pepsinogen A3	0.850
lung	*SFTPA1*	surfactant protein A1	1.276
stomach	*GKN2*	gastrokine 2	1.393
lung	*SFTPD*	surfactant protein D	2.020
lung	*SLC34A2*	solute carrier family 34 member 2	2.389
lung	*SCGB3A2*	secretoglobin family 3A member 2	2.592
lung	*NAPSA*	napsin A aspartic peptidase	13.318
lung	*MS4A15* [249, 250]	membrane spanning 4-domains A15	24.137
lung	*RTKN2* [251, 252]	rhotekin 2	45.110
lung	*SFTPB*	surfactant protein B	57.718
bladder	*IL23A* [253]	interleukin 23 subunit alpha	68.238
bladder	*MMP13* [254, 255]	matrix metallopeptidase 13	139.716
bladder	*UPK2*	uroplakin 2	1486.626
stomach	*MUC5AC* [256]	mucin 5AC, oligomeric mucus/gel-forming	2054.960
stomach	*TFF1*	trefoil factor 1	2889.759

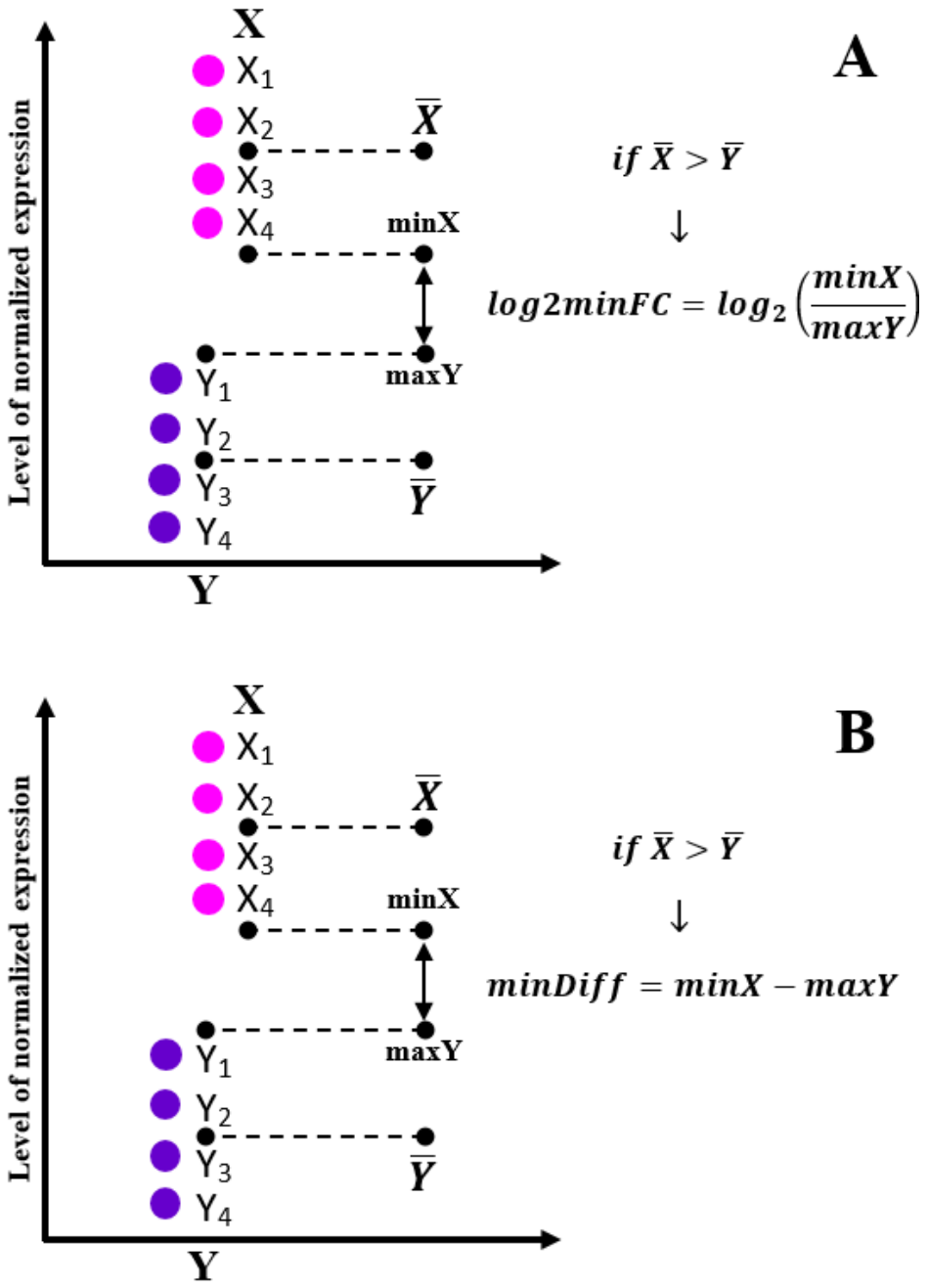


Figure 13. Scheme of the (A) *minFC*, and (B) *minDiff* measures.

XI.3.2. Pipeline: PREDICT

The PREDICT (Figure 14) is a pipeline that is applied to the results of standard RNA-seq data analysis as described in chapters X.1. and XI.2. Having DEGs identified using the DESeq2 package [195] we calculate measures, namely *minFC* (minimal Fold Change), and *minDiff* (minimal Difference) (Figure 13). Having such a list of DEGs with the fold change (FC) values, q values, *minFC*, and *minDiff* values we filter out DEGs with log2FC below threshold 0.500, and then we filter out DEGs with q value below threshold 0.050. Next for that list, we select results over the *minFC* threshold of 0.100, and then those that are over the *minDiff* threshold of 100 (Figure 14).

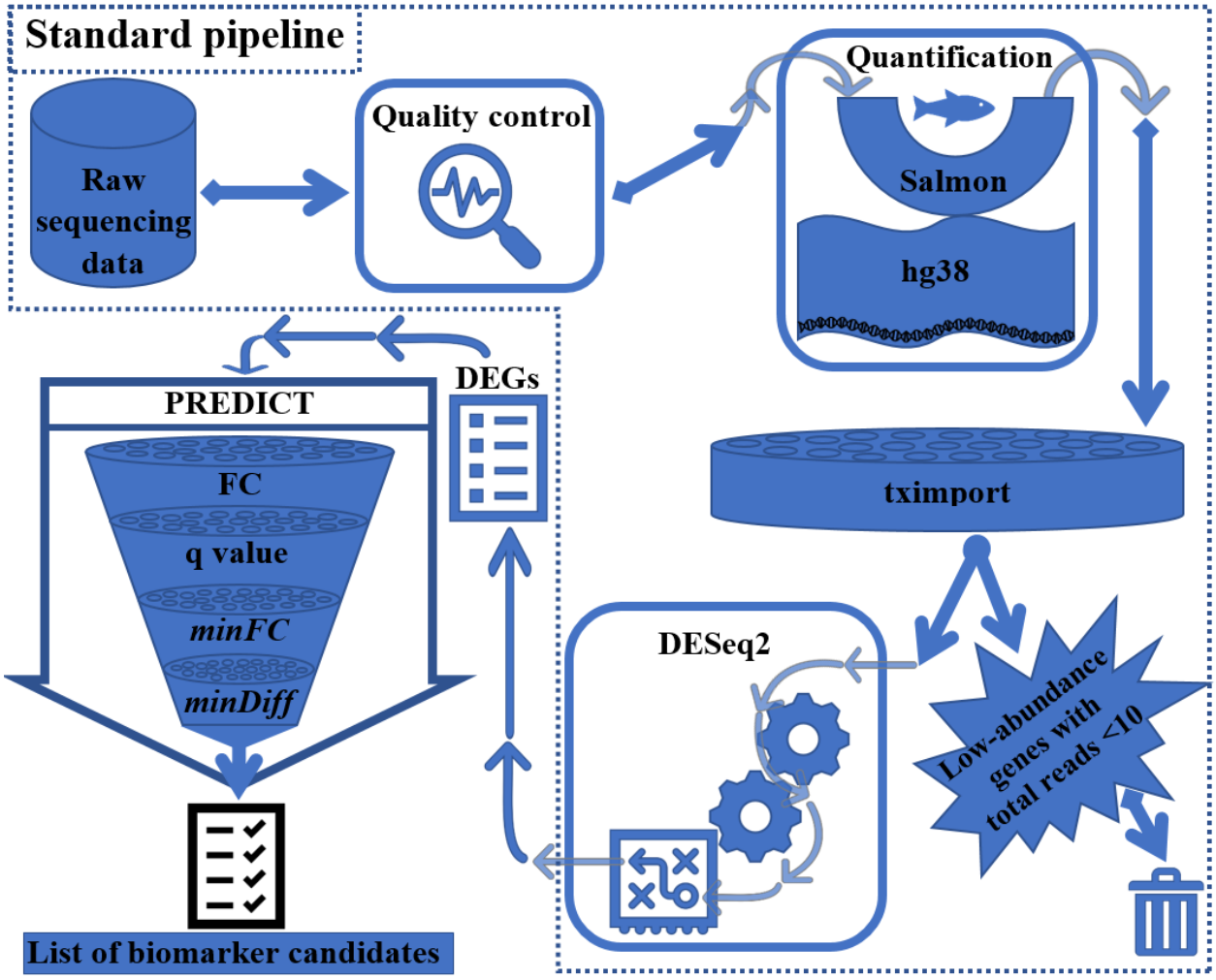


Figure 14. Scheme of the standard pipeline for RNA-seq data analysis, and scheme of Pipeline for Rapid Evaluation and Discovery of Important biomarker CandidaTes (PREDICT).

When selecting a biomarker candidate, it is also important, as I mentioned earlier in chapter XI.2., that the direction of the detected change in the biomarker level is always the same, thus

enabling the reproducibility of testing results. Taking into account the measures *minFC*, and *minDiff* with their thresholds in the PREDICT pipeline it ensures that this condition is met. Additionally, both these measures enable this pipeline to guarantee that the expression value intervals for the groups do not overlap.

Simultaneous use of both *minDiff*, and *minFC* measures with their thresholds (sequentially after considering fold change, and q value threshold) is necessary because their single use does not allow filtering out all incorrect results (Figure 15). As we can see in the Venn diagram presented below (Figure 15) that for example in the case of stomach cancer if we would apply just the *minFC* threshold we would have 166 unwanted results (by results I mean DEGs), and in the case of the *minDiff* threshold we would have 9 undesirable results. The *minFC* measure ensures the retention of results with a greater log2*minFC* value than log2*minFC* = 0.100, but at the same time may pass, for example, biomarkers with a result of normalized reads as S: 0.01 vs R: 0.03, which are likely to be difficult to be applied in clinical practice, as it would be impossible to detect them. In the case of the *minDiff* measure alone, biomarkers with results such as S: 12,000 vs R: 12,120 may pass, for which such a difference may not have biological significance. Therefore, the incorporation of both measures in the pipeline allows for the removal of all unwanted results and the selection of candidates who are more likely to exhibit appropriately high sensitivity and specificity.

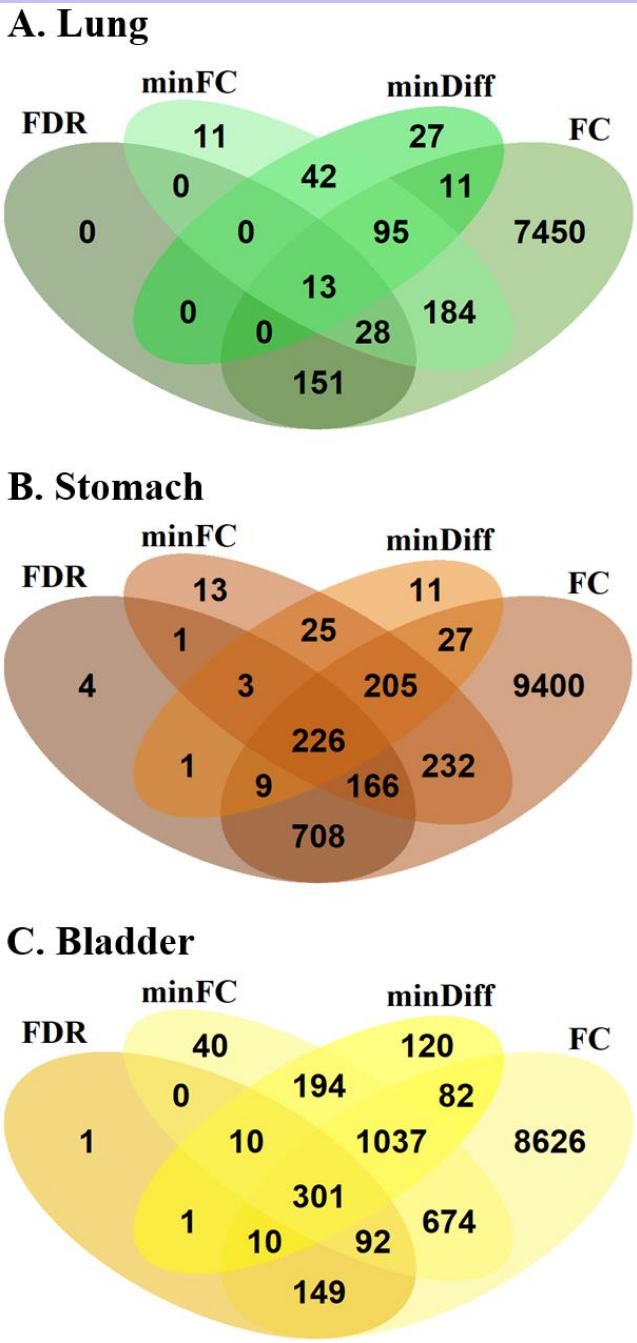


Figure 15. Venn diagram presenting a comparison of lists of DEGs selected with the standard method: with q value < 0.050 (FDR), or with log2 fold change (FC) < 0.500, or with applying single measure *minFC* or *minDiff* threshold in (A) lung, (B) stomach, and (C) bladder data set.

By selecting a candidate biomarker while maintaining the conditions introduced in the PREDICT pipeline, it is possible to identify potential candidates that may have the appropriate level of specificity required for predictive biomarkers. With this pipeline by sequentially applying thresholds of log2FC > 0.500, q value < 0.050, log2*minFC* > 0.100, and *minDiff* > 100 to the results obtained from the standard differential analysis method, it led to filtering out the unwanted results, and the numbers of DEGs obtained by this method are presented in Figure 16.

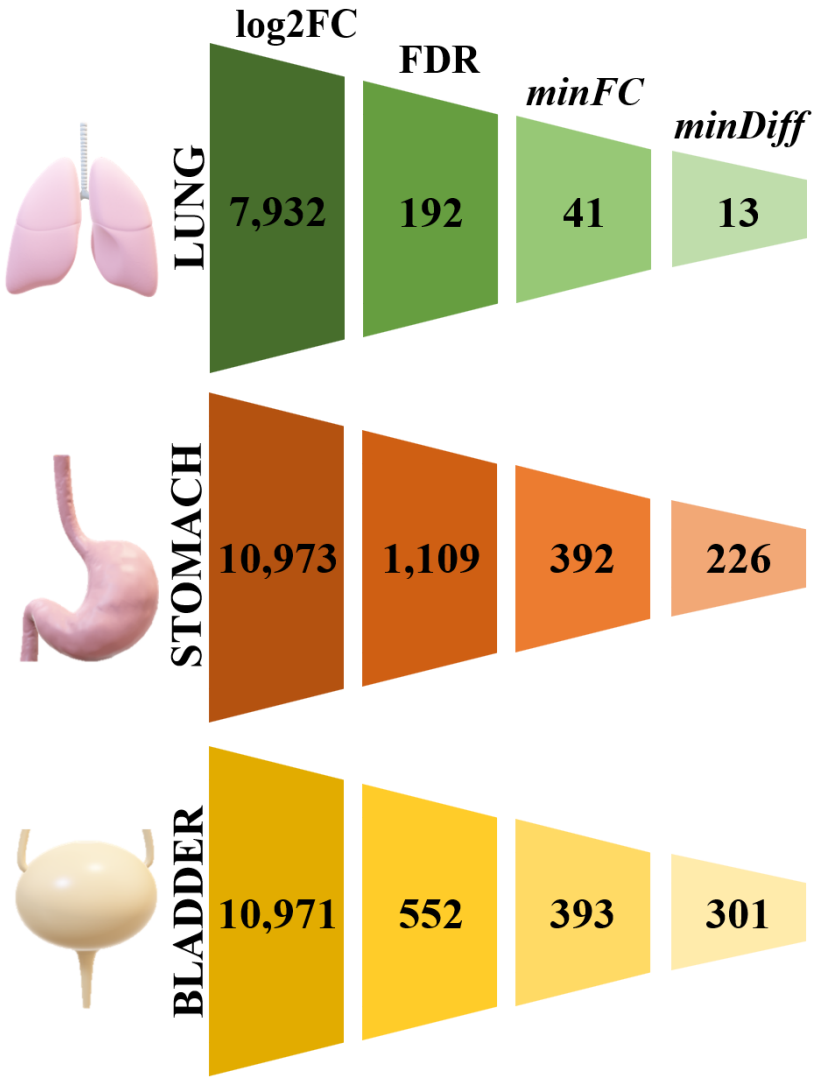
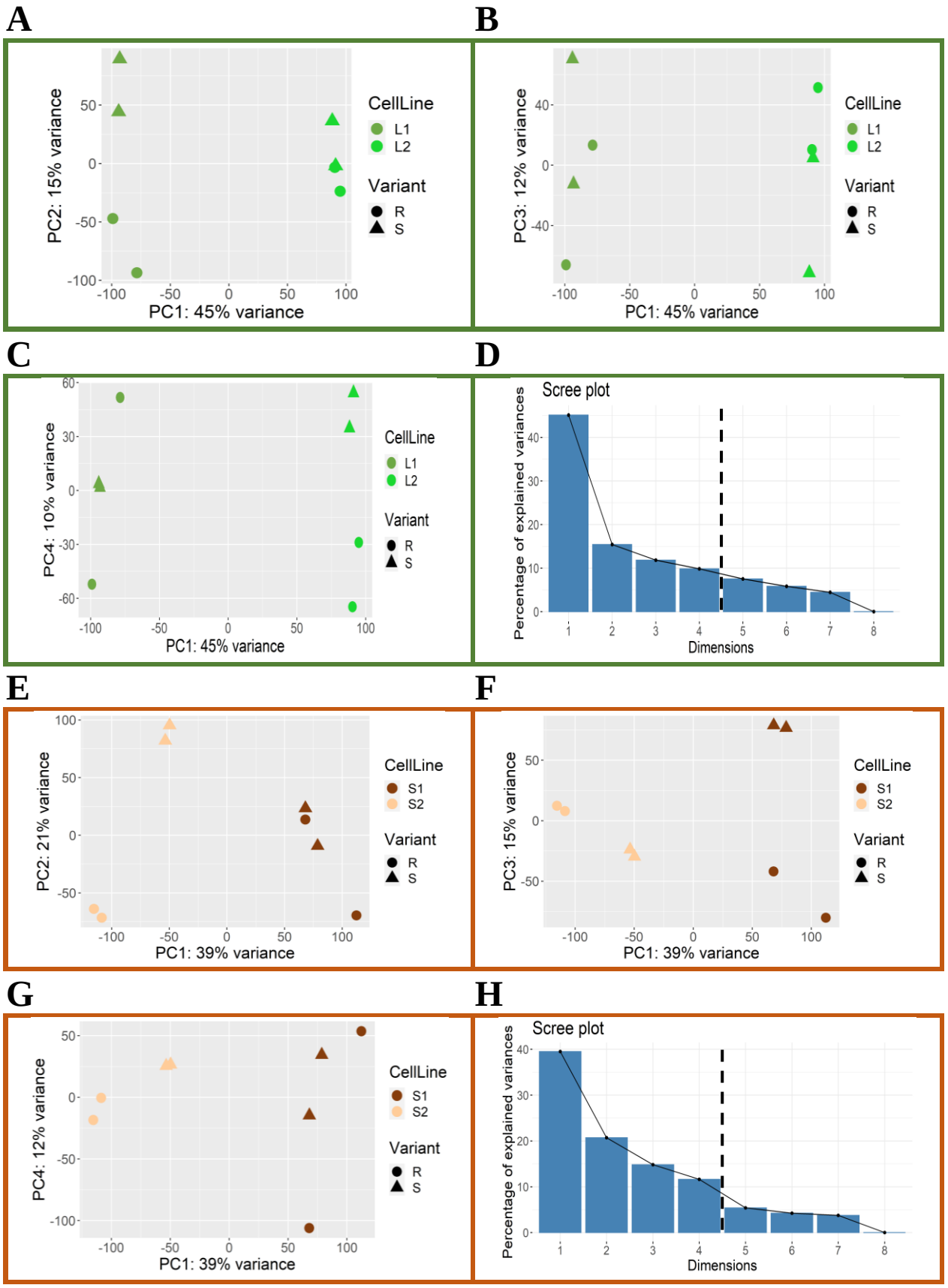


Figure 16. Numbers of identified DEGs obtained by sequentially applying measures thresholds: log2 fold change (log2FC) < 0.500, q value (FDR) < 0.050, log2*minFC* > 0.100, and *minDiff* > 100.

To evaluate inter-tumor diversity, Principal Component Analysis (PCA) (Figure 17) was applied to the regularized logarithm (rlog) transformed data. Based on this unsupervised analysis, we can see that the main sources of variability in the data are related to the cell lines, which mimic the inter-tumor diversity in my experimental design. The lung cancer cell lines exhibit the greatest difference, while the bladder cancer cell lines exhibit the smallest difference (Figure 17). This diversity and low sample size are the main reasons why so many DEGs lack proper biomarker features. By applying the PREDICT pipeline, we can address these imperfections, and select more suitable candidates for biomarkers. Additionally, by removing candidates with a lower probability of entering clinical practice, we reduce the number of potential candidates for the validation phase (described more in Chapter VI.). As this phase is costly, it is easier to decide which candidates to consider when the number of potential biomarkers is smaller and more suitable.


A
PC2: 15% variance
PC1: 45% variance
CellLine
L1
L2
Variant
R
S
B
PC3: 12% variance
PC1: 45% variance
CellLine
L1
L2
Variant
R
S
C
PC4: 10% variance
PC1: 45% variance
CellLine
L1
L2
Variant
R
S
D
Scree plot
Percentage of explained variances
Dimensions
E
PC2: 21% variance
PC1: 39% variance
CellLine
S1
S2
Variant
R
S
F
PC3: 15% variance
PC1: 39% variance
CellLine
S1
S2
Variant
R
S
G
PC4: 12% variance
PC1: 39% variance
CellLine
S1
S2
Variant
R
S
H
Scree plot
Percentage of explained variances
Dimensions

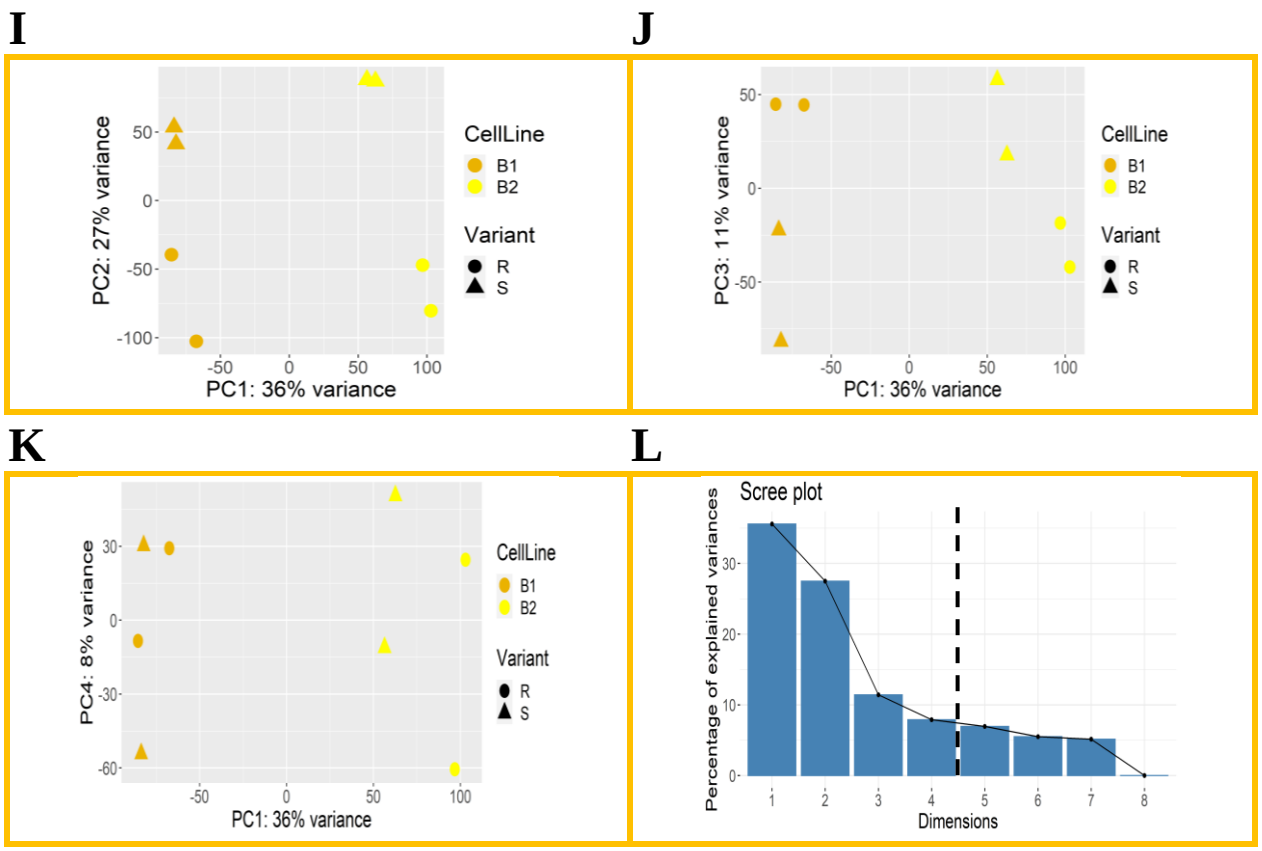


Figure 17. Principal Component Analysis (PCA) was carried out on three data sets: lung (A, B, C, D), stomach (E, F, G, H), and bladder (I, J, K, L). For each analysis, there is a scree plot (lung: D, stomach: H, and bladder: L) representing the percentage contribution of principal components (PC) in explaining variability in the data. The dashed line on this plot marks a threshold of ~85 % of variability up to which I was assessing the acquired PC. There is four PC for each dataset, and they are presented on three plots PC1 versus PC2 or PC3, or PC4 accordingly (lung (A, B, C), stomach (E, F, G), and bladder (I, J, K)).

To assess whether the observed inter-tumor diversity in the PCA analysis presented in Figure 17 is influenced by batch effects, I conducted a Principal Component Analysis using 15 housekeeping genes (Table 11) as a control. This allowed me to distinguish between true biological variability and experimental artifacts.

Table 11. Housekeeping genes [257, 258].

Gene ID	Gene symbol	Gene name
ENSG00000111640	*GAPDH*	Glyceraldehyde-3-phosphate dehydrogenase
ENSG00000075624	*ACTB*	Beta-actin (β-actin)
ENSG00000273686	*B2M*	Beta-2-microglobulin
ENSG00000112339	*HBS1L*	HBS1-like protein
ENSG00000165704	*HPRT1*	Hypoxanthine guanine phosphoribosyl transferase I
ENSG00000073578	*SDHA*	Succinate dehydrogenase complex, subunit A, flavoprotein
ENSG00000196565	*HBG2*	Gamma globin (γ-globin)
ENSG00000184009	*ACTG1*	Actin Gamma 1
ENSG00000227794	*RPS18*	Ribosomal Protein S18
ENSG00000272391	*POM121C*	Nuclear Pore Membrane Protein 121-2
ENSG00000112110	*MRPL18*	Mitochondrial Ribosomal Protein L18
ENSG00000175768	*TOMM5*	Translocase Of Outer Mitochondrial Membrane 5
ENSG00000149658	*YTHDF1*	Dermatomyositis Associated With Cancer Putative Autoantigen 1
ENSG00000133112	*TPT1*	Tumor Protein, Translationally-Controlled 1
ENSG00000177954	*RPS27*	Ribosomal Protein S27

With the results of this PCA analysis (Figure 18) based on the expression of housekeeping genes (Table 11), we can see that there is no significant difference between cell lines thus the observed diversity is not related to the batch effect.

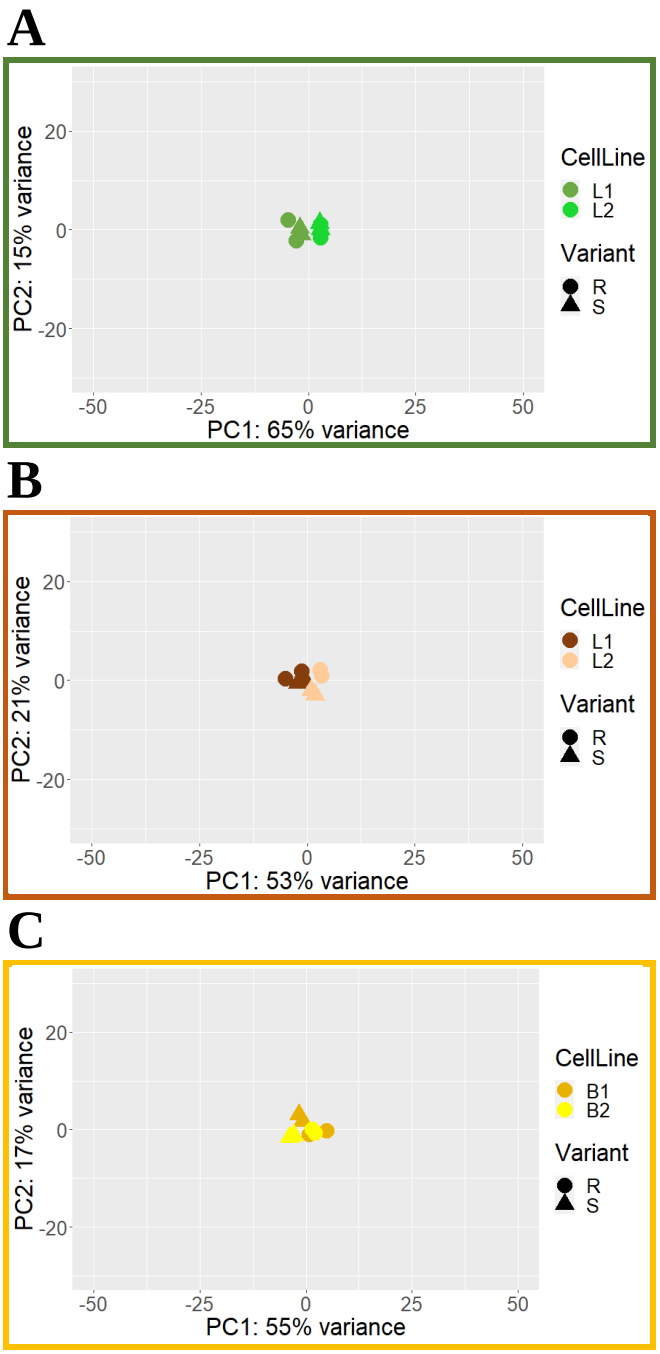


Figure 18. Principal Component Analysis (PCA) was carried out on three data sets of housekeeping genes (Table 11) from lung (A), stomach (B), and bladder (C).

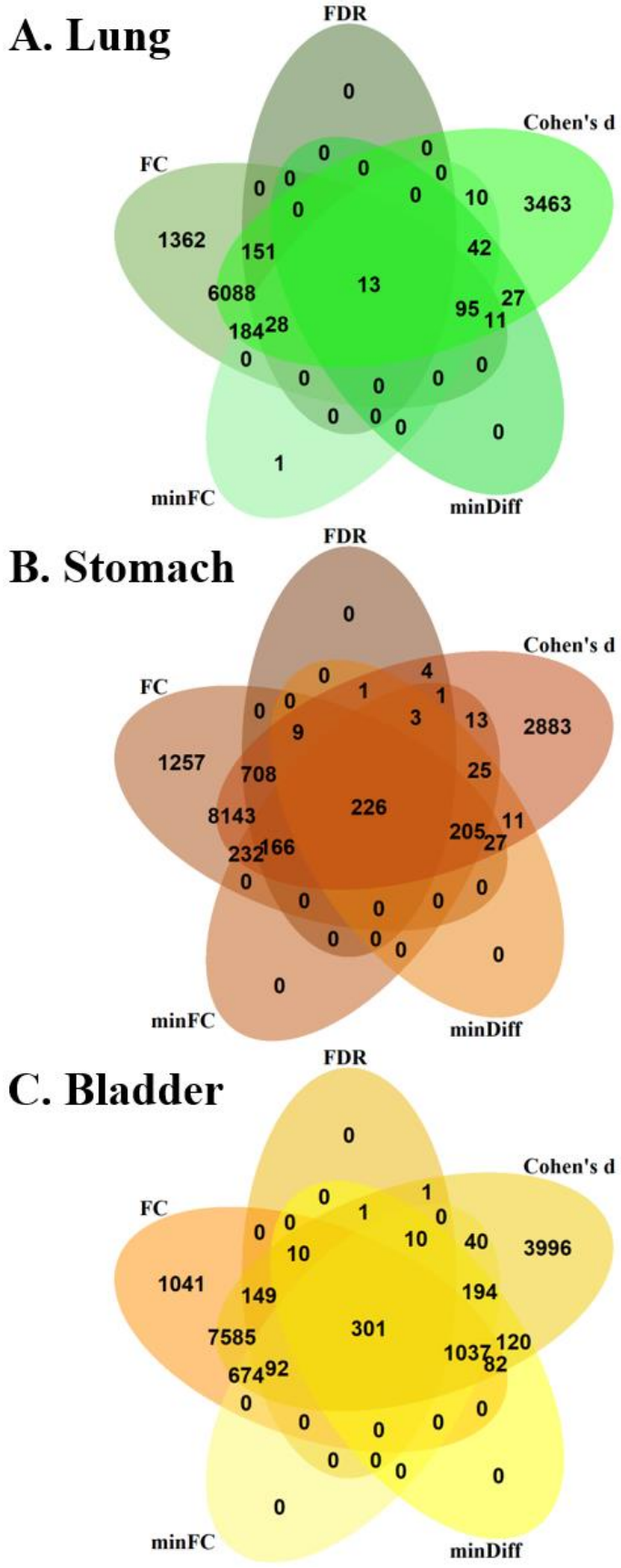


Figure 19. Venn diagram presenting a comparison of lists of DEGs selected with the standard method: with q value < 0.050 (FDR), or with log2 fold change (FC) < 0.500, or with Cohen's d > 0.300, or with applying single measure minFC or minDiff threshold in (A) lung, (B) stomach, and (C) bladder data set.

The fold change (FC) is a commonly used approach to detect differential gene expression, but it has limitations due to the lack of consideration of the uncertainty associated with gene expression measurements in different conditions. This is especially true for genes with low read counts, which tend to have less reliable fold changes than those with high read counts. Furthermore, it is not appropriate to compare fold changes based on read counts for genes that vary in expression levels or have different lengths [195, 259, 260]. Due to the limitations of the FC measure, I decided to assess whether replacing this measure with a more recommended measure of effect size, such as Cohen's d, could significantly improve the results obtained using the PREDICT pipeline. Cohen's d differs from FC in measuring effect size, such as by indicating that it takes into account the variance of each condition, which can provide a more robust measure of differential expression.

When assessing the Venn diagram (Figure 19) in terms of replacing the FC measure with the Cohen's d measure in the PREDICT pipeline, we can see that for the lung dataset, the number of selected DEGs remains the same and for the stomach and bladder datasets, there would be an increase of 3 and 10 DEGs, respectively. Therefore, the replacement of FC with Cohen's d would not significantly alter the results. Moreover, FC is a widely used measure and in many cases preferred by scientists in the different fields of biology as well as it has utility for various downstream analysis tasks, such as prioritizing genes for further investigation and linking FC with other variables of interest. For all these reasons, I have decided not to replace FC in the PREDICT pipeline.

Additionally, we can see that by using Cohen's d instead of FC measure we would still receive some undesirable results (DEGs), thus *minFC* and *minDiff* measures are necessary to filter out all incorrect results.

XII. Assessment of biological context of genes selected with PREDICT

XII.1. Three generations of pathway analysis

In 2012, Khatri et al. propose a classification system for pathway analysis methods, which includes over-representation analysis (First Generation), functional class scoring (Second Generation), and pathway topology-based analysis (Third Generation). The classification is based on the specific type of analysis each method performs, and there are significant differences in the input datasets and computational analyses used between the different classes [261].

First Generation

In the first generation of pathway analysis methods, Over-Representation Analysis (ORA) is based on the hypothesis that relevant pathways can be identified by detecting an over-representation of differentially expressed genes within a given pathway compared to what would be expected by chance. ORA methods evaluate the fraction of pathway components found within a user-selected list of biological components based on certain criteria such as log fold change and statistical significance. Statistical methods such as the hypergeometric distribution, chi-square, binomial probability, or Fisher's exact test are used to calculate a confidence value and rank pathways. Multiple testing correction is performed to avoid false positives. The output from ORA methods is generally a list of relevant pathways ordered by their P value or multiple-hypothesis-test-corrected P value. The advantage of ORA methodologies is that they provide Omic data with biological context, facilitating hypothesis generation and subsequent experimental testing, which is in line with the Systems Biology approach [261].

ORA methods have some limitations, despite their ability to quickly identify significant biological meaning from large datasets. First, these methods often omit potentially important components near the selected cut-off threshold, leading to reduced information and unstable results. Additionally, cut-off thresholds are arbitrarily selected, making it difficult to establish a standard threshold [262]. Second, ORA methods treat all components in a pathway equally, disregarding any inherent information about interactions between genes such as gene expression level or position in the pathway [263]. Therefore, two pathways with the same genes but different topologies would yield the same result. Finally, ORA methods assume that pathways are independent of each other, ignoring the fact that pathways can interact and overlap with each other [264].

Second Generation

Functional Class Scoring (FCS) methods are the second generation of pathway analysis (PA) and operate under the hypothesis that even small but coordinated changes in gene expression can significantly affect a pathway's overall state. Unlike ORA methods, FCS methods use all available measurements in high-throughput biological data (HTBD) and do not have a cut-off threshold limitation. FCS methods still use pathways as gene sets to perform their computations. The workflow for FCS methods involves three steps: (1) calculating a basal-level statistic using all HTBD to compute differential expressions of individual components; (2) aggregating basal-level statistics from each pathway's components into a single pathway-level statistic; and (3) calculating the statistical significance of pathway-level statistics using methods such as the Kolmogorov-Smirnov statistic, Wilcoxon sum rank statistic, max-mean statistic, or chi-squared test [261].

FCS methods offer several advantages over ORA methods [261]:

- FCS methods do not require an arbitrary cut-off threshold for differentially expressed genes because they utilize all available information.
- They can identify variances between pathways that are only marginally passing the differentially expressed thresholds and those that are significantly passing them.
- FCS methods can identify subtle but coordinated relations between gene-expression levels of molecules and their associated pathways.
- Some FCS methods can detect the most significant genes in any given pathway, known as the core of the pathway, such as in the case of GSEA.

Gene Set Enrichment Analysis (GSEA) is a widely used FCS method for analyzing gene expression data. The approach involves ranking genes based on their differential expression between two phenotypic classes, using the signal-to-noise ratio basal-level statistic. The ranked gene list is then evaluated against a pre-defined set of genes (e.g. gene sets from the MSigDB) to calculate an enrichment score (ES) for each gene set, using a Kolmogorov-Smirnov pathway-level statistic. Finally, the significance of the ES is determined, and adjustments are made for multiple hypothesis testing [261].

Although FCS methods address some of the limitations of ORA analyses, they still have some drawbacks, primarily because they use pathways as gene sets instead of networks. Some of these limitations include [261]:

- Many FCS methods still assume that all components within a pathway have equal weight when determining the pathway statistic, regardless of prior knowledge about the pathway. This can result in inaccurate pathway rankings.

- These methods often do not consider the relationships between components within pathways, leading to an underutilization of information from Pathway Databases (PDBs) and a decreased ability to identify important pathways.

- FCS methods still analyze pathways independently of one another and do not consider the overlap between pathways or the potential influence that one pathway may have on another.

Third Generation

Advancements in pathway annotation from Pathway Databases have led to the integration of pathway topology into pathway analysis (PA) methodologies, giving rise to the pathway topology based (PTB) analysis. PTB analysis assumes that interactions in pathway topology contain information for interpreting correlated changes between pathway components. PTB methods extend ORA and FCS methods, by adding pathway topology for assessing the statistical relevance of the pathways. For ORA extended methods, user-selected genes are mapped onto the pathway topology, followed by network and statistical analyses. For FCS extended methods, the HTBD and topology are used to calculate basal-level statistics and continue further in an analogous way to an FCS approach [261].

Pathway topology-based (PTB) analysis can overcome the limitations of ORA and FCS methods in several ways [261]:

- By considering the topology of pathways, PTB methods can assign different weights to pathway components based on their biological significance, resulting in more accurate pathway statistics.

- PTB methods can also incorporate diverse topological information to analyze the same set of pathway components, providing a more refined analysis.

- PTB methods consider causal interactions within pathways, allowing for analysis of upstream and downstream effects, and identifying potential key drivers of pathway behavior.

PTB methodologies have limitations that are difficult to address as they challenge the current paradigm in life sciences, which recognizes that life components work together as a dynamic, adaptable, and robust system. Despite this, some limitations can be identified that will likely be addressed in future methods as experimental and annotation barriers are overcome. These limitations include [261]:

- some PTB methods do not consider the direction of associations among pathway components, potentially missing chain effects of deregulation,
- PTB methods do not account for interconnections between pathways, which may result in a failure to detect relevant pathways,
- PTB methods do not consider the time and spatial distribution of pathway components, which may be dependent on biomolecule compartmentalization,
- moreover, molecular regulation in a time-scale manner is also important for understanding pathway mechanisms, and
- most methods cannot distinguish the multiple states and variants that a pathway element can have, which may affect phenotype and pathway functioning,
- the sample size must be sufficient to apply PTB methods.

XII.2. Biological context assessment

As mentioned in earlier chapters the main limitation of the experimental design covered in my dissertation is related to the small sample size. Therefore, in this case analysis using third-generation Pathway-Topology Based methods is not recommended [265, 266]. An attempt to conduct such an analysis was made using the hiPathia tool v 1.7.4 (HIgh throughput PATHway Interpretation and Analysis) [267], but no significant results were obtained. Based on literature data [265, 266], a sample size of 10-15 repetitions per group would be needed for this type of analysis. In my case, there are only four repetitions per group. Therefore, due to the nature of the experimental design of the data used as an example in my doctoral dissertation, the assessment of biological context was performed using first and second generation methods.

Enrichment analysis was performed against Reactome pathways using the ReactomePA package v 1.16.2 [268]. Two methods were employed: over-representation analysis (ORA) with the gene list selected with the PREDICT pipeline, and gene set enrichment analysis (GSEA) with genes ranked according to the Wald test statistic. P values were adjusted for multiple testing with Benjamini-Hochberg correction (q value).

Lung

Using the ORA method on genes selected with the PREDICT pipeline for the lung DEGs data set, I have identified 37 significant pathways (q value < 0.050; see Table 12). Of these 37 pathways, 9 may suggest a potential association with the mechanism of resistance to FGFR inhibitor (highlighted in blue in Table 12). Signaling pathways such as „ERK/MAPK targets", „MAPK targets/ Nuclear events mediated by MAP kinases", „Gastrin-CREB signaling pathway via PKC and MAPK", and „TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation" may indicate a potential resistance mechanism in the form of compensatory action. Since FGFR inhibition can result in the inhibition of proliferation (described more in chapters II. and IV.2.), a resistant cell may compensate for this event by activating the MAPK (Mitogen-activated protein kinase) signaling pathway, which stimulates proliferation, among other [269]. Similarly, the increased activity of signaling pathways associated with RAS (Rat sarcoma virus) in my case indicated as the pathway of "CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling" can be explained in the same way [270]. The "Glucuronidation" signaling pathway is associated with drug metabolism, so its increased activity may be related to a resistant cell's ability to cope with the presence of an FGFR-TKI [271]. The role of signaling pathways

associated with RND1 (Rho Family GTPase 1; our indicated pathway being "RND1 GTPase cycle") is poorly understood. My analysis indicates a change in the activity of this signaling pathway due to altered regulation of the GRB7 (Growth Factor Receptor Bound Protein 7) protein. Most reports suggest that the GRB7 protein is associated with increased proliferation, among others [272-274]. However, there are also reports indicating inhibitory action in interaction with the RND1 protein [275]. In our case, the *GRB7* gene, which was identified as downregulated in lung cancer cell lines resistant to FGFR inhibitor, may suggest that the lack of its interaction with the signaling pathway associated with RND1 reduces its activity, and in this way, we can observe a mechanism of compensatory activation of proliferation inhibited by FGFR inhibitor [276, 277].

Table 12. The list of significant pathways from ORA analysis performed based on the lung gene set selected with the PREDICT pipeline.

No *	Description *	Gene count *	qvalue *
1	Toll Like Receptor 10 (TLR10) Cascade	2	0.013
2	Toll Like Receptor 5 (TLR5) Cascade	2	0.013
3	MyD88 cascade initiated on plasma membrane	2	0.013
4	TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	2	0.013
5	MyD88 dependent cascade initiated on endosome	2	0.013
6	Toll Like Receptor 7/8 (TLR7/8) Cascade	2	0.013
7	Toll Like Receptor 9 (TLR9) Cascade	2	0.013
8	MyD88:MAL(TIRAP) cascade initiated on plasma membrane	2	0.013
9	Toll Like Receptor TLR6:TLR2 Cascade	2	0.013
10	Toll Like Receptor TLR1:TLR2 Cascade	2	0.013
11	Toll Like Receptor 2 (TLR2) Cascade	2	0.013
12	Toll Like Receptor 4 (TLR4) Cascade	2	0.019
13	Toll-like Receptor Cascades	2	0.024
14	Metallothioneins bind metals	1	0.029
15	Response to metal ions	1	0.032
16	TRAF6 mediated IRF7 activation in TLR7/8 or 9 signaling	1	0.032

17	Tie2 Signaling	1	0.034
18	IRAK4 deficiency (TLR2/4)	1	0.034
19	Gastrin-CREB signalling pathway via PKC and MAPK	1	0.034
20	ERK/MAPK targets	1	0.040
21	Telomere Extension By Telomerase	1	0.040
22	Glucuronidation	1	0.040
23	WNT ligand biogenesis and trafficking	1	0.040
24	Downstream signal transduction	1	0.040
25	CREB1 phosphorylation through NMDA receptor-mediated activation of RAS signaling	1	0.040
26	MAPK targets/ Nuclear events mediated by MAP kinases	1	0.040
27	Diseases of Immune System	1	0.040
28	Diseases associated with the TLR signaling cascade	1	0.040
29	Metabolic disorders of biological oxidation enzymes	1	0.042
30	Synthesis of PA	1	0.047
31	RET signaling	1	0.047
32	RND1 GTPase cycle	1	0.047
33	Signaling by SCF-KIT	1	0.047
34	Transport of vitamins, nucleosides, and related molecules	1	0.047
35	Recycling pathway of L1	1	0.048
36	Signaling by ERBB2	1	0.050
37	Extension of Telomeres	1	0.050

***** blue color indicates pathway potential association with the mechanism of resistance to FGFR inhibitor.

A major limitation of this ORA analysis is the very small number of genes used, which is only 13. Only this many genes could be selected when applying the PREDICTpipeline. With such a small number of genes, the appearance of even one gene in the signaling pathway from the analyzed list can indicate the pathway as statistically significant. Therefore, in this particular situation, it is more beneficial to evaluate the available literature on these genes to assess their biological

context. Based on such an evaluation, using the MEDLINE (through PubMed) and Scopus repositories, as well as GeneCards, I associated 11 out of 13 genes with a potential mechanism of resistance to anti-FGFR treatment. For two genes, accounting for their direction of change, I could not link the protein's role to a potential resistance mechanism based on the available literature. This may be due to the different protein's tissue-specific activity (moonlighting proteins) [278, 279]. Taking as an example from the list of genes selected using the PREDICT pipeline (Table 13), the protein GRB7 (as described above) may indirectly stimulate proliferation, and in my case, its downregulation would indicate the opposite role in lung cancer. Therefore, we can assume that the role of the GRB7 protein may be tissue-specific. Another example of a moonlighting protein in my analysis is the WLS (Wnt Ligand Secretion Mediator) protein (Table 13), which interacts with the Wnt proteins and indirectly stimulates proliferation [280-283]. On the other hand, according to Yang et al. [284], in melanoma, overexpression of WLS inhibits cell proliferation. Based on the literature review conducted for the mentioned 13 genes (Table 13), the proteins MT2A [285], GRB7 [275], CDC14B [286], RPS6KA3 [287], IRAK4 [288], ZCCHC14 [289], WLS [284], ANKRD28 [290], and LCLAT1 [291] may participate in a compensatory mechanism of proliferation induction. The SLC35D1 (Solute Carrier Family 35 Member D1) protein is associated with the glucuronidation mechanism, which as mentioned above, is related to drug metabolism. Therefore, the cells may cope with the presence of the FGFR inhibitor by metabolizing it.

Table 13. The list of the lung genes selected with the PREDICT pipeline.

No *	Gene symbol *	Gene name *	Potential mechanism of resistance to FGFR inhibitor
1	***MT2A***	Metallothionein 2A	Compensation mechanism: downregulation promotes proliferation and migration [285].
2	***GRB7***	Growth Factor Receptor Bound Protein 7	Compensation mechanism: with downregulation of GRB7 there is no proliferation inhibitory effect mediated by interaction with Rnd1 [275].
3	***CDC14B***	Cell Division Cycle 14B	Compensation mechanism: by role in cell cycle control affects cell proliferation [286].
4	***CFAP36***	Cilia And Flagella Associated Protein 36	Compensation mechanism: CFAP36 directly interacts with ARL3, causing it to be blocked from its membrane localization and preventing ARL3 from performing its function. ARL3 is most likely a tumor suppressor by inducing into tumor immune cell infiltration. Therefore, by blocking ARL3, CFAP36 promotes cell survival [292, 293].

5	***RPS6KA3***	Ribosomal Protein S6 Kinase A3	Compensation mechanism: by interacting with MAPK promotes proliferation and migration [287].
6	***IRAK4***	Interleukin 1 Receptor Associated Kinase 4	Compensation mechanism: by interacting with NF-kappaB promotes proliferation [288].
7	***SLC35D1***	Solute Carrier Family 35 Member D1	Drug metabolism: participate in the glucuronidation process that is involved in the metabolism of drugs [271].
8	***CDCA7L***	Cell Division Cycle Associated 7 Like	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that CDCA7L promotes cell proliferation, and inhibited cell apoptosis [294].
9	***ISOC2***	Isochorismatase Domain Containing 2	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that ISOC2 inhibits the expression of p16(INK4a). p16(INK4a) is a multiple tumor suppressor, playing an important role in proliferation and tumorigenesis [295].
10	***ZCCHC14***	Zinc Finger CCHC-Type Containing 14	Compensation mechanism: downregulation promotes proliferation [289].
11	***WLS***	Wnt Ligand Secretion Mediator	Compensation mechanism: downregulation promotes proliferation [284].
12	***ANKRD28***	Ankyrin Repeat Domain 28	Compensation mechanism: downregulation leads to a decrease in the inhibition of NFKBIE dephosphorylation, resulting in its direction towards ubiquitination pathway. As a result, the NFkB signaling pathway is not inhibited, which leads to the stimulation of proliferation [290].
13	***LCLAT1***	Lysocardiolipin Acyltransferase 1	Compensation mechanism: by interacting with EGF and AKT promotes proliferation [291].

***** blue and red color indicates that this gene is downregulated or upregulated respectively.

From the second-generation pathway analysis methods, the GSEA (Gene Set Enrichment Analysis) was selected as it is the most popular and can be used for small sample size data [296]. For the GSEA analysis, we use the entire list of genes ranked accordingly (in my case, genes ranked according to the Wald test statistic), so we no longer have information from the PREDICT pipeline. As a result of the GSEA analysis, we identified three statistically significant pathways presented in Table 14. Two of these pathways, "Metallothioneins bind metals" and "Response to metal ions," are consistent with the results of the ORA analysis. The protein MT2A, which gene expression is downregulated in my case, is associated with these pathways and has been docu-

mented by Lui et al. [285] to inhibit proliferation when overexpressed. Therefore, it can be assumed that its downregulation is associated with decreased activity of these signaling pathways and thus linked to the mechanism of resistance to FGFR inhibitor therapy through compensatory activation of proliferation.

Table 14. The list of significant pathways from GSEA analysis performed based on the lung DEGs data set.

No *	Description *	Gene count *	qvalue *
1	Metallothioneins bind metals	6	0.018
2	Response to metal ions	6	0.018
3	Keratinization	16	0.040

***** blue color indicates pathway potential association with the mechanism of resistance to FGFR inhibitor.

Stomach

Based on the data set selected with the PREDICT pipeline no significant pathways were identified with the over-representation analysis method. However, using the gene set enrichment analysis I have identified 220 significant pathways (q value < 0.050) (the list of significant pathways from GSEA analysis can be found in Table S3 in Supplementary Materials placed on CD-R attached at the back of the dissertation). Since with a closer look, it was possible to distinguish groups of similar pathways, thus using the enrichplot tool v 1.13.1.994 [297], pathways were clustered (an arbitrary number of clusters was chosen) based on the Jaccard similarity coefficient. There was possible to distinguish 10 different groups of pathways (Figure 20), specifically related to: i) nuclear transport, ii) viral infection, iii) defective HDR, iv) DNA repair, v) m phase, kinetochores, vi) mismatch repair, telomeres, vii) cell cycle, viii) mitochondrial translation, ix) mRNA, and x) extracellular matrix organization.

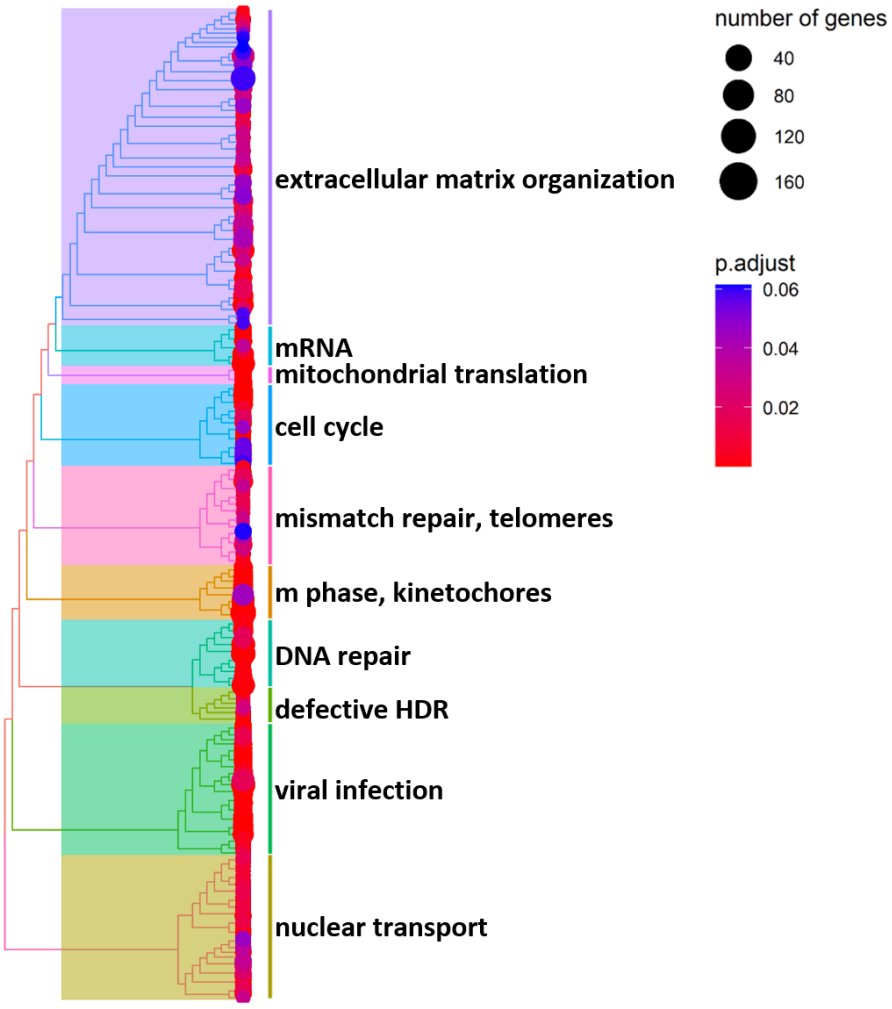


Figure 20. Hierarchical clustering of statistically significant (q value < 0.05) pathways identified with GSEA analysis performed based on the stomach DEGs data set. The number of the stomach DEGs identified in the pathways is represented using the size of the circles as indicated in the right panel.

Since FGFR indirectly participate in regulating proliferation, and cell survival as well as maintaining tissue homeostasis, thus potential FGFR-TKIs resistant mechanism could lie in v, vi, vii, and x cluster (Figure 20). Therefore, in the gene set that came out related to these pathways I have selected genes that meet PREDICT statistical properties. As there were 57 identified genes I have assessed them based on available literature (Table 15) similar to the case of lung genes described above.

With the performed literature assessment I was able to connect 44 out of 57 genes with a potential mechanism of FGFR-TKIs resistance. Most genes might be related to a sort of compensatory activation of proliferation (32 genes), migration (15 genes), and invasion (5 genes) as a response to inhibition of these processes caused by blocking of FGF receptor/s signaling (Table 15). This is potentially mainly mediated through MAPK and AKT signaling pathways but also like in the case of RHEB (Ras Homolog, MTORC1 Binding) protein by interacting with mTORC1 pathway [298] or in the case of APOL1 (Apolipoprotein L1) protein by activating NOTCH1 (Neurogenic locus notch homolog protein 1) signaling pathway [299]. 18 genes could be related to the antiapoptotic mechanism (Table 15). Like for example in the case of IQGAP2 (IQ Motif Containing GTPase Activating Protein 2) protein, Song et al. [300] report that it promotes apoptosis via activating the p38-p53 pathway, triggered by an increase in reactive oxygen species (ROS). Downregulation of this gene in my case may suggest a potential antiapoptotic protective mechanism. The rest of the genes are briefly described in Table 15 and more information can be found under provided citations.

Table 15. The list of genes selected with the PREDICT pipeline from the stomach DEGs data set that came out related to the v or vi or vii or x cluster of pathways from GSEA analysis.

No *	Gene symbol *	Gene name *	Potential mechanism of resistance to FGFR inhibitor
1	***TOM1***	Target Of Myb1 Membrane Trafficking Protein	Antiapoptotic mechanism: act as a negative regulator of IL-1beta and TNF-alpha-induced signaling pathways [301].
2	***IFIT3***	Interferon Induced Protein With Tetratricopeptide Repeats 3	Antiapoptotic mechanism: negative regulation of apoptotic process [302].
3	***IFIT2***	Interferon Induced Protein With Tetratricopeptide Repeats 2	Antiapoptotic mechanism: negative regulation of apoptotic process [302].
4	***SSRP1***	Structure Specific Recognition Protein 1	Antiapoptotic mechanism: when it is not downregulated like in our case it may potentiate cisplatin-induced cell death by blocking replication and repair of modified DNA [303].

5	***IFIT1***	Interferon Induced Protein With Tetratricopeptide Repeats 1	Antiapoptotic mechanism: negative regulation of apoptotic process [302].
6	***TFDP1***	Transcription Factor Dp-1	Antiapoptotic mechanism: downregulation prevents from activated E2F mediated apoptosis [304, 305].
7	***CLCN7***	Chloride Voltage-Gated Channel 7	Antiautophagy mechanism: CLCN7 promotes autophagy resistance through TGFβ signaling [306].
8	***APOL1***	Apolipoprotein L1	Compensation and antiapoptotic mechanism: promotes proliferation and inhibits apoptosis via activating NOTCH1 signaling pathway [299].
9	***RHEB***	Ras Homolog, MTORC1 Binding	Compensation and antiapoptotic mechanism: activation of cell growth through stimulation of mTORC1 activity, protecting cancer cells from apoptosis induced by metabolic or oxidative stress by reciprocal regulation between Rheb and AMPK [307].
10	***UBC***	Ubiquitin C	Compensation and antiapoptotic mechanism: promotes cell growth, cell survival, and antiapoptotic effects [308].
11	***HSPB8***	Heat Shock Protein Family B (Small) Member 8	Compensation and antiapoptotic mechanism: promotes cell growth, cell survival, antiapoptotic effects, and autophagy pro-survival [309].
12	***BATF***	Basic Leucine Zipper ATF-Like Transcription Factor	Compensation and antiapoptotic mechanism: promotes cell growth, cell survival, and antiapoptotic effects [310, 311].
13	***IL17RC***	Interleukin 17 Receptor C	Compensation and antiapoptotic mechanism: promotes cell growth, cell survival, and antiapoptotic effects by activation of NF-kappa-B and MAPkinase pathways, and protecting from TNFα-induced apoptosis [312].
14	***C3***	Complement C3	Compensation and antiapoptotic mechanism: promotes cell growth, migration, cell survival, and antiapoptotic effects via activation of the PLC, MAPK and AKT signaling pathways [313-315].
15	***IQGAP2***	IQ Motif Containing GTPase Activating Protein 2	Compensation and antiapoptotic mechanism: IQGAP2 acts as a tumor-suppressor where with its downregulation it promotes cell growth, migration, cell survival and inhibit apoptosis by not disrupting signaling pathways, such as the MAPK/ERK, receptor tyrosine kinase (RTK)-activated phosphatidylinositol 3-kinase/AKT (PI3K-AKT), transforming growth factor β (TGF-β), and Wnt/-catenin pathways [300].
16	***CD55***	CD55 Molecule (Cromer Blood Group)	Compensation and antiapoptotic mechanism: promotes cell growth, metastasis, and antiapoptotic effects [316].
17	***LAMTOR2***	Late Endosomal/Lysosomal Adaptor, MAPK And MTOR Activator 2	Compensation mechanism: promotes cell growth, migration, and cell survival as LAMTOR2 is an adapter protein that enhances the efficiency of the MAP kinase cascade facilitating the activation of MAPK2 [317].

18	***PDS5B***	PDS5 Cohesin Associated Factor B	Compensation mechanism: This protein is a negative regulator of cell proliferation, migration, and invasion via upregulation of LATS1 [318].
19	***LGALS1***	Galectin 1	Compensation mechanism: promotes cell growth, cell migration, and cell invasion by interacting with MAPK and MMP-9 path ways [319].
20	***ULK1***	Unc-51 Like Autophagy Activating Kinase 1	Compensation mechanism: promotes cell survival, migration, and invasion as ULK1 is involved in autophagy in response to starvation. It acts as a upstream of phosphatidylinositol 3-kinase PIK3C3 to regulate the formation of autophagophores. Since the dual role of autophagy in cancers, ULK1 can promote cancer development and increase the survival, invasion, and metastasis of tumor cells [320].
21	***CTSA***	Cathepsin A	Compensation mechanism: play an important role in the growth and metastasis by promoting proliferation and migration [321].
22	***DDX58***	DExH-Box Helicase 58	Compensation mechanism: DDX58 expression is significantly associated with immune cell infiltration. Evidence suggests that monocytes/macrophages are involved in tumor growth, metastasis, and tumor vascularization by regulating the tumor microenvironment [322].
23	***MTA2***	Metastasis Associated 1 Family Member 2	Compensation mechanism: downregulation promotes proliferation and migration [285].
24	***PSAP***	Prosaposin	Compensation mechanism: PSAP promotes progression by decreasing tumor-infiltrating lymphocytes [323] and also is involved in proliferation, tumorigenesis, and metastasis. PSAP regulates the invasion and migration through the TGF-β1/Smad signaling pathway [324].
25	***CDC37L1***	Cell Division Cycle 37 Like 1, HSP90 Cochaperone	Compensation mechanism: promotes cell growth, and cell survival by interacting with Hsp90 [325].
26	***MCM3***	Minichromosome Maintenance Complex Component 3	Compensation mechanism: This protein is a negative regulator of cell proliferation, by directly binding to the HIF-1α subunit and synergistically inhibit HIF-1 transcriptional activity via distinct O2-dependent mechanisms [326].
27	***RAB9A***	RAB9A, Member RAS Oncogene Family	Compensation mechanism: promotes cell growth by activating the AKT/mTOR signaling pathway [327].
28	***TERF2***	Telomeric Repeat Binding Factor 2	Compensation mechanism: its downregulation promotes cell growth by not inhibiting EGFR [328].
29	***ATP6V1G1***	ATPase H+ Transporting V1 Subunit G1	Compensation mechanism: promotes cell growth by probably interacting with the MAPK/Erk pathway [329].
30	***ACTR2***	Actin Related Protein 2	Compensation mechanism: its downregulation promotes cell growth by not interacting with MMD and CFL1 proteins [330].

31	***SLC15A4***	Solute Carrier Family 15 Member 4	Compensation mechanism: promotes cell growth through regulating cell cycle related pathway, mainly participate in the cell cycle and division [331].
32	***CHEK1***	Checkpoint Kinase 1	Compensation mechanism: CHEK1 is essential component to delay cell cycle progression [332].
33	***ATP6V1C1***	ATPase H+ Transporting V1 Subunit C1	Compensation mechanism: promotes cell growth by probably interacting with the mTORC1 pathway [298].
34	***TRIM68***	Tripartite Motif Containing 68	Compensation mechanism: promotes cell growth as TRIM68 is coactivator of androgen receptor [333].
35	***HDAC2***	Histone Deacetylase 2	Compensation mechanism: acts as tumor-suppressor by inhibiting tumor cell growth [334].
36	***BLM***	BLM RecQ Like Helicase	Compensation mechanism: acts as tumor-suppressor by inhibiting tumor cell growth [335].
37	***E2F4***	E2F Transcription Factor 4	Compensation mechanism: downregulation in E2F4 may be allowing cells to progress through the cell cycle and escape quiescence/dormancy, and activation of cell cycle regulator CDK6, promoting cell proliferation [336].
38	***DCPS***	Decapping Enzyme, Scavenger	Compensation mechanism: acts as tumor-suppressor by inhibiting tumor cell growth [337].
39	***CFL1***	Cofilin 1	Compensation mechanism: downregulation of CFL1 is related with increase proliferation and invasion rate [338].
40	***CYB5R1***	Cytochrome B5 Reductase 1	Drug metabolism: involved in oxidative stress protection and drug metabolism [339].
41	***IFITM1***	Interferon Induced Transmembrane Protein 1	Compensation mechanism: promotes distant metastasis [340].
42	***KRT16***	Keratin 16	Compensation mechanism: promotes migration [341].
43	***ATP2B4***	ATPase Plasma Membrane Ca2+ Transporting 4	Compensation and antiapoptotic mechanism: ATP2B4 promotes cell migration, and apoptotic resistance [342].
44	***IDS***	Iduronate 2-Sulfatase	Compensation mechanism: IDS enhances invasiveness through the ID2 – SNAIL axis [343].
45	***TBC1D25***	TBC1 Domain Family Member 25	TBC1D25 in cancer development have not been explored [344].
46	***NDUFS1***	NADH:Ubiquinone Oxidoreductase Core Subunit S1	No clear evidence about connecting this gene with potential resistance mechanism to FGFR inhibitor there is only information that patients with low NDUFS1 levels had poor overall survival [345].

47	***DIS3***	DIS3 Homolog, Exosome Endoribonuclease And 3'-5' Exoribonuclease	No clear evidence about connecting this gene with potential resistance mechanism to FGFR inhibitor there is only information that in myeloma DIS3 can be a driving force for tumorigenesis via DNA:RNA hybrid-dependent enhanced genome instability and increased mutational rate [346].
48	***HJURP***	Holliday Junction Recognition Protein	No clear evidence about connecting this gene with potential resistance mechanism to FGFR inhibitor there is only information that in triple-negative breast cancer HJURP regulates cell proliferation and chemo-resistance via YAP1/NDRG1 transcriptional axis. However in our case we are dealing with downregulation of this gene so it might be a moonlighting protein but I don't know in what this gene could be a negative regulator [347].
49	***RORC***	RAR Related Orphan Receptor C	No clear evidence about connecting this gene with potential resistance mechanism to FGFR inhibitor there is only information that RORC upregulation correlates with a poor prognosis [348].
50	***POLA2***	DNA Polymerase Alpha 2, Accessory Subunit	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that POLA2 promotes cell proliferation [349].
51	***PSMD7***	Proteasome 26S Subunit, Non-ATPase 7	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that PSMD7 promotes proliferation and apoptotic resistance by regulating the p53 pathway [350].
52	***EXOSC8***	Exosome Component 8	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that EXOSC8 promotes proliferation [351].
53	***CENPO***	Centromere Protein O	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that CENPO expression regulates gastric cancer cell proliferation [352].
54	***CCNB2***	Cyclin B2	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that CCNB2 promotes invasion and metastasis [353].
55	***CDT1***	Chromatin Licensing And DNA Replication Factor 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that Cdt1 overexpression most likely contributes to tumorigenecity by causing genomic instability [354].
56	***HIBCH***	3-Hydroxyisobutyryl-CoA Hydrolase	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that HIBCH promotes cell growth, resistant to apoptosis, and autophagy [355].
57	***MVD***	Mevalonate Diphosphate Decarboxylase	?

***** blue and red color indicates that this gene is downregulated or upregulated respectively.

Bladder

Using the ORA method on genes selected with the PREDICT pipeline, I have identified 37 significant pathways (q value < 0.050) (the list of significant pathways from ORA analysis can be found in Table S4 in Supplementary Materials placed on CD-R attached at the back of the dissertation). Similarly, in the case of the stomach pathways I have used the enrichplot tool v 1.13.1.994 [297], to identify pathway clusters. There was possible to distinguish 10 different groups of pathways (Figure 21), specifically related to: i) chaperone and protein folding, ii) rRNA processing, iii) mitosis, iv) cell cycle, and v) cell cycle and nuclear envelope.

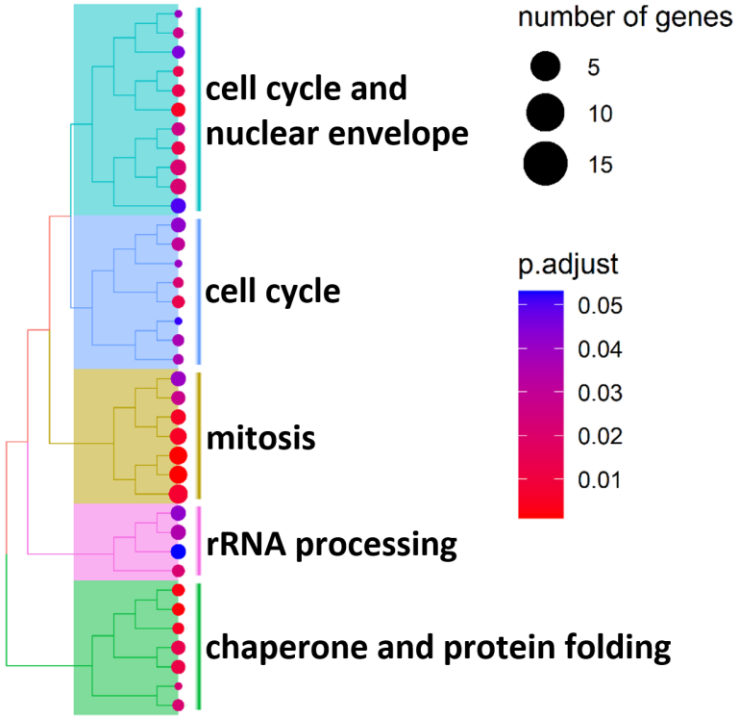


Figure 21. Hierarchical clustering of statistically significant (q value < 0.05) pathways identified with ORA analysis performed based on genes selected with the PREDICT pipeline from the bladder DEGs data set. The number of bladder DEGs identified in the pathways is represented using the size of the circles as indicated in the right panel.

As I mentioned earlier FGFR indirectly participate in regulating proliferation, and cell survival as well as maintaining tissue homeostasis, thus potential FGFR-TKIs resistant mechanism could lie in iii, iv, and v cluster (Figure 21). Therefore again as I did in the stomach case, in the gene set that came out related to these pathways I have selected genes that meet PREDICT statistical properties. That provided me with 39 genes (Table 16).

Based on the GSEA analysis I have identified 493 significant pathways (q value < 0.050) (the list of significant pathways from GSEA analysis can be found in Table S5 in Supplementary Materials placed on CD-R attached at the back of the dissertation). With enrichplot tool v

1.13.1.994 [297], there was possible to distinguish 10 different groups of pathways (Figure 22), specifically related to: i) NF-κB, MAPK, PTEN, migration, proliferation, apoptosis, ii) SUMOylation of proteins interacting with DNA and RNA, viral infection, mRNA processing, iii) RNA Polymerase II, HIV, iv) translation, v) HDR and DNA repair, vi) DNA repair, telomeres, translation, vii) centrosome, viii) kinetochore, ix) RNA Polymerase I & III, rRNA expression, and x) mitosis regulation, protein folding, mitochondrial translation, metabolism.

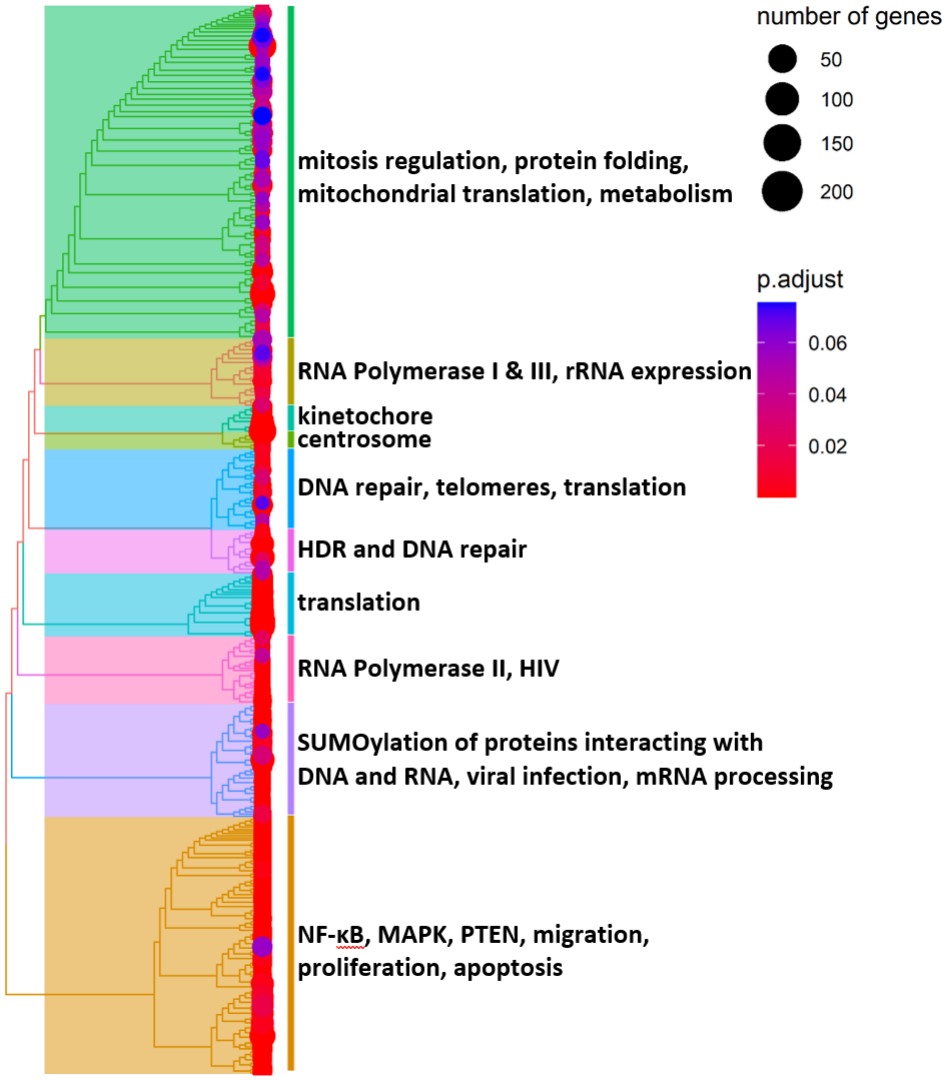


Figure 22. Hierarchical clustering of statistically significant (q value < 0.05) pathways identified with GSEA analysis performed based on the bladder DEGs data set. The number of bladder DEGs identified in the pathways is represented using the size of the circles as indicated in the right panel.

As I have described in chapter II. FGFR can indirectly affect migration, apoptosis, and proliferation, thus cluster i (Figure 22) may include pathways related to potential resistant mechanisms toward FGFR inhibitors. Therefore, I selected genes that met the PREDICT statistical properties from the gene set related to these pathways. This allows the identification of 40 genes (Table 16).

Since I have acquired a lot of common genes between gene sets identified based on this ORA and GSEA analysis (Figure 23), therefore for literature assessment I have extracted common genes (25 genes) between them and unique genes (14 and 15) at ORA and GSEA analysis respectively (Table 16).

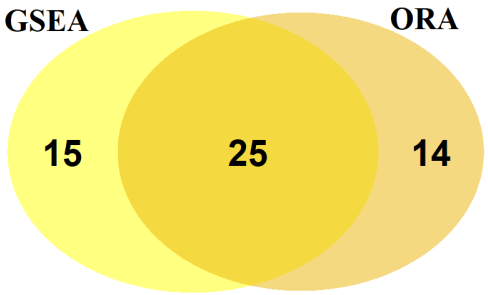


Figure 23. Venn diagram presenting a comparison between the gene lists selected with the PREDICT pipeline from a particular cluster of signaling pathways that were identified with the ORA or GSEA analysis (ORA: iii, iv, and v; GSEA: i).

Table 16. Common and unique lists of genes selected with the PREDICT pipeline from the bladder DEGs data set that came out related with the: iii or iv or v cluster of pathways from ORA analysis or i cluster of pathways from GSEA analysis.

Common genes between the both gene sets acquired based on ORA and GSEA analysis

No *	Gene symbol *	Gene name *	Potential mechanism of resistance to FGFR inhibitor
1	***FEN1***	Flap Structure-Specific Endonuclease 1	Compensation mechanism: downregulation of this genes maintain cell survival by preventing from its tumor-suppressor effect [356].
2	***AURKA***	Aurora Kinase A	Compensation mechanism: Most research reports that this protein promotes tumor progression when is upregulated [357]. In my case downregulation is difficult to explain. However there is report that reduced expression of AURKA is associated with increase invasiveness of breast cancer [358].
3	***PLK1***	Polo Like Kinase 1	Compensation mechanism: downregulated PLK1 increases proliferation rate whereas its overexpression leads to cell proliferation defects at least partially due to aberrant mitosis and the activation of the spindle assembly checkpoint [359].
4	***TUBB4B***	Tubulin Beta 4B Class IVb	Compensation mechanism: TUBB4B downregulation Is critical for increasing migration [360].
5	***UBE2C***	Ubiquitin Conjugating Enzyme E2 C	Compensation and drug metabolism mechanism: downregulation of this genes prevents from blocking migration and invasion, and increase drug resistance with induction metabolism-related pathways [361].

6	***HAUS7***	HAUS Augmin Like Complex Subunit 7	Antiapoptosis mechanism: p53 could be stabilized as a result of deubiquitination by HAUSP and suggested that HAUSP may thereby act as a tumour suppressor [362].
7	***LMNB1***	Lamin B1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein prevents senescence [363].
8	***TUBG1***	Tubulin Gamma 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that TUBG1 promotes cell proliferation, migration, and invasion and inhibited cell apoptosis [364].
9	***RFC3***	Replication Factor C Subunit 3	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes cell proliferation, migration, and invasion [365].
10	***KIF2C***	Kinesin Family Member 2C	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein proliferation, migration, and invasion [366].
11	***CDT1***	Chromatin Licensing And DNA Replication Factor 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation, migration, and invasion [367].
12	***CCNA2***	Cyclin A2	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes migration, and proliferation, and inhibited their apoptosis [368].
13	***POLE3***	DNA Polymerase Epsilon 3, Accessory Subunit	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes migration, and proliferation, and inhibited their apoptosis [369].
14	***TCP1***	T-Complex 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that downregulation of TCP1 expression can significantly reduce cell viability, inhibited cell proliferation, and migration [370].
15	***ERCC6L***	ERCC Excision Repair 6 Like, Spindle Assembly Checkpoint Helicase	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes cell growth and invasion [371].
16	***PSMD12***	Proteasome 26S Subunit, Non-ATPase 12	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation and invasion by interacting with Nrf2 [372].
17	***TPX2***	TPX2 Microtubule Nucleation Factor	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein increase proliferation and decrease the apoptosis [373].

18	***CDCA5***	Cell Division Cycle Associated 5	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation, and inhibited their apoptosis [374].
19	***CENPO***	Centromere Protein O	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that CENPO expression regulates gastric cancer cell proliferation [352].
20	***EML4***	EMAP Like 4	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein proliferation [375].
21	***CCNB2***	Cyclin B2	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation [376].
22	***TUBA1C***	Tubulin Alpha 1c	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation [377].
23	***CDK1***	Cyclin Dependent Kinase 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation [378].
24	***CENPM***	Centromere Protein M	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein increase migration and invasion [379].
25	***RRM2***	Ribonucleotide Reductase Regulatory Subunit M2	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein evokes cell invasion, migration and VEGF expression through the PI3K/AKT signaling pathway [380].

Unique genes for the gene set acquired based on ORA analysis

No *	Gene symbol *	Gene name *	Potential mechanism of resistance to FGFR inhibitor
1	***RTEL1***	Regulator Of Telomere Elongation Helicase 1	Compensation mechanism: depletion of RTEL1 leads to increased levels of TERRA RNA which is essential for cell viability [381].
2	***WRAP53***	WD Repeat Containing Antisense To TP53	Compensation mechanism: downregulated WRAP53-1α increase the mRNA and protein levels of p53 and there is increased proliferation [382].
3	***RBL2***	RB Transcriptional Corepressor Like 2	Compensation and antiapoptotic mechanism: promotes proliferation and inhibits apoptosis via interacting with AKT signaling pathway [383].
4	***PLK3***	Polo Like Kinase 3	Compensation and antiapoptotic mechanism: downregulation of this genes prevents from blocking cell proliferation and inducing apoptosis [384].

5	***PPP6C***	Protein Phosphatase 6 Catalytic Subunit	Compensation and antiapoptotic mechanism: PPP6C negatively regulates oncogenic RAF-MEK-ERK signaling. With PPP6C downregulation there is increased growth, prevention of apoptosis, and induction of drug resistance [385].
6	***IFT140***	Intraflagellar Transport 140	Compensation mechanism: promotes migration [386].
7	***SSRP1***	Structure Specific Recognition Protein 1	Antiapoptotic mechanism: when it is not downregulated it may potentiate cisplatin-induced cell death by blocking replication and repair of modified DNA [303].
8	***BBS2***	Bardet-Biedl Syndrome 2	No clear evidence about connecting this gene upregulation with potential resistance mechanism to FGFR inhibitor there is only information that patients with high BBS2 levels had poor survival in mesothelioma patients [387].
9	***BANF1***	BAF Nuclear Assembly Factor 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferative and migratory activity [388].
10	***CCT4***	Chaperonin Containing TCP1 Subunit 4	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein increase proliferation and decrease the apoptosis [389].
11	***NUP35***	Nucleoporin 35	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes proliferation [390].
12	***HMGA1***	High Mobility Group AT-Hook 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that downregulation of HMGA1 mediates autophagy and inhibits migration and invasion in bladder cancer via miRNA-221/TP53INP1/p-ERK Axis [391].
13	***TP53INP1***	Tumor Protein P53 Inducible Nuclear Protein 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein negatively regulates the metastasis [392].
14	***IST1***	IST1 Factor Associated With ESCRT-III	?

Unique genes for the gene set acquired based on GSEA analysis

No *	**Gene symbol** *	**Gene name** *	**Potential mechanism of resistance to FGFR inhibitor**
1	***COPS2***	COP9 Signalosome Subunit 2	Compensation mechanism: downregulation of this genes maintain cell survival by preventing from its tumor-suppressor effect [393].
2	***ERO1A***	Endoplasmic Reticulum Oxidoreductase 1 Alpha	Compensation mechanism: ERO1A drives the production of tumor-promoting myeloid-derived suppressor cells via oxidative protein folding [394].

3	***CCNF***	Cyclin F	Probably compensation mechanism: lower expression is related with worse survival [395].
4	***CHUK***	Component Of Inhibitor Of Nuclear Factor Kappa B Kinase Complex	Compensation mechanism: CHUK/IKK-α loss enhances growth associated with HIF up-regulation [396].
5	***RBM14***	RNA Binding Motif Protein 14	Compensation mechanism: RBM14 loss enhances proliferation. This protein suppress cell growth in part by down-regulating c-myc and its downstream effectors ccnd1 and skp2 and causing accumulation of p27/Kip1 protein [397].
6	***DUSP6***	Dual Specificity Phosphatase 6	Compensation mechanism: downregulation of this genes maintain cell growth and survival by preventing from its tumor-suppressor effect. This protein is negative regulator of kinase ERK1/2 and affects EGFR, TGF-β and WNT signaling pathways [398].
7	***CYCS***	Cytochrome C, Somatic	Compensation and antiapoptotic mechanism: CYCS by interacting with USP53 inhibits proliferation, migration and invasion, and induced apoptosis [399].
8	***SPRED1***	Sprouty Related EVH1 Domain Containing 1	Compensation and antiapoptotic mechanism: downregulation of SPRED1 is related with hyperactivation of the MAP/ERK pathway that augmented Bcl-2 expression and stability leading to increase of proliferation and apoptosis inhibition [400].
9	***AMD1***	Adenosylmethionine Decarboxylase 1	Antiapoptotic mechanism: downregulated AMD1 is related with lower apoptotic rate [401].
10	***PTRH2***	Peptidyl-TRNA Hydrolase 2	Antianoikis mechanism: PTRH2 by interacting with AES initiate anoikis. Downregulation of PTRH2 acts as pro-survival mechanism [402].
11	***SPRED2***	Sprouty Related EVH1 Domain Containing 2	Antiautophagy mechanism: Spred2 interaction with LC3 promotes autophagosome maturation and induces autophagy-dependent cell death [403].
12	***STAM***	Signal Transducing Adaptor Molecule	Compensation mechanism: STAM inhibits cell viability, invasion, and migration [404].
13	***NCBP1***	Nuclear Cap Binding Protein Subunit 1	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein promotes cell growth, wound healing ability, migration and epithelial-mesenchymal transition [405].
14	***ETV4***	ETS Variant Transcription Factor 4	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that ETV4 promotes metastasis in response to activation of PI3-kinase and Ras signaling [406].
15	***SRM***	Spermidine Synthase	I found no evidence of a link to a potential mechanism of resistance to the FGFR inhibitor. The available literature reports that this protein inhibit apoptosis [407].

***** blue and red color indicates that this gene is downregulated or upregulated respectively.

Based on the performed literature assessment I was able to connect 25 out of 54 genes with a potential mechanism of FGFR-TKIs resistance. Similarly, in the case of the stomach most of the genes might be related with a sort of compensatory activation of proliferation (12 genes), migration (6 genes), cell survival (5 genes), and invasion (3 genes) as a response to inhibition of this processes caused by blocking of FGF receptor/s signaling (Table 16). Here similarly to in the case of the stomach this is potentially mainly mediated through MAPK and AKT signaling pathways. However, in the case of this gene set presented in Table 16, we have much more examples that are tumor-suppressors like for example *RBM14* (RNA Binding Motif Protein 14) gene where its loss expression enhances proliferation. This protein suppresses cell growth in part by down-regulating c-myc and its downstream effectors ccnd1 and skp2 and causing accumulation of p27/Kip1 protein [397]. Another example can be *DUSP6* (Dual Specificity Phosphatase 6) gene. Where downregulation of this gene maintains cell growth and survival by preventing its tumor-suppressor effect. This protein is a negative regulator of kinase ERK1/2 and affects EGFR, TGF-β, and WNT signaling pathways [398]. The rest of the genes are briefly described in Table 16 and more information can be found under provided citations.

What is interesting is that there are 4 common genes between the stomach and bladder gene sets (Figure 24), specifically SSRP1 (Structure Specific Recognition Protein 1), CCNB2 (Cyclin B2), CDT1 (Chromatin Licensing And DNA Replication Factor 1), and CENPO (Centromere Protein O). Additionally what is more interesting is that both organs have the same direction of change which suggests that they are potential universal candidates for a predictive biomarker for both cancer types. Unfortunately, there is no common candidate with the lung gene set with PREDICT properties (Figure 24).

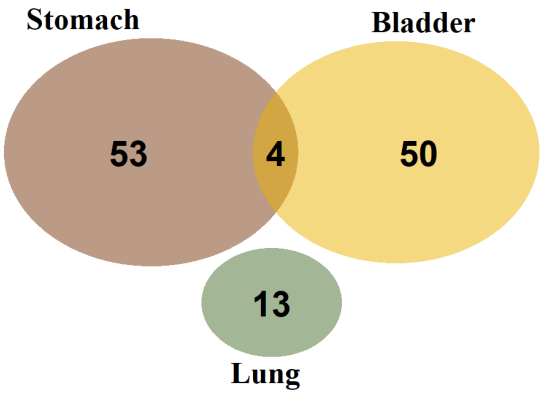


Figure 24. Venn diagram presenting a comparison between the stomach, bladder, and lung gene sets selected with the PREDICT pipeline from a particular cluster of signaling pathways that were identified with the ORA and/or GSEA analysis.

XIII. CONCLUSIONS

Using the RNA sequencing (RNA-seq) data, a comprehensive analysis of gene expression in cell lines from three different cancer types (lung, stomach, and bladder) was performed to identify potential predictive biomarker candidates related to mechanisms of FGFR tyrosine kinase inhibitors (FGFR-TKIs) resistance.

The standard method of selecting DEGs (Differentially Expressed Genes) can produce errors leading to potential biomarker candidates with an inconsistent direction of change which leads to a lack of reproducibility, as well as those with a small detectable difference, or a lack of biological effect, which can contribute to reduced sensitivity and specificity. This is particularly relevant for experiments with low sample sizes.

In order to overcome these limitations, the "Pipeline for Rapid Evaluation, and Discovery of Important biomarker CandidaTes" (PREDICT) was developed, based on sequentially applying thresholds of log2FC > 0.500, q value < 0.050, log2minFC > 0.100, and minDiff > 100 to the results obtained from the standard differential analysis method.

Utilizing the statistical properties implemented in the PREDICT pipeline, led to filtering out the unwanted results. Thus, the numbers of DEGs obtained by this method were 13, 226, and 301 for the lung, stomach, and bladder data set respectively. The selected biomarker candidates possessed characteristics suitable for a biomarker that can be applied in clinical settings.

- Based on differential expression analysis performed by the DESeq2 tool, the statistically significant DEGs were identified. Adopted threshold of q value = 0.050. Genes with the q value below the threshold were filtered out.
- The selected biomarker candidates were characterized by log2 fold change (log2FC) value > 0.500.
- There was minimal Fold Change (*minFC*) between minimal and maximal values of a particular gene expression measurement between two non-overlapping groups of measurements, respectively. Adopted threshold of log2*minFC* = 0.100. Genes with the log2minFC value below the threshold were filtered out.
- There was minimal Difference (*minDiff*) between minimal and maximal values of a particular gene expression measurement between two non-overlapping groups of measurements,

respectively. Adopted threshold of *minDiff* = 100. Genes with the *minDiff* value below the threshold were filtered out.

- By implementing *minFC* and *minDiff* measures, expression value intervals for the groups do not overlap. Therefore providing us with potential candidates with the desired reproducibility characteristics.
- By implementing adopted thresholds for log2FC, q value, *minFC*, and *minDiff* measures, enables the selection of potential candidates with the desired sensitivity and specificity characteristics.

Due to the limitations of the fold change measure, it was assessed whether replacing this measure with a more recommended measure of effect size, such as Cohen's d, could significantly improve the results obtained using the PREDICT pipeline. However, as this replacement would not significantly alter the results (number of selected DEGs), and given that the fold change measure is widely used and in many cases preferred by scientists across various biological fields, as well as it has utility for various downstream analysis tasks, such as prioritizing genes for further investigation and linking it with other variables of interest, it was decided against replacing it within the PREDICT pipeline.

The DEGs identified with the DESeq2 tool and followed PREDICT pipeline were further used to assess the biological context. The context was assessed by signaling pathway enrichment analysis, where two methods were employed: over-representation analysis (ORA) with the gene list selected with the PREDICT pipeline, and gene set enrichment analysis (GSEA) with genes ranked according to the Wald test statistic. In the stomach and bladder data set significant pathways were clustered to distinguish groups of similar pathways and to select groups that were potentially related to FGFR-TKIs resistant mechanism. Then, in the gene set that came out related to selected clusters of pathways, genes were selected that met PREDICT statistical properties. As 57 and 54 genes were identified for the stomach and bladder, respectively, they were assessed based on the published literature. In the case of the lung data set, only 13 DEGs were selected with the PREDICT pipeline, so a literature assessment was performed for all of these genes.

Based on signaling pathway analysis, combined with the use of PREDICT pipeline and literature search, it was possible to uncover the link with potential resistance mechanisms towards FGFR-TKIs for majority of selected genes. These findings indicate that resistant tumors exhibit compensatory activation of pathways regulating cell proliferation, migration rate, survival, invasiveness, and antiapoptotic properties, in response to FGFR-TKIs treatment.

By comparing the selected gene sets between the three different cancer types, several potential universal biomarkers of FGFR-TKIs resistance were identified, including *SSRP1* (Structure Specific Recognition Protein 1), *CCNB2* (Cyclin B2), *CDT1* (Chromatin Licensing And DNA Replication Factor 1), and *CENPO* (Centromere Protein O). These genes were commonly dysregulated in both stomach and bladder cancer and showed the same direction of change in expression in these two cancer types. They may serve as universal biomarkers for predicting FGFR-TKIs resistance in patients with diagnosed stomach or bladder cancer.

In conclusion, the use of the PREDICT pipeline led to the filtering out the unwanted results, and the selected biomarker candidates possess characteristics suitable for a biomarker that can be applied in clinical settings. An extensive literature search uncovered the link with potential resistance mechanisms towards FGFR-TKIs for the majority of selected genes. The next step in biomarker development would be validation/qualification phase to confirm that the differential expression observed in the discovery phase can be seen using other methods and on the different biological material.

Acknowledgments

This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029.

REFERENCES

1. World Health Organization, *Cancer.* 2023.
2. Dasgupta, A. and A. Wahed, *Chapter 13 - Tumor markers*, in *Clinical Chemistry, Immunology and Laboratory Quality Control (Second Edition)*, A. Dasgupta and A. Wahed, Editors. 2021, Elsevier. p. 269-293.
3. Füzéry, A.K., et al., *Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges.* Clinical Proteomics, 2013. **10**(1): p. 13.
4. McDermott, J.E., et al., *Challenges in Biomarker Discovery: Combining Expert Insights with Statistical Analysis of Complex Omics Data.* Expert Opin Med Diagn, 2013. **7**(1): p. 37-51.
5. Xue, W.J., et al., *Recent developments and advances of FGFR as a potential target in cancer.* Future Med Chem, 2018. **10**(17): p. 2109-2126.
6. Dieci, M.V., et al., *Fibroblast growth factor receptor inhibitors as a cancer treatment: from a biologic rationale to medical perspectives.* Cancer Discov, 2013. **3**(3): p. 264-79.
7. Krause, D.S. and R.A. Van Etten, *Tyrosine kinases as targets for cancer therapy.* N Engl J Med, 2005. **353**(2): p. 172-87.
8. Regeenes, R., et al., *Fibroblast growth factor receptor 5 (FGFR5) is a co-receptor for FGFR1 that is up-regulated in beta-cells by cytokine-induced inflammation.* J Biol Chem, 2018. **293**(44): p. 17218-17228.
9. Trueb, B., *Biology of FGFRL1, the fifth fibroblast growth factor receptor.* Cell Mol Life Sci, 2011. **68**(6): p. 951-64.
10. Turner, N. and R. Grose, *Fibroblast growth factor signalling: from development to cancer.* Nature Reviews Cancer, 2010. **10**(2): p. 116-129.
11. Mikhaylenko, D.S., et al., *Structural Alterations in Human Fibroblast Growth Factor Receptors in Carcinogenesis.* Biochemistry (Mosc), 2018. **83**(8): p. 930-943.
12. Beenken, A. and M. Mohammadi, *The FGF family: biology, pathophysiology and therapy.* Nature Reviews Drug Discovery, 2009. **8**(3): p. 235-253.
13. Babina, I.S. and N.C. Turner, *Advances and challenges in targeting FGFR signalling in cancer.* Nat Rev Cancer, 2017. **17**(5): p. 318-332.
14. Futami, T., et al., *Identification of a novel oncogenic mutation of FGFR4 in gastric cancer.* Sci Rep, 2019. **9**(1): p. 14627.
15. Nyeng, P., et al., *FGF10 signaling controls stomach morphogenesis.* Dev Biol, 2007. **303**(1): p. 295-310.
16. Cardoso, W.V., et al., *FGF-1 and FGF-7 induce distinct patterns of growth and differentiation in embryonic lung epithelium.* Dev Dyn, 1997. **208**(3): p. 398-405.
17. Yin, Y., F. Wang, and D.M. Ornitz, *Mesothelial- and epithelial-derived FGF9 have distinct functions in the regulation of lung development.* Development, 2011. **138**(15): p. 3169-77.
18. Ikeda, Y., et al., *Fgfr2 is integral for bladder mesenchyme patterning and function.* Am J Physiol Renal Physiol, 2017. **312**(4): p. F607-f618.
19. Xie, Y., et al., *FGF/FGFR signaling in health and disease.* Signal Transduct Target Ther, 2020. **5**(1): p. 181.
20. Park, H.K., et al., *Distinct association of genetic variations of vascular endothelial growth factor, transforming growth factor-beta, and fibroblast growth factor receptors with atopy and airway hyperresponsiveness.* Allergy, 2008. **63**(4): p. 447-53.

21. Helsten, T., et al., *The FGFR Landscape in Cancer: Analysis of 4,853 Tumors by Next-Generation Sequencing.* Clin Cancer Res, 2016. **22**(1): p. 259-67.
22. Dienstmann, R., et al., *Genomic aberrations in the FGFR pathway: opportunities for targeted therapies in solid tumors.* Ann Oncol, 2014. **25**(3): p. 552-563.
23. Wu, Y.M., et al., *Identification of targetable FGFR gene fusions in diverse cancers.* Cancer Discov, 2013. **3**(6): p. 636-47.
24. Hierro, C., et al., *Targeting the fibroblast growth factor receptor 2 in gastric cancer: promise or pitfall?* Ann Oncol, 2018. **29**(7): p. 1605.
25. Campbell, J.D., et al., *Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas.* Nat Genet, 2016. **48**(6): p. 607-16.
26. Peifer, M., et al., *Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer.* Nat Genet, 2012. **44**(10): p. 1104-10.
27. Weiss, J., et al., *Frequent and focal FGFR1 amplification associates with therapeutically tractable FGFR1 dependency in squamous cell lung cancer.* Sci Transl Med, 2010. **2**(62): p. 62ra93.
28. Luo, J., et al., *An mRNA Gene Expression-Based Signature to Identify FGFR1-Amplified Estrogen Receptor-Positive Breast Tumors.* J Mol Diagn, 2017. **19**(1): p. 147-161.
29. Tchaicha, J.H., et al., *Kinase domain activation of FGFR2 yields high-grade lung adenocarcinoma sensitive to a Pan-FGFR inhibitor in a mouse model of NSCLC.* Cancer Res, 2014. **74**(17): p. 4676-84.
30. Turner, N., et al., *Integrative molecular profiling of triple negative breast cancers identifies amplicon drivers and potential therapeutic targets.* Oncogene, 2010. **29**(14): p. 2013-23.
31. Cancer Genome Atlas Research, N., *Comprehensive molecular characterization of urothelial bladder carcinoma.* Nature, 2014. **507**(7492): p. 315-22.
32. Gao, Q., et al., *Driver Fusions and Their Implications in the Development and Treatment of Human Cancers.* Cell Rep, 2018. **23**(1): p. 227-238.e3.
33. Taylor, J.G.t., et al., *Identification of FGFR4-activating mutations in human rhabdomyosarcomas that promote metastasis in xenotransplanted models.* J Clin Invest, 2009. **119**(11): p. 3395-407.
34. Morimoto, Y., et al., *Single nucleotide polymorphism in fibroblast growth factor receptor 4 at codon 388 is associated with prognosis in high-grade soft tissue sarcoma.* Cancer, 2003. **98**(10): p. 2245-50.
35. Krook, M.A., et al., *Fibroblast growth factor receptors in cancer: genetic alterations, diagnostics, therapeutic targets and mechanisms of resistance.* Br J Cancer, 2021. **124**(5): p. 880-892.
36. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
37. Zarczynska, I., et al., *p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer.* Cells, 2021. **10**(12).
38. Toumazis, I., et al., *Risk-Based lung cancer screening: A systematic review.* Lung Cancer, 2020. **147**: p. 154-186.
39. Ferlay, J., et al., *Cancer statistics for the year 2020: An overview.* Int J Cancer, 2021.
40. Global Burden of Disease Cancer, C., et al., *Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017: A Systematic Analysis for the Global Burden of Disease Study.* JAMA Oncol, 2019. **5**(12): p. 1749-1768.
41. Bray, F., et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.* CA Cancer J Clin, 2018. **68**(6): p. 394-424.
42. Siegel, R.L., K.D. Miller, and A. Jemal, *Cancer statistics, 2016.* CA Cancer J Clin, 2016. **66**(1): p. 7-30.
43. Asplund, J., et al., *Survival Trends in Gastric Adenocarcinoma: A Population-Based Study in Sweden.* Ann Surg Oncol, 2018. **25**(9): p. 2693-2702.
44. Van Cutsem, E., et al., *Gastric cancer.* Lancet, 2016. **388**(10060): p. 2654-2664.

45. Ilic, M. and I. Ilic, *Epidemiology of stomach cancer.* World J Gastroenterol, 2022. **28**(12): p. 1187-1203.
46. Deng, W., et al., *Alcohol consumption and risk of stomach cancer: A meta-analysis.* Chem Biol Interact, 2021. **336**: p. 109365.
47. Collatuzzo, G., et al., *Exploring the interactions between Helicobacter pylori (Hp) infection and other risk factors of gastric cancer: A pooled analysis in the Stomach cancer Pooling (StoP) Project.* Int J Cancer, 2021. **149**(6): p. 1228-1238.
48. Karimi, P., et al., *Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention.* Cancer Epidemiol Biomarkers Prev, 2014. **23**(5): p. 700-13.
49. Shariat, S.F., M. Milowsky, and M.J. Droller, *Bladder cancer in the elderly.* Urol Oncol, 2009. **27**(6): p. 653-67.
50. Dobruch, J., et al., *Gender and Bladder Cancer: A Collaborative Review of Etiology, Biology, and Outcomes.* Eur Urol, 2016. **69**(2): p. 300-10.
51. Shariat, S.F., et al., *The effect of age and gender on bladder cancer: a critical review of the literature.* BJU Int, 2010. **105**(3): p. 300-8.
52. Tran, L., et al., *Advances in bladder cancer biology and therapy.* Nat Rev Cancer, 2021. **21**(2): p. 104-121.
53. Weijers, Y., et al., *Management of low-risk and intermediate-risk non-muscle-invasive bladder carcinoma.* Hematol Oncol Clin North Am, 2015. **29**(2): p. 219-25, vii.
54. Cohen, P., *Protein kinases--the major drug targets of the twenty-first century?* Nat Rev Drug Discov, 2002. **1**(4): p. 309-15.
55. Witte, O.N., A. Dasgupta, and D. Baltimore, *Abelson murine leukaemia virus protein is phosphorylated in vitro to form phosphotyrosine.* Nature, 1980. **283**(5750): p. 826-31.
56. Cohen, P., D. Cross, and P.A. Jänne, *Kinase drug discovery 20 years after imatinib: progress and future directions.* Nat Rev Drug Discov, 2021. **20**(7): p. 551-569.
57. Tirumani, S.H., et al., *Imatinib and beyond in gastrointestinal stromal tumors: A radiologist's perspective.* AJR Am J Roentgenol, 2013. **201**(4): p. 801-10.
58. Wertheimer, C., et al., *EGFR inhibitor Gefitinib attenuates posterior capsule opacification in vitro and in the ex vivo human capsular bag model.* Graefes Arch Clin Exp Ophthalmol, 2015. **253**(3): p. 409-17.
59. Burotto, M., et al., *Gefitinib and erlotinib in metastatic non-small cell lung cancer: a meta-analysis of toxicity and efficacy of randomized clinical trials.* Oncologist, 2015. **20**(4): p. 400-10.
60. Kuczynski, E.A., et al., *Effects of Sorafenib Dose on Acquired Reversible Resistance and Toxicity in Hepatocellular Carcinoma.* Cancer Res, 2015. **75**(12): p. 2510-9.
61. Cheng, A.L., et al., *Sunitinib versus sorafenib in advanced hepatocellular cancer: results of a randomized phase III trial.* J Clin Oncol, 2013. **31**(32): p. 4067-75.
62. De Silva, N., et al., *Molecular effects of Lapatinib in the treatment of HER2 overexpressing oesophago-gastric adenocarcinoma.* Br J Cancer, 2015. **113**(9): p. 1305-12.
63. Pusztai, L., et al., *Gene signature-guided dasatinib therapy in metastatic breast cancer.* Clin Cancer Res, 2014. **20**(20): p. 5265-71.
64. Sankar, K., S.M. Gadgeel, and A. Qin, *Molecular therapeutic targets in non-small cell lung cancer.* Expert Rev Anticancer Ther, 2020. **20**(8): p. 647-661.
65. Jiao, Q., et al., *Advances in studies of tyrosine kinase inhibitors and their acquired resistance.* Mol Cancer, 2018. **17**(1): p. 36.
66. Touat, M., et al., *Targeting FGFR Signaling in Cancer.* Clin Cancer Res, 2015. **21**(12): p. 2684-94.
67. Loriot, Y., et al., *Erdafitinib in Locally Advanced or Metastatic Urothelial Carcinoma.* N Engl J Med, 2019. **381**(4): p. 338-348.
68. Montazeri, K. and J. Bellmunt, *Erdafitinib for the treatment of metastatic bladder cancer.* Expert Rev Clin Pharmacol, 2020. **13**(1): p. 1-6.

69. Abou-Alfa, G.K., et al., *Pemigatinib for previously treated, locally advanced or metastatic cholangiocarcinoma: a multicentre, open-label, phase 2 study.* Lancet Oncol, 2020. **21**(5): p. 671-684.
70. Hoy, S.M., *Pemigatinib: First Approval.* Drugs, 2020. **80**(9): p. 923-929.
71. ClinicalTrials.gov.
72. Yue, S., et al., *FGFR-TKI resistance in cancer: current status and perspectives.* J Hematol Oncol, 2021. **14**(1): p. 23.
73. Lacouture, M.E., et al., *Dermatologic Adverse Events Associated with Selective Fibroblast Growth Factor Receptor Inhibitors: Overview, Prevention, and Management Guidelines.* Oncologist, 2021. **26**(2): p. e316-e326.
74. Zhou, Y., et al., *FGF/FGFR signaling pathway involved resistance in various cancer types.* J Cancer, 2020. **11**(8): p. 2000-2007.
75. Sohl, C.D., et al., *Illuminating the molecular mechanisms of tyrosine kinase inhibitor resistance for the FGFR1 gatekeeper mutation: the Achilles' heel of targeted therapy.* ACS Chem Biol, 2015. **10**(5): p. 1319-29.
76. Yoza, K., et al., *Biophysical characterization of drug-resistant mutants of fibroblast growth factor receptor 1.* Genes Cells, 2016. **21**(10): p. 1049-1058.
77. Chell, V., et al., *Tumour cell responses to new fibroblast growth factor receptor tyrosine kinase inhibitors and identification of a gatekeeper mutation in FGFR3 as a mechanism of acquired resistance.* Oncogene, 2013. **32**(25): p. 3059-70.
78. Goyal, L., et al., *Polyclonal Secondary FGFR2 Mutations Drive Acquired Resistance to FGFR Inhibition in Patients with FGFR2 Fusion-Positive Cholangiocarcinoma.* Cancer Discov, 2017. **7**(3): p. 252-263.
79. Kas, S.M., et al., *Transcriptomics and Transposon Mutagenesis Identify Multiple Mechanisms of Resistance to the FGFR Inhibitor AZD4547.* Cancer Res, 2018. **78**(19): p. 5668-5679.
80. Datta, J., et al., *Akt Activation Mediates Acquired Resistance to Fibroblast Growth Factor Receptor Inhibitor BGJ398.* Mol Cancer Ther, 2017. **16**(4): p. 614-624.
81. Wang, L., et al., *A Functional Genetic Screen Identifies the Phosphoinositide 3-kinase Pathway as a Determinant of Resistance to Fibroblast Growth Factor Receptor Inhibitors in FGFR Mutant Urothelial Cell Carcinoma.* Eur Urol, 2017. **71**(6): p. 858-862.
82. Herrera-Abreu, M.T., et al., *Parallel RNA interference screens identify EGFR activation as an escape mechanism in FGFR3-mutant cancer.* Cancer Discov, 2013. **3**(9): p. 1058-71.
83. Gozgit, J.M., et al., *Combined targeting of FGFR2 and mTOR by ponatinib and ridaforolimus results in synergistic antitumor activity in FGFR2 mutant endometrial cancer models.* Cancer Chemother Pharmacol, 2013. **71**(5): p. 1315-23.
84. Grygielewicz, P., et al., *Epithelial-mesenchymal transition confers resistance to selective FGFR inhibitors in SNU-16 gastric cancer cells.* Gastric Cancer, 2016. **19**(1): p. 53-62.
85. Pandha, H.S. and J. Waxman, *Tumour markers.* QJM, 1995. **88**(4): p. 233-41.
86. Gold, P. and S.O. Freedman, *Demonstration of Tumor-Specific Antigens in Human Colonic Carcinomata by Immunological Tolerance and Absorption Techniques.* J Exp Med, 1965. **121**(3): p. 439-62.
87. Dasgupta, A. and A. Wahed, *Clinical Chemistry, Immunology and Laboratory Quality Control: A Comprehensive Review for Board Preparation, Certification and Clinical Practice*. Clinical Chemistry, Immunology and Laboratory Quality Control: A Comprehensive Review for Board Preparation, Certification and Clinical Practice. 2014: Elsevier Inc. 1-483.
88. Fda-Nih_Biomarker_Working_Group, in *BEST (Biomarkers, EndpointS, and other Tools) Resource*. 2016, Food and Drug Administration (US) National Institutes of Health (US): Silver Spring (MD) Bethesda (MD).
89. Califf, R.M., *Biomarker definitions and their applications.* Exp Biol Med (Maywood), 2018. **243**(3): p. 213-221.

90. Sharma, S., *Tumor markers in clinical practice: General principles and guidelines.* Indian J Med Paediatr Oncol, 2009. **30**(1): p. 1-8.
91. New York State Department of Health, *Disease Screening - Statistics Teaching Tools.* 2023.
92. Byrnes, S.A. and B.H. Weigl, *Selecting analytical biomarkers for diagnostic applications: a first principles approach.* Expert Rev Mol Diagn, 2018. **18**(1): p. 19-26.
93. Wishart, D.S., et al., *MarkerDB: an online database of molecular biomarkers.* Nucleic Acids Res, 2021. **49**(D1): p. D1259-D1267.
94. Nalejska, E., E. Maczynska, and M.A. Lewandowska, *Prognostic and predictive biomarkers: tools in personalized oncology.* Mol Diagn Ther, 2014. **18**(3): p. 273-84.
95. Ray, P., et al., *Statistical evaluation of a biomarker.* Anesthesiology, 2010. **112**(4): p. 1023-40.
96. Simon, R., *Sensitivity, Specificity, PPV, and NPV for Predictive Biomarkers.* J Natl Cancer Inst, 2015. **107**(8).
97. Yang, Z.Y., et al., *Promising biomarkers for predicting the outcomes of patients with KRAS wild-type metastatic colorectal cancer treated with anti-epidermal growth factor receptor monoclonal antibodies: a systematic review with meta-analysis.* Int J Cancer, 2013. **133**(8): p. 1914-25.
98. Dahabreh, I.J., et al., *Systematic review: Anti-epidermal growth factor receptor treatment effect modification by KRAS mutations in advanced colorectal cancer.* Ann Intern Med, 2011. **154**(1): p. 37-49.
99. Hoffman, R.M., et al., *Prostate-specific antigen testing accuracy in community practice.* BMC Fam Pract, 2002. **3**: p. 19.
100. Wolf, A.M., et al., *American Cancer Society guideline for the early detection of prostate cancer: update 2010.* CA Cancer J Clin, 2010. **60**(2): p. 70-98.
101. Ray, P., et al., *Usefulness of B-type natriuretic peptide in elderly patients with acute dyspnea.* Intensive Care Med, 2004. **30**(12): p. 2230-6.
102. Aye, P.S., et al., *Development and validation of a predictive model for estimating EGFR mutation probabilities in patients with non-squamous non-small cell lung cancer in New Zealand.* BMC Cancer, 2020. **20**(1): p. 658.
103. Chau, C.H., et al., *Validation of analytic methods for biomarkers used in drug development.* Clin Cancer Res, 2008. **14**(19): p. 5967-76.
104. Kang, Y., S. Vijay, and T.S. Gujral, *Deep neural network modeling identifies biomarkers of response to immune-checkpoint therapy.* iScience, 2022. **25**(5): p. 104228.
105. Piorino, F., A.T. Patterson, and M.P. Styczynski, *Low-cost, point-of-care biomarker quantification.* Curr Opin Biotechnol, 2022. **76**: p. 102738.
106. Ong, S.E. and M. Mann, *Mass spectrometry-based proteomics turns quantitative.* Nat Chem Biol, 2005. **1**(5): p. 252-62.
107. Rifai, N., M.A. Gillette, and S.A. Carr, *Protein biomarker discovery and validation: the long and uncertain path to clinical utility.* Nat Biotechnol, 2006. **24**(8): p. 971-83.
108. Anderson, N.L., *The roles of multiple proteomic platforms in a pipeline for new diagnostics.* Mol Cell Proteomics, 2005. **4**(10): p. 1441-4.
109. Atashpaz-Gargari, E., U.M. Braga-Neto, and E.R. Dougherty, *Modeling and systematic analysis of biomarker validation using selected reaction monitoring.* EURASIP J Bioinform Syst Biol, 2014. **2014**: p. 17.
110. Li, A. and R.C. Bergan, *Clinical trial design: Past, present, and future in the context of big data and precision medicine.* Cancer, 2020. **126**(22): p. 4838-4846.
111. Korn, E.L., et al., *Meta-analysis of phase II cooperative group trials in metastatic stage IV melanoma to determine progression-free and overall survival benchmarks for future phase II trials.* J Clin Oncol, 2008. **26**(4): p. 527-34.
112. Simon, R., *Optimal two-stage designs for phase II clinical trials.* Control Clin Trials, 1989. **10**(1): p. 1-10.
113. Walker, E. and A.S. Nowacki, *Understanding equivalence and noninferiority testing.* J Gen Intern Med, 2011. **26**(2): p. 192-6.

114. Simon, R., *Biomarker based clinical trial design.* Chin Clin Oncol, 2014. **3**(3): p. 39.
115. Pusztai, L. and K.R. Hess, *Clinical trial design for microarray predictive marker discovery and assessment.* Ann Oncol, 2004. **15**(12): p. 1731-7.
116. Jones, C.L. and E. Holmgren, *An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies.* Contemp Clin Trials, 2007. **28**(5): p. 654-61.
117. Freidlin, B., et al., *Randomized phase II trial designs with biomarkers.* J Clin Oncol, 2012. **30**(26): p. 3304-9.
118. Kim, E.S., et al., *The BATTLE trial: personalizing therapy for lung cancer.* Cancer Discov, 2011. **1**(1): p. 44-53.
119. Simon, R. and E. Polley, *Clinical trials for precision oncology using next-generation sequencing.* Per Med, 2013. **10**(5): p. 485-495.
120. Shak, S., *Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. Herceptin Multinational Investigator Study Group.* Semin Oncol, 1999. **26**(4 Suppl 12): p. 71-7.
121. Chapman, P.B., et al., *Improved survival with vemurafenib in melanoma with BRAF V600E mutation.* N Engl J Med, 2011. **364**(26): p. 2507-16.
122. Shaw, A.T., et al., *Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis.* Lancet Oncol, 2011. **12**(11): p. 1004-12.
123. Karuri, S.W. and R. Simon, *A two-stage Bayesian design for co-development of new drugs and companion diagnostics.* Stat Med, 2012. **31**(10): p. 901-14.
124. Hong, F. and R. Simon, *Run-in phase III trial design with pharmacodynamics predictive biomarkers.* J Natl Cancer Inst, 2013. **105**(21): p. 1628-33.
125. Simon, R.M., S. Paik, and D.F. Hayes, *Use of archived specimens in evaluation of prognostic and predictive biomarkers.* J Natl Cancer Inst, 2009. **101**(21): p. 1446-52.
126. Jiang, W., B. Freidlin, and R. Simon, *Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect.* J Natl Cancer Inst, 2007. **99**(13): p. 1036-43.
127. Matsui, S., et al., *Developing and validating continuous genomic signatures in randomized clinical trials for predictive medicine.* Clin Cancer Res, 2012. **18**(21): p. 6065-73.
128. Gu, X., G. Yin, and J.J. Lee, *Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints.* Contemp Clin Trials, 2013. **36**(2): p. 642-50.
129. Stollmeier, F., T. Geisel, and J. Nagler, *Possible origin of stagnation and variability of earth's biodiversity.* Phys Rev Lett, 2014. **112**(22): p. 228101.
130. Sala, O.E., et al., *Global biodiversity scenarios for the year 2100.* Science, 2000. **287**(5459): p. 1770-4.
131. Pettibone, L., K. Vohland, and D. Ziegler, *Understanding the (inter)disciplinary and institutional diversity of citizen science: A survey of current practice in Germany and Austria.* PLoS One, 2017. **12**(6): p. e0178778.
132. Couee, I., *Linguistic diversity in a new agenda for equity in science.* Proc Natl Acad Sci U S A, 2022. **119**(23): p. e2204376119.
133. Goetzmann, W.N. and A. Kumar, *Equity portfolio diversification.* Review of Finance, 2008. **12**(3): p. 433-463.
134. Purvis, A. and A. Hector, *Getting the measure of biodiversity.* Nature, 2000. **405**(6783): p. 212-9.
135. Whittaker, R.J., K.J. Willis, and R. Field, *Scale and species richness: Towards a general, hierarchical theory of species diversity.* Journal of Biogeography, 2001. **28**(4): p. 453-470.
136. Xu, S., L. Bottcher, and T. Chou, *Diversity in biology: definitions, quantification and models.* Phys Biol, 2020. **17**(3): p. 031001.
137. Sarkar, S., *Ecological diversity and biodiversity as concepts for conservation planning: Comments on Ricotta.* Acta Biotheoretica, 2006. **54**(2): p. 133-140.

138. Duclos, J.Y., J. Esteban, and D. Ray, *Polarization: Concepts, measurement, estimation.* Econometrica, 2004. **72**(6): p. 1737-1772.
139. Mäs, M., et al., *In the short term we divide, in the long term we unite: Demographic crisscrossing and the effects of faultlines on subgroup polarization.* Organization Science, 2013. **24**(3): p. 716-736.
140. Grubb, M., L. Butler, and P. Twomey, *Diversity and security in UK electricity generation: The influence of low-carbon objectives.* Energy Policy, 2006. **34**(18): p. 4050-4062.
141. Vogelstein, B., et al., *Cancer genome landscapes.* Science, 2013. **339**(6127): p. 1546-58.
142. Sawyers, C., *Targeted cancer therapy.* Nature, 2004. **432**(7015): p. 294-7.
143. Haber, D.A., et al., *Molecular targeted therapy of lung cancer: EGFR mutations and response to EGFR inhibitors.* Cold Spring Harb Symp Quant Biol, 2005. **70**: p. 419-26.
144. Huang, M., et al., *Molecularly targeted cancer therapy: some lessons from the past decade.* Trends Pharmacol Sci, 2014. **35**(1): p. 41-50.
145. Leiserson, M.D., et al., *Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes.* Nat Genet, 2015. **47**(2): p. 106-14.
146. Kaye, S.B., *Progress in the treatment of ovarian cancer-lessons from homologous recombination deficiency-the first 10 years.* Ann Oncol, 2016. **27 Suppl 1**(Suppl 1): p. i1-i3.
147. Livraghi, L. and J.E. Garber, *PARP inhibitors in the management of breast cancer: current data and future prospects.* BMC Med, 2015. **13**: p. 188.
148. Iorio, F., et al., *A Landscape of Pharmacogenomic Interactions in Cancer.* Cell, 2016. **166**(3): p. 740-754.
149. Mroz, E.A. and J.W. Rocco, *The challenges of tumor genetic diversity.* Cancer, 2017. **123**(6): p. 917-927.
150. Flaherty, K.T., et al., *Molecular Landscape and Actionable Alterations in a Genomically Guided Cancer Clinical Trial: National Cancer Institute Molecular Analysis for Therapy Choice (NCI-MATCH).* J Clin Oncol, 2020. **38**(33): p. 3883-3894.
151. Herold, F., et al., *Causes and Consequences of Interindividual Response Variability: A Call to Apply a More Rigorous Research Design in Acute Exercise-Cognition Studies.* Front Physiol, 2021. **12**: p. 682891.
152. Zolotovskaia, M.A., et al., *Disparity between Inter-Patient Molecular Heterogeneity and Repertoires of Target Drugs Used for Different Types of Cancer in Clinical Oncology.* Int J Mol Sci, 2020. **21**(5).
153. Schell, R.F., et al., *Meta-analysis of inter-patient pharmacokinetic variability of liposomal and non-liposomal anticancer agents.* Nanomedicine, 2014. **10**(1): p. 109-17.
154. Ilan, Y., *Next-Generation Personalized Medicine: Implementation of Variability Patterns for Overcoming Drug Resistance in Chronic Diseases.* J Pers Med, 2022. **12**(8).
155. Marusyk, A., V. Almendro, and K. Polyak, *Intra-tumour heterogeneity: a looking glass for cancer?* Nat Rev Cancer, 2012. **12**(5): p. 323-34.
156. Mazor, T., et al., *Intratumoral Heterogeneity of the Epigenome.* Cancer Cell, 2016. **29**(4): p. 440-451.
157. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.
158. Yap, T.A., et al., *Intratumor heterogeneity: seeing the wood for the trees.* Sci Transl Med, 2012. **4**(127): p. 127ps10.
159. Fidler, I.J. and M.L. Kripke, *Metastasis results from preexisting variant cells within a malignant tumor.* Science, 1977. **197**(4306): p. 893-5.
160. Dexter, D.L., et al., *Heterogeneity of tumor cells from a single mouse mammary tumor.* Cancer Res, 1978. **38**(10): p. 3174-81.
161. Håkansson, L. and C. Tropé, *On the presence within tumours of clones that differ in sensitivity to cytostatic drugs.* Acta Pathol Microbiol Scand A, 1974. **82**(1): p. 35-40.
162. Mroz, E.A., et al., *Intra-tumor genetic heterogeneity and mortality in head and neck cancer: analysis of data from the Cancer Genome Atlas.* PLoS Med, 2015. **12**(2): p. e1001786.

163. Mroz, E.A., et al., *High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma.* Cancer, 2013. **119**(16): p. 3034-42.
164. Pereira, B., et al., *The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes.* Nat Commun, 2016. **7**: p. 11479.
165. Bozic, I. and M.A. Nowak, *Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers.* Proc Natl Acad Sci U S A, 2014. **111**(45): p. 15964-8.
166. Roskoski, R., Jr., *A historical overview of protein kinases and their targeted small molecule inhibitors.* Pharmacol Res, 2015. **100**: p. 1-23.
167. Bozic, I., B. Allen, and M.A. Nowak, *Dynamics of targeted cancer therapy.* Trends Mol Med, 2012. **18**(6): p. 311-6.
168. Van Allen, E.M., et al., *The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma.* Cancer Discov, 2014. **4**(1): p. 94-109.
169. Bozic, I., et al., *Evolutionary dynamics of cancer in response to targeted combination therapy.* Elife, 2013. **2**: p. e00747.
170. Yamani, A., et al., *Discovery and optimization of novel pyrazole-benzimidazole CPL304110, as a potent and selective inhibitor of fibroblast growth factor receptors FGFR (1-3).* Eur J Med Chem, 2021. **210**: p. 112990.
171. Yadav, S.P., *The wholeness in suffix -omics, -omes, and the word om.* J Biomol Tech, 2007. **18**(5): p. 277.
172. Yang, X., et al., *High-Throughput Transcriptome Profiling in Drug and Biomarker Discovery.* Front Genet, 2020. **11**: p. 19.
173. Green, E.D., J.D. Watson, and F.S. Collins, *Human Genome Project: Twenty-five years of big biology.* Nature, 2015. **526**(7571): p. 29-31.
174. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
175. Powell, J., *Enhanced concatemer cloning-a modification to the SAGE (Serial Analysis of Gene Expression) technique.* Nucleic Acids Res, 1998. **26**(14): p. 3445-6.
176. Brenner, S., et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays.* Nat Biotechnol, 2000. **18**(6): p. 630-4.
177. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
178. Islam, S., et al., *Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq.* Genome Res, 2011. **21**(7): p. 1160-7.
179. Eid, J., et al., *Real-time DNA sequencing from single polymerase molecules.* Science, 2009. **323**(5910): p. 133-8.
180. Howorka, S., S. Cheley, and H. Bayley, *Sequence-specific detection of individual DNA strands using engineered nanopores.* Nat Biotechnol, 2001. **19**(7): p. 636-9.
181. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
182. Marioni, J.C., et al., *RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.* Genome Res, 2008. **18**(9): p. 1509-17.
183. Miller, J.A., et al., *Improving reliability and absolute quantification of human brain microarray data by filtering and scaling probes using RNA-Seq.* BMC Genomics, 2014. **15**(1): p. 154.
184. Nagalakshmi, U., et al., *The transcriptional landscape of the yeast genome defined by RNA sequencing.* Science, 2008. **320**(5881): p. 1344-9.
185. van Dijk, E.L., et al., *Ten years of next-generation sequencing technology.* Trends Genet, 2014. **30**(9): p. 418-26.
186. Finotello, F. and B. Di Camillo, *Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis.* Brief Funct Genomics, 2015. **14**(2): p. 130-42.
187. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.

188. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.
189. Li, D., et al., *An evaluation of RNA-seq differential analysis methods.* PLoS One, 2022. **17**(9): p. e0264246.
190. Di, Y., et al., *The NBP negative binomial model for assessing differential gene expression from RNA-Seq.* Statistical Applications in Genetics and Molecular Biology, 2011. **10**(1).
191. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
192. Law, C.W., et al., *voom: Precision weights unlock linear model analysis tools for RNA-seq read counts.* Genome Biol, 2014. **15**(2): p. R29.
193. Hardcastle, T.J. and K.A. Kelly, *baySeq: empirical Bayesian methods for identifying differential expression in sequence count data.* BMC Bioinformatics, 2010. **11**: p. 422.
194. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome Biol, 2010. **11**(10): p. R106.
195. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
196. Leng, N., et al., *EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.* Bioinformatics, 2013. **29**(8): p. 1035-43.
197. Auer, P.L. and R.W. Doerge, *A two-stage poisson model for testing RNA-Seq data.* Statistical Applications in Genetics and Molecular Biology, 2011. **10**(1).
198. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
199. Van De Wiel, M.A., et al., *Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors.* Biostatistics, 2013. **14**(1): p. 113-28.
200. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nat Biotechnol, 2013. **31**(1): p. 46-53.
201. Tarazona, S., et al., *Differential expression in RNA-seq: a matter of depth.* Genome Res, 2011. **21**(12): p. 2213-23.
202. Li, J. and R. Tibshirani, *Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data.* Stat Methods Med Res, 2013. **22**(5): p. 519-36.
203. Oshlack, A., M.D. Robinson, and M.D. Young, *From RNA-seq reads to differential expression results.* Genome Biol, 2010. **11**(12): p. 220.
204. Pertea, M., et al., *Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.* Nat Protoc, 2016. **11**(9): p. 1650-67.
205. Wu, J., et al., *OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds.* Nucleic Acids Res, 2013. **41**(10): p. 5149-63.
206. Bonfert, T., et al., *ContextMap 2: fast and accurate context-based RNA-seq mapping.* BMC Bioinformatics, 2015. **16**: p. 122.
207. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
208. Yuan, L., et al., *GAAP: Genome-organization-framework-Assisted Assembly Pipeline for prokaryotic genomes.* BMC Genomics, 2017. **18**(Suppl 1): p. 952.
209. Ye, C., et al., *DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies.* Sci Rep, 2016. **6**: p. 31900.
210. Goodwin, S., et al., *Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome.* Genome Res, 2015. **25**(11): p. 1750-6.
211. Tello-Ruiz, M.K., et al., *Gramene 2016: comparative plant genomics and pathway resources.* Nucleic Acids Res, 2016. **44**(D1): p. D1133-40.
212. Stelpflug, S.C., et al., *An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development.* Plant Genome, 2016. **9**(1).

213. Yang, J., et al., *The genome sequence of allopolyploid Brassica juncea and analysis of differential homoeolog gene expression influencing selection.* Nat Genet, 2016. **48**(10): p. 1225-32.
214. Tang, W., et al., *Tumor origin detection with tissue-specific miRNA and DNA methylation markers.* Bioinformatics, 2018. **34**(3): p. 398-406.
215. Niu, S.Y., et al., *Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes.* Brief Bioinform, 2018. **19**(6): p. 1415-1429.
216. Trapnell, C., et al., *Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.* Nat Protoc, 2012. **7**(3): p. 562-78.
217. Pimentel, H., et al., *Differential analysis of RNA-seq incorporating quantification uncertainty.* Nat Methods, 2017. **14**(7): p. 687-690.
218. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
219. Sonia Tarazona, F.G., Alberto Ferrer, Joaquín Dopazo, Ana Conesa, *NOIseq: a RNA-seq differential expression method robust for sequencing depth biases.* EMBnet.journal, 2012: p. 18-19.
220. Wang, L., et al., *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.* Bioinformatics, 2010. **26**(1): p. 136-8.
221. Seyednasrollah, F., A. Laiho, and L.L. Elo, *Comparison of software packages for detecting differential expression in RNA-seq studies.* Brief Bioinform, 2015. **16**(1): p. 59-70.
222. Jiang, Z., et al., *Whole transcriptome analysis with sequencing: methods, challenges and potential solutions.* Cell Mol Life Sci, 2015. **72**(18): p. 3425-39.
223. Arowolo, M.O., et al., *A survey of dimension reduction and classification methods for RNA-Seq data on malaria vector.* Journal of Big Data, 2021. **8**(1): p. 50.
224. Xiang, R., et al., *A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data.* Front Genet, 2021. **12**: p. 646936.
225. Elssied, N.O.F., O. Ibrahim, and A.H. Osman, *A novel feature selection based on one-way ANOVA F-test for e-mail spam classification.* Research Journal of Applied Sciences, Engineering and Technology, 2014. **7**(3): p. 625-638.
226. V. Arul Kumar, N.E., *A Survey on Dimensionality Reduction Technique.* International Journal of Emerging Trends & Technology in Computer Science, 2014. **3**(6): p. 036-041.
227. Nguyen, L.H. and S. Holmes, *Ten quick tips for effective dimensionality reduction.* PLoS Comput Biol, 2019. **15**(6): p. e1006907.
228. Sun, S., et al., *Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis.* Genome Biol, 2019. **20**(1): p. 269.
229. Jain, D. and V. Singh, *Feature selection and classification systems for chronic disease prediction: A review.* Egyptian Informatics Journal, 2018. **19**(3): p. 179-189.
230. Priyanka Jindal, D.K., *A Review on Dimensionality Reduction Techniques.* International Journal of Computer Applications, 2017. **173**(2).
231. Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philos Trans A Math Phys Eng Sci, 2016. **374**(2065): p. 20150202.
232. Liebermeister, W., *Linear modes of gene expression determined by independent component analysis.* Bioinformatics, 2002. **18**(1): p. 51-60.
233. Pierson, E. and C. Yau, *ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.* Genome Biol, 2015. **16**: p. 241.
234. Ahmed, S., M. Rattray, and A. Boukouvalas, *GrandPrix: scaling up the Bayesian GPLVM for single-cell data.* Bioinformatics, 2019. **35**(1): p. 47-54.
235. Laurens van der Maaten, G.H., *Visualizing Data using t-SNE.* Journal of Machine Learning Research, 2008. **9**(86): p. 2579-2605.
236. Geoffrey Hinton, S.R., *Stochastic Neighbor Embedding.* NIPS, 2002. **15**: p. 833–840.
237. Leland McInnes, J.H., James Melville,, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* arXiv, 2018.

238. Dong, W., M. Charikar, and K. Li. *Efficient K-nearest neighbor graph construction for generic similarity measures*. in *20th International Conference on World Wide Web, WWW 2011*. 2011. Hyderabad.
239. Eraslan, G., et al., *Single-cell RNA-seq denoising using a deep count autoencoder.* Nature Communications, 2019. **10**(1): p. 390.
240. Salehi, M. and M. Roudbari, *Zero inflated Poisson and negative binomial regression models: Application in education.* Medical Journal of the Islamic Republic of Iran, 2015. **29**(1): p. 1177-1183.
241. Patro, R., et al., *Salmon provides fast and bias-aware quantification of transcript expression.* Nat Methods, 2017. **14**(4): p. 417-419.
242. Soneson, C., M.I. Love, and M.D. Robinson, *Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.* F1000Res, 2015. **4**: p. 1521.
243. Storey, J.D. and R. Tibshirani, *Statistical significance for genomewide studies.* Proc Natl Acad Sci U S A, 2003. **100**(16): p. 9440-5.
244. McCarthy, D.J. and G.K. Smyth, *Testing significance relative to a fold-change threshold is a TREAT.* Bioinformatics, 2009. **25**(6): p. 765-71.
245. The Human Protein Atlas, *The lung-specific proteome.* 2023.
246. The Human Protein Atlas, *The urinary bladder-specific proteome.* 2023.
247. The Human Protein Atlas, *The stomach-specific proteome.* 2023.
248. GeneCards®: The Human Gene Database, 2023.
249. Cen, C., et al., *Systemic analysis the expression, prognostic, and immune infiltrates significance of MS4A family in lung cancer.* All Life, 2022. **15**(1): p. 134-146.
250. Zijun Zheng, H.L., Hui Guo,, *Role of the membrane spanning 4A (MS4A) gene family in lung adenocarcinoma.* Research Square, 2022.
251. Steele, M.P., et al., *Relationship between gene expression and lung function in Idiopathic Interstitial Pneumonias.* BMC Genomics, 2015. **16**: p. 869.
252. Ji, L., et al., *RTKN2 is Associated with Unfavorable Prognosis and Promotes Progression in Non-Small-Cell Lung Cancer.* Onco Targets Ther, 2020. **13**: p. 10729-10738.
253. Wang, P., et al., *IL-23 concentration -dependently regulates T24 cell proliferation, migration and invasion and is associated with prognosis in patients with bladder cancer.* Oncol Rep, 2018. **40**(6): p. 3685-3693.
254. Bostrom, P.J., et al., *Expression of collagenase-3 (matrix metalloproteinase-13) in transitional-cell carcinoma of the urinary bladder.* Int J Cancer, 2000. **88**(3): p. 417-23.
255. Nagumo, Y., et al., *PLD1 promotes tumor invasion by regulation of MMP-13 expression via NF-κB signaling in bladder cancer.* Cancer Lett, 2021. **511**: p. 15-25.
256. Jia, Y., et al., *A comprehensive analysis of common genetic variation in MUC1, MUC5AC, MUC6 genes and risk of stomach cancer.* Cancer Causes Control, 2010. **21**(2): p. 313-21.
257. Silver, N., et al., *Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR.* BMC Mol Biol, 2006. **7**: p. 33.
258. Caracausi, M., et al., *Systematic identification of human housekeeping genes possibly useful as references in gene expression studies.* Mol Med Rep, 2017. **16**(3): p. 2397-2410.
259. Feng, J., et al., *GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data.* Bioinformatics, 2012. **28**(21): p. 2782-8.
260. Guo, Z., et al., *An empirical bayesian approach for testing gene expression fold change and its application in detecting global dosage effects.* NAR Genom Bioinform, 2020. **2**(3): p. lqaa072.
261. Garcia-Campos, M.A., J. Espinal-Enriquez, and E. Hernandez-Lemus, *Pathway Analysis: State of the Art.* Front Physiol, 2015. **6**: p. 383.
262. Pavlidis, P., et al., *Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex.* Neurochem Res, 2004. **29**(6): p. 1213-22.

263. Khatri, P., M. Sirota, and A.J. Butte, *Ten years of pathway analysis: current approaches and outstanding challenges.* PLoS Comput Biol, 2012. **8**(2): p. e1002375.
264. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
265. Ma, J., A. Shojaie, and G. Michailidis, *A comparative study of topology-based pathway enrichment analysis methods.* BMC Bioinformatics, 2019. **20**(1): p. 546.
266. Mubeen, S., et al., *On the influence of several factors on pathway enrichment analysis.* Briefings in Bioinformatics, 2022. **23**(3).
267. Hidalgo, M.R., et al., *High throughput estimation of functional cell activities reveals disease mechanisms and predicts relevant clinical outcomes.* Oncotarget, 2017. **8**(3): p. 5160-5178.
268. Yu, G. and Q.Y. He, *ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization.* Mol Biosyst, 2016. **12**(2): p. 477-9.
269. Cargnello, M. and P.P. Roux, *Activation and function of the MAPKs and their substrates, the MAPK-activated protein kinases.* Microbiol Mol Biol Rev, 2011. **75**(1): p. 50-83.
270. Castellano, E. and J. Downward, *Role of RAS in the regulation of PI 3-kinase.* Curr Top Microbiol Immunol, 2010. **346**: p. 143-69.
271. Yang, G., et al., *Glucuronidation: driving factors and their impact on glucuronide disposition.* Drug Metab Rev, 2017. **49**(2): p. 105-138.
272. Zheng, Y., et al., *Upregulated GRB7 promotes proliferation and tumorigenesis of Bladder Cancer via Phospho-AKT Pathway.* Int J Biol Sci, 2020. **16**(16): p. 3221-3230.
273. Chu, P.Y., et al., *EGF-induced Grb7 recruits and promotes Ras activity essential for the tumorigenicity of Sk-Br3 breast cancer cells.* J Biol Chem, 2010. **285**(38): p. 29279-85.
274. Chu, P.Y., Y.L. Tai, and T.L. Shen, *Grb7, a Critical Mediator of EGFR/ErbB Signaling, in Cancer Development and as a Potential Therapeutic Target.* Cells, 2019. **8**(5).
275. Vayssiere, B., et al., *Interaction of the Grb7 adapter protein with Rnd1, a new member of the Rho family.* FEBS Lett, 2000. **467**(1): p. 91-6.
276. Qin, C.D., et al., *The Rho GTPase Rnd1 inhibits epithelial-mesenchymal transition in hepatocellular carcinoma and is a favorable anti-metastasis target.* Cell Death Dis, 2018. **9**(5): p. 486.
277. Mouly, L., et al., *The RND1 Small GTPase: Main Functions and Emerging Role in Oncogenesis.* Int J Mol Sci, 2019. **20**(15).
278. Emanuele, S. and M. Giuliano, *Dual Function Molecules and Processes in Cell Fate Decision: A Preface to the Special Issue.* Int J Mol Sci, 2020. **21**(24).
279. Jeffery, C.J., *Moonlighting proteins.* Trends Biochem Sci, 1999. **24**(1): p. 8-11.
280. Bland, T., et al., *WLS-Wnt signaling promotes neuroendocrine prostate cancer.* iScience, 2021. **24**(1): p. 101970.
281. Lu, D., et al., *Wls promotes the proliferation of breast cancer cells via Wnt signaling.* Med Oncol, 2015. **32**(5): p. 140.
282. Lu, D., et al., *Golgi Phosphoprotein 3 Promotes Wls Recycling and Wnt Secretion in Glioma Progression.* Cell Physiol Biochem, 2018. **47**(6): p. 2445-2457.
283. Chai, G., et al., *A Human Pleiotropic Multiorgan Condition Caused by Deficient Wnt Secretion.* N Engl J Med, 2021. **385**(14): p. 1292-1301.
284. Yang, P.T., et al., *WLS inhibits melanoma cell proliferation through the β-catenin signalling pathway and induces spontaneous metastasis.* EMBO Mol Med, 2012. **4**(12): p. 1294-307.
285. Liu, X., et al., *Metallothionein 2A (MT2A) controls cell proliferation and liver metastasis by controlling the MST1/LATS2/YAP1 signaling pathway in colorectal cancer.* Cancer Cell Int, 2022. **22**(1): p. 205.
286. Wu, J., et al., *Cdc14B depletion leads to centriole amplification, and its overexpression prevents unscheduled centriole duplication.* J Cell Biol, 2008. **181**(3): p. 475-83.
287. Cho, Y.Y., *RSK2 and its binding partners in cell proliferation, transformation and cancer development.* Arch Pharm Res, 2017. **40**(3): p. 291-303.

288. Kawagoe, T., et al., *Essential role of IRAK-4 protein and its kinase activity in Toll-like receptor-mediated immune responses but not in TCR signaling.* J Exp Med, 2007. **204**(5): p. 1013-24.
289. Shi, X., et al., *ZCCHC14 regulates proliferation and invasion of non-small cell lung cancer through the MAPK-P38 signalling pathway.* J Cell Mol Med, 2021. **25**(3): p. 1406-1414.
290. GeneCards®: The Human Gene Database, *ANKRD28 Gene - Ankyrin Repeat Domain 28.* 2023.
291. Chan, V., et al., *The LCLAT1/LYCAT acyltransferase supports EGF-mediated phosphatidylinositol-3,4,5-trisphosphate and Akt signaling.* bioRxiv, 2023: p. 2023.01.26.524308.
292. Lokaj, M., et al., *The Interaction of CCDC104/BARTL1 with Arl3 and Implications for Ciliary Function.* Structure, 2015. **23**(11): p. 2122-32.
293. Casalou, C., A. Ferreira, and D.C. Barral, *The Role of ARF Family Proteins and Their Regulators and Effectors in Cancer Progression: A Therapeutic Perspective.* Front Cell Dev Biol, 2020. **8**: p. 217.
294. Tsai, C.-K., et al., *Cell division cycle-associated 7-like gene: A novel biomarker for adverse survival in human high-grade gliomas.* Journal of Medical Sciences, 2016. **36**(6): p. 224-228.
295. Huang, X., et al., *Identification and characterization of a novel protein ISOC2 that interacts with p16INK4a.* Biochem Biophys Res Commun, 2007. **361**(2): p. 287-93.
296. Maleki, F., et al., *Size matters: how sample size affects the reproducibility and specificity of gene set analysis.* Hum Genomics, 2019. **13**(Suppl 1): p. 42.
297. Guangchuang Yu, E.H., Chun-Hui Gao,, *enrichplot: Visualization of Functional Enrichment Result.* 2022.
298. McConnell, M., et al., *Osteoclast proton pump regulator Atp6v1c1 enhances breast cancer growth by activating the mTORC1 pathway and bone metastasis by increasing V-ATPase activity.* Oncotarget, 2017. **8**(29): p. 47675-47690.
299. Lin, J., et al., *Oncogene APOL1 promotes proliferation and inhibits apoptosis via activating NOTCH1 signaling pathway in pancreatic cancer.* Cell Death Dis, 2021. **12**(8): p. 760.
300. Song, F., et al., *The Antithetic Roles of IQGAP2 and IQGAP3 in Cancers.* Cancers (Basel), 2023. **15**(4).
301. Oglesby, I.K., et al., *miR-126 is downregulated in cystic fibrosis airway epithelial cells and regulates TOM1 expression.* J Immunol, 2010. **184**(4): p. 1702-9.
302. Pidugu, V.K., et al., *Emerging Functions of Human IFIT Proteins in Cancer.* Frontiers in Molecular Biosciences, 2019. **6**.
303. Gene Cards, *SSRP1 Gene - Structure Specific Recognition Protein 1.* 2023.
304. Huang, J., et al., *TFDP3 as E2F Unique Partner, Has Crucial Roles in Cancer Cells and Testis.* Frontiers in Oncology, 2021. **11**.
305. Dynlacht, B.D., *E2F and p53 make a nice couple: converging pathways in apoptosis.* Cell Death & Differentiation, 2005. **12**(4): p. 313-314.
306. Maurizi, A., et al., *Extra-skeletal manifestations in mice affected by Clcn7-dependent autosomal dominant osteopetrosis type 2 clinical and therapeutic implications.* Bone Res, 2019. **7**: p. 17.
307. Armijo, M.E., et al., *Rheb signaling and tumorigenesis: mTORC1 and new horizons.* Int J Cancer, 2016. **138**(8): p. 1815-23.
308. Tang, Y., et al., *Downregulation of ubiquitin inhibits the proliferation and radioresistance of non-small cell lung cancer cells in vitro and in vivo.* Sci Rep, 2015. **5**: p. 9476.
309. Cristofani, R., et al., *The Role of HSPB8, a Component of the Chaperone-Assisted Selective Autophagy Machinery, in Cancer.* Cells, 2021. **10**(2).
310. Jia, C., et al., *Evidence of Omics, Immune Infiltration, and Pharmacogenomics for BATF in a Pan-Cancer Cohort.* Front Mol Biosci, 2022. **9**: p. 844721.
311. Feng, Y., et al., *BATF acts as an oncogene in non-small cell lung cancer.* Oncol Lett, 2020. **19**(1): p. 205-210.
312. Iyoda, M., et al., *IL-17A and IL-17F stimulate chemokines via MAPK pathways (ERK1/2 and p38 but not JNK) in mouse cultured mesangial cells: synergy with TNF-alpha and IL-1beta.* Am J Physiol Renal Physiol, 2010. **298**(3): p. F779-87.
313. Maranto, J., J. Rappaport, and P.K. Datta, *Role of C/EBP-β, p38 MAPK, and MKK6 in IL-1β-mediated C3 gene regulation in astrocytes.* J Cell Biochem, 2011. **112**(4): p. 1168-75.

314. Yuan, K., et al., *Complement C3 overexpression activates JAK2/STAT3 pathway and correlates with gastric cancer progression.* J Exp Clin Cancer Res, 2020. **39**(1): p. 9.
315. von Elsner, L., et al., *C3 exoenzyme impairs cell proliferation and apoptosis by altering the activity of transcription factors.* Naunyn Schmiedebergs Arch Pharmacol, 2016. **389**(9): p. 1021-31.
316. Bharti, R., et al., *CD55 in cancer: Complementing functions in a non-canonical manner.* Cancer Lett, 2022. **551**: p. 215935.
317. Scheffler, J.M., et al., *LAMTOR2 regulates dendritic cell homeostasis through FLT3-dependent mTOR signalling.* Nature Communications, 2014. **5**(1): p. 5138.
318. Xu, H., et al., *PDS5B inhibits cell proliferation, migration, and invasion via upregulation of LATS1 in lung cancer cells.* Cell Death Discov, 2021. **7**(1): p. 168.
319. Li, J.M., et al., *Upregulation of LGALS1 is associated with oral cancer metastasis.* Ther Adv Med Oncol, 2018. **10**: p. 1758835918794622.
320. Liu, L., et al., *A Review of ULK1-Mediated Autophagy in Drug Resistance of Cancer.* Cancers (Basel), 2020. **12**(2).
321. Hu, B., X. Zhu, and J. Lu, *Cathepsin A knockdown decreases the proliferation and invasion of A549 lung adenocarcinoma cells.* Mol Med Rep, 2020. **21**(6): p. 2553-2559.
322. Yu, P., et al., *The Function, Role and Process of DDX58 in Heart Failure and Human Cancers.* Front Oncol, 2022. **12**: p. 911309.
323. Miyahara, Y., et al., *Prosaposin, tumor-secreted protein, promotes pancreatic cancer progression by decreasing tumor-infiltrating lymphocytes.* Cancer Sci, 2022. **113**(8): p. 2548-2559.
324. Wen, Z., et al., *Pan-cancer analysis of PSAP identifies its expression and clinical relevance in gastric cancer.* Pathol Res Pract, 2022. **238**: p. 154027.
325. National Library of Medicine, *CDC37L1 cell division cycle 37 like 1, HSP90 cochaperone.* 2023.
326. Hubbi, M.E., et al., *MCM proteins are negative regulators of hypoxia-inducible factor 1.* Mol Cell, 2011. **42**(5): p. 700-12.
327. Sun, P., L. Li, and Z. Li, *RAB9A Plays an Oncogenic Role in Human Liver Cancer Cells.* Biomed Res Int, 2020. **2020**: p. 5691671.
328. Benhamou, Y., et al., *Telomeric repeat-binding factor 2: a marker for survival and anti-EGFR efficacy in oral carcinoma.* Oncotarget, 2016. **7**(28): p. 44236-44251.
329. A. Storaci, I.B., M. Caroli, S. Ferrero, V. Vaira,, *V-ATPase G1 expression in human glioma stem cells correlates with ERK activation.* BMJ, 2018. **3**(SUPPLEMENT 2).
330. Wang, X., et al., *Proteomic Profiling of Exosomes From Hemorrhagic Moyamoya Disease and Dysfunction of Mitochondria in Endothelial Cells.* Stroke, 2021. **52**(10): p. 3351-3361.
331. Huang, H., et al., *SLC15A4 Serves as a Novel Prognostic Biomarker and Target for Lung Adenocarcinoma.* Frontiers in Genetics, 2021. **12**.
332. Feijoo, C., et al., *Activation of mammalian Chk1 during DNA replication arrest: a role for Chk1 in the intra-S phase checkpoint monitoring replication origin firing.* J Cell Biol, 2001. **154**(5): p. 913-23.
333. Yang, S., et al., *Androgen receptor differentially regulates the proliferation of prostatic epithelial cells in vitro and in vivo.* Oncotarget, 2016. **7**(43): p. 70404-70419.
334. Li, Y. and E. Seto, *HDACs and HDAC Inhibitors in Cancer Development and Therapy.* Cold Spring Harb Perspect Med, 2016. **6**(10).
335. Kaur, E., R. Agrawal, and S. Sengupta, *Functions of BLM Helicase in Cells: Is It Acting Like a Double-Edged Sword?* Frontiers in Genetics, 2021. **12**.
336. Ruppender, N., et al., *Cellular Adhesion Promotes Prostate Cancer Cells Escape from Dormancy.* PLoS One, 2015. **10**(6): p. e0130565.
337. Kelsey M. Gray, A.M.D., Jessica L. Fleming, Amanda E. Toland,, *DCPS as a cutaneous squamous cell carcinoma susceptibility gene.* Cancer Research, 2012. **72**(8_Supplement): p. 103.
338. Tsai, C.H., et al., *Over-expression of cofilin-1 suppressed growth and invasion of cancer cells is associated with up-regulation of let-7 microRNA.* Biochim Biophys Acta, 2015. **1852**(5): p. 851-61.

339. Woischke, C., et al., *CYB5R1 links epithelial-mesenchymal transition and poor prognosis in colorectal cancer.* Oncotarget, 2016. **7**(21): p. 31350-60.
340. Sakamoto, S., et al., *Interferon-Induced Transmembrane Protein 1 (IFITM1) Promotes Distant Metastasis of Small Cell Lung Cancer.* Int J Mol Sci, 2020. **21**(14).
341. Elazezy, M., et al., *Emerging Insights into Keratin 16 Expression during Metastatic Progression of Breast Cancer.* Cancers (Basel), 2021. **13**(15).
342. Sritangos, P., et al. *Plasma Membrane Ca2+ ATPase Isoform 4 (PMCA4) Has an Important Role in Numerous Hallmarks of Pancreatic Cancer*. Cancers, 2020. **12**, DOI: 10.3390/cancers12010218.
343. Kamata, Y.U., et al., *Introduction of ID2 Enhances Invasiveness in ID2-null Oral Squamous Cell Carcinoma Cells via the SNAIL Axis.* Cancer Genomics Proteomics, 2016. **13**(6): p. 493-497.
344. Han, Z., et al., *Model-based analysis uncovers mutations altering autophagy selectivity in human cancer.* Nat Commun, 2021. **12**(1): p. 3258.
345. Su, C.Y., et al., *The opposite prognostic effect of NDUFS1 and NDUFS8 in lung cancer reflects the oncojanus role of mitochondrial complex I.* Sci Rep, 2016. **6**: p. 31357.
346. Gritti, I., et al., *Loss of ribonuclease DIS3 hampers genome integrity in myeloma by disrupting DNA:RNA hybrid metabolism.* EMBO J, 2022. **41**(22): p. e108040.
347. Mao, M., et al., *HJURP regulates cell proliferation and chemo-resistance via YAP1/NDRG1 transcriptional axis in triple-negative breast cancer.* Cell Death Dis, 2022. **13**(4): p. 396.
348. He, S., et al., *A comprehensive pancancer analysis reveals the potential value of RAR-related orphan receptor C (RORC) for cancer immunotherapy.* Front Genet, 2022. **13**: p. 969476.
349. Wang, H., et al., *Identification of specific susceptibility loci for the early-onset colorectal cancer.* Genome Med, 2023. **15**(1): p. 13.
350. Xu, X., et al., *PSMD7 downregulation suppresses lung cancer progression by regulating the p53 pathway.* J Cancer, 2021. **12**(16): p. 4945-4957.
351. Cui, K., et al., *EXOSC8 promotes colorectal cancer tumorigenesis via regulating ribosome biogenesis-related processes.* Oncogene, 2022. **41**(50): p. 5397-5410.
352. Cao, Y., et al., *CENPO expression regulates gastric cancer cell proliferation and is associated with poor patient prognosis.* Mol Med Rep, 2019. **20**(4): p. 3661-3670.
353. Wang, D., et al., *CCNB2 is a novel prognostic factor and a potential therapeutic target in low-grade glioma.* Biosci Rep, 2022. **42**(1).
354. Seo, J., et al., *Cdt1 transgenic mice develop lymphoblastic lymphoma in the absence of p53.* Oncogene, 2005. **24**(55): p. 8176-86.
355. Shan, Y., et al., *Targeting HIBCH to reprogram valine metabolism for the treatment of colorectal cancer.* Cell Death Dis, 2019. **10**(8): p. 618.
356. Henneke, G., E. Friedrich-Heineken, and U. Hubscher, *Flap endonuclease 1: a novel tumour suppresser protein.* Trends Biochem Sci, 2003. **28**(7): p. 384-90.
357. Mou, P.K., et al., *Aurora kinase A, a synthetic lethal target for precision cancer medicine.* Exp Mol Med, 2021. **53**(5): p. 835-847.
358. LUCKY POH WAH GOH, E.U.H.S., KEK HENG CHUA, PING-CHIN LEE,, *Reduced expression of AURKA in peripheral blood of breast cancer patients.* Journal of Biotechnology, Computational Biology and Bionanotechnology, 2018. **99**(1): p. 83–90.
359. de Carcer, G., et al., *Plk1 overexpression induces chromosomal instability and suppresses tumor development.* Nat Commun, 2018. **9**(1): p. 3012.
360. Sobierajska, K., et al., *TUBB4B Downregulation Is Critical for Increasing Migration of Metastatic Colon Cancer Cells.* Cells, 2019. **8**(8).
361. Xiang, C. and H.C. Yan, *Ubiquitin conjugating enzyme E2 C (UBE2C) may play a dual role involved in the progression of thyroid carcinoma.* Cell Death Discov, 2022. **8**(1): p. 130.
362. Cummins, J.M., et al., *Tumour suppression: disruption of HAUSP gene stabilizes p53.* Nature, 2004. **428**(6982): p. 1 p following 486.

363. Lammerhirt, L., et al., *Knockdown of Lamin B1 and the Corresponding Lamin B Receptor Leads to Changes in Heterochromatin State and Senescence Induction in Malignant Melanoma.* Cells, 2022. **11**(14).
364. Zhang, K., et al., *Upregulated TUBG1 expression is correlated with poor prognosis in hepatocellular carcinoma.* PeerJ, 2022. **10**: p. e14415.
365. He, Z.-Y., et al., *Up-Regulation of RFC3 Promotes Triple Negative Breast Cancer Metastasis and is Associated With Poor Prognosis Via EMT.* Translational Oncology, 2017. **10**(1): p. 1-9.
366. Mo, S., et al., *Down regulated oncogene KIF2C inhibits growth, invasion, and metastasis of hepatocellular carcinoma through the Ras/MAPK signaling pathway and epithelial-to-mesenchymal transition.* Annals of Translational Medicine, 2022. **10**(3): p. 151.
367. Cai, C., et al., *CDT1 Is a Novel Prognostic and Predictive Biomarkers for Hepatocellular Carcinoma.* Frontiers in Oncology, 2021. **11**.
368. Li, X., et al., *Downregulation of CCNA2 disturbs trophoblast migration, proliferation, and apoptosis during the pathogenesis of recurrent miscarriage.* Am J Reprod Immunol, 2019. **82**(1): p. e13144.
369. Zhang, P., et al., *POLE2 facilitates the malignant phenotypes of glioblastoma through promoting AURKA-mediated stabilization of FOXM1.* Cell Death & Disease, 2022. **13**(1): p. 61.
370. Tang, N., et al., *TCP1 regulates Wnt7b/β-catenin pathway through P53 to influence the proliferation and migration of hepatocellular carcinoma cells.* Signal Transduction and Targeted Therapy, 2020. **5**(1): p. 169.
371. Xie, Y., et al., *ERCC6L promotes cell growth and invasion in human colorectal cancer.* Oncol Lett, 2019. **18**(1): p. 237-246.
372. Wang, Z., et al., *PSMD12 promotes glioma progression by upregulating the expression of Nrf2.* Ann Transl Med, 2021. **9**(8): p. 700.
373. Koike, Y., et al., *TPX2 is a prognostic marker and promotes cell proliferation in neuroblastoma.* Oncol Lett, 2022. **23**(4): p. 136.
374. Chen, H., et al., *CDCA5, Transcribed by E2F1, Promotes Oncogenesis by Enhancing Cell Proliferation and Inhibiting Apoptosis via the AKT Pathway in Hepatocellular Carcinoma.* J Cancer, 2019. **10**(8): p. 1846-1854.
375. Li, Y., et al., *EML4-ALK-mediated activation of the JAK2-STAT pathway is critical for non-small cell lung cancer transformation.* BMC Pulm Med, 2021. **21**(1): p. 190.
376. Wu, S., R. Su, and H. Jia, *Cyclin B2 (CCNB2) Stimulates the Proliferation of Triple-Negative Breast Cancer (TNBC) Cells In Vitro and In Vivo.* Dis Markers, 2021. **2021**: p. 5511041.
377. Gui, S., et al., *TUBA1C expression promotes proliferation by regulating the cell cycle and indicates poor prognosis in glioma.* Biochem Biophys Res Commun, 2021. **577**: p. 130-138.
378. Sofi, S., et al., *Targeting cyclin-dependent kinase 1 (CDK1) in cancer: molecular docking and dynamic simulations of potential CDK1 inhibitors.* Med Oncol, 2022. **39**(9): p. 133.
379. Zheng, C., et al., *Upregulation of CENPM facilitates tumor metastasis via the mTOR/p70S6K signaling pathway in pancreatic cancer.* Oncol Rep, 2020. **44**(3): p. 1003-1012.
380. Zhuang, S., et al., *RRM2 elicits the metastatic potential of breast cancer cells by regulating cell invasion, migration and VEGF expression via the PI3K/AKT signaling.* Oncol Lett, 2020. **19**(4): p. 3349-3355.
381. Ghisays, F., et al., *RTEL1 influences the abundance and localization of TERRA RNA.* Nat Commun, 2021. **12**(1): p. 3016.
382. Zhu, Y., et al., *Differential effects of WRAP53 transcript variants on non-small cell lung cancer cell behaviors.* PLoS One, 2023. **18**(1): p. e0281132.
383. Pentimalli, F., et al., *RBL2/p130 is a direct AKT target and is required to induce apoptosis upon AKT inhibition in lung cancer and mesothelioma cell lines.* Oncogene, 2018. **37**(27): p. 3657-3671.
384. Yang, Y., et al., *Polo-like kinase 3 functions as a tumor suppressor and is a negative regulator of hypoxia-inducible factor-1 alpha under hypoxic conditions.* Cancer Res, 2008. **68**(11): p. 4077-85.
385. Cho, E., et al., *PPP6C negatively regulates oncogenic ERK signaling through dephosphorylation of MEK.* Cell Rep, 2021. **34**(13): p. 108928.

386. Mai, Z., et al., *Integration of Tumor Heterogeneity for Recurrence Prediction in Patients with Esophageal Squamous Cell Cancer.* Cancers (Basel), 2021. **13**(23).
387. Rouka, E., et al., *Effect of primary cilium-associated genes expression on the survival of mesothelioma patients: In silico investigation of TCGA data.* European Respiratory Journal, 2020. **56**(suppl 64): p. 1135.
388. Ren, Z., et al., *Downregulation of VRK1 reduces the expression of BANF1 and suppresses the proliferative and migratory activity of esophageal cancer cells.* Oncol Lett, 2020. **20**(2): p. 1163-1170.
389. Li, F., et al., *CCT4 suppression inhibits tumor growth in hepatocellular carcinoma by interacting with Cdc20.* Chin Med J (Engl), 2021. **134**(22): p. 2721-2729.
390. Simon, D.N. and M.P. Rout, *Cancer and the nuclear pore complex.* Adv Exp Med Biol, 2014. **773**: p. 285-307.
391. Liu, X., et al., *Downregulation of HMGA1 Mediates Autophagy and Inhibits Migration and Invasion in Bladder Cancer via miRNA-221/TP53INP1/p-ERK Axis.* Front Oncol, 2020. **10**: p. 589.
392. Wang, Y., et al., *TP53INP1 inhibits hypoxia-induced vasculogenic mimicry formation via the ROS/snail signalling axis in breast cancer.* J Cell Mol Med, 2018. **22**(7): p. 3475-3488.
393. Leal, J.F., et al., *Cellular senescence bypass screen identifies new putative tumor suppressor genes.* Oncogene, 2008. **27**(14): p. 1961-70.
394. Tanaka, T., et al., *Cancer-associated oxidoreductase ERO1-alpha drives the production of tumor-promoting myeloid-derived suppressor cells via oxidative protein folding.* J Immunol, 2015. **194**(4): p. 2004-10.
395. Fu, J., et al., *Low cyclin F expression in hepatocellular carcinoma associates with poor differentiation and unfavorable prognosis.* Cancer Sci, 2013. **104**(4): p. 508-15.
396. Chavdoula, E., et al., *CHUK/IKK-alpha loss in lung epithelial cells enhances NSCLC growth associated with HIF up-regulation.* Life Sci Alliance, 2019. **2**(6).
397. Kang, Y.K., et al., *Dual roles for coactivator activator and its counterbalancing isoform coactivator modulator in human kidney cell tumorigenesis.* Cancer Res, 2008. **68**(19): p. 7887-96.
398. Moncho-Amor, V., et al., *Role of Dusp6 Phosphatase as a Tumor Suppressor in Non-Small Cell Lung Cancer.* Int J Mol Sci, 2019. **20**(8).
399. Yao, Y., et al., *USP53 plays an antitumor role in hepatocellular carcinoma through deubiquitination of cytochrome c.* Oncogenesis, 2022. **11**(1): p. 31.
400. Qiao, J., et al., *Spred1 deficit promotes treatment resistance and transformation of chronic phase CML.* Leukemia, 2022. **36**(2): p. 492-506.
401. Shen, S., L. Zeng, and H. Huang, *Effect of Methionine on AMD1 Gene Expression in Prostate Cancer Cells.* Nutr Cancer, 2021. **73**(9): p. 1804-1815.
402. Corpuz, A.D., J.W. Ramos, and M.L. Matter, *PTRH2: an adhesion regulated molecular switch at the nexus of life, death, and differentiation.* Cell Death Discovery, 2020. **6**(1): p. 124.
403. Jiang, K., et al., *Tumor suppressor Spred2 interaction with LC3 promotes autophagosome maturation and induces autophagy-dependent cell death.* Oncotarget, 2016. **7**(18): p. 25652-67.
404. Deng, T., et al., *STAM Prolongs Clear Cell Renal Cell Carcinoma Patients' Survival via Inhibiting Cell Growth and Invasion.* Front Oncol, 2021. **11**: p. 611081.
405. Zhang, H., et al., *NCBP1 promotes the development of lung adenocarcinoma through up-regulation of CUL4B.* J Cell Mol Med, 2019. **23**(10): p. 6965-6977.
406. Aytes, A., et al., *ETV4 promotes metastasis in response to activation of PI3-kinase and Ras signaling in a mouse model of advanced prostate cancer.* Proc Natl Acad Sci U S A, 2013. **110**(37): p. E3506-15.
407. Guo, Y., et al., *Spermine synthase and MYC cooperate to maintain colorectal cancer cell survival by repressing Bim expression.* Nat Commun, 2020. **11**(1): p. 3243.

List of Figures

List of Tables

List of scientific achievements

Publications

1. Marcin Kubeczko, Dorota Gabryś, Marzena Gawkowska, Anna Polakiewicz-Gilowska, Alexander J. Cortez, Aleksandra Krzywon, Grzegorz Woźniak, Tomasz Latusek, Aleksandra Leśniak, Katarzyna Świderska, Marta Mianowska-Malec, Barbara Łanoszka, Konstanty Chomik, Mateusz Gajek, Anna Michalik, Elżbieta Nowicka, Rafał Tarnawski, Tomasz Rutkowski and Michał Jarząb. *Safety and feasibility of radiation therapy combined with CDK 4/6 inhibitors in the management of advanced breast cancer*. Cancers. 2023. 15(3), 690. Punktacja MEiN: 140. IF: 6.575.
2. Katarzyna Aleksandra Kujawa, Ewa Zembala-Nożynska, Joanna Syrkis, Alexander Jorge Cortez, Jolanta Kupryjańczyk and Katarzyna Marta Lisowska. *Microfibril associated protein 5 (MFAP5) is related to survival of ovarian cancer patients but not useful as prognostic biomarker*. Int. J. Mol. Sci. 2022. 23(24), 15994. punktacja MEiN 140. IF 6.208.
3. Marcin Zeman, Władysław Skałba, Agata Małgorzata Wilk, Alexander Jorge Cortez, Adam Maciejewski, Agnieszka Czarniecka. *Impact of renin-angiotensin system inhibitors on the survival of patients with rectal cancer*. BMC Cancer. 2022. 25;22(1):815. punktacja MEiN: 100. IF: 4.430.
4. Agnieszka Kotecka-Blicharz, Marcela Krzempek, Alexander Jorge Cortez, Małgorzata Oczko-Wojciechowska, Agnieszka Czarniecka, Ewa Nożyńska, Ewa Chmielik, Barbara Jarząb, Jolanta Krajewska. *The role of thyroid sonographic malignancy risk features, when the fine needle aspiration biopsy result turns to be indeterminate*. Endokrynologia Polska. 2022. 73(2):316-324. punktacja MEiN: 70. IF: 1.582.
5. Izabela Zarczynska, Monika Gorska-Arcisz, Alexander Jorge Cortez, Katarzyna Aleksandra Kujawa, Agata Małgorzata Wilk, Andrzej C. Skladanowski, Aleksandra Stanczak, Monika Skupinska, Maciej Wieczorek, Katarzyna Lisowska, Rafal Sadej, Kamila Kitowska. *p38 mediates resistance to FGFR inhibition in Non-Small Cell Lung Cancer*. Cells. 2021. 10(12), 3363. punkty MEiN: 140. IF: 7.666.
6. Natalia Vydra, Patryk Janus, Paweł Kuś, Tomasz Stokowy, Katarzyna Mrowiec, Agnieszka Toma-Jonik, Aleksandra Krzywon, Alexander Jorge Cortez, Bartosz Wojtaś, Bartłomiej Gielniewski, Roman Jaksik, Marek Kimmel, Wiesława Widłak. *Heat Shock*

Factor 1 (HSF1) cooperates with estrogen receptor α (ERα) in the regulation of estrogen action in breast cancer cells. eLife. 2021. 16;10:e69843. punktacja MEiN: 200. IF: 8.713.

7. Magdalena Olbryt, Marcin Rajczykowski, Wiesław Bal, Anna Fiszer-Kierzkowska, Alexander Jorge Cortez, Magdalena Mazur, Rafał Suwiński, Wiesława Widłak. *NGS Analysis of Liquid Biopsy (LB) and Formalin-Fixed Paraffin-Embedded (FFPE) Melanoma Samples Using Oncomine™ Pan-Cancer Cell-Free Assay*. genes. 2021. 12(7):1080. punktacja MEiN: 100. IF: 4.141.
8. Przemysław Soczomski, Beata Jurecka-Lubieniecka, Aleksandra Krzywon, Alexander Jorge Cortez, Stanisław Zgliczyński, Natalia Rogozik, Małgorzata Oczko-Wojciechowska, Agnieszka Pawlaczek, Tomasz Bednarczuk, Barbara Maria Jarzab. *A direct comparison of patients with hereditary and sporadic pancreatic neuroendocrine tumors: evaluation of clinical course, prognostic factors and genotype-phenotype correlations*. Frontiers in Endocrinology. 2021. 12:681013. punktacja MEiN: 100. IF: 6.055.
9. Alexander Jorge Cortez, Katarzyna Aleksandra Kujawa, Agata Małgorzata Wilk, Damian Robert Sojka, Joanna Patrycja Syrkis, Magdalena Olbryt and Katarzyna Marta Lisowska. *Evaluation of the Role of ITGBL1 in Ovarian Cancer*. Cancers. 2020. 12(9):2676. punktacja MEiN: 140. IF: 6.639.
10. Pawel Rajwa, Mikolaj Przydacz, Wojciech Krajewski, Blazej Kuffel, Piotr Zapala, Aleksandra Krzywon, Alexander J. Cortez, Bartosz Dybowski, Remigiusz Stamirowski, Marcin Jarzemski, Rafal B. Drobot, Pawel Stelmach, Krystyna Mlynarek, Mateusz Marcinek, Maciej Przudzik, Wiktor Krawczyk, Jakub Ryszawy, Dominik Choragwicki, Lukasz Zapala, Marcin Lipa, Michal Pozniak, Dawid Janczak, Szymon Słomian, Jan Łaszkiewicz, Marcel Nowak, Marcin Miszczyk, Marek Roslan, Michał Tkocz, Romuald Zdrojowy, Andrzej Potyka, Tomasz Szydełko, Tomasz Drewa, Piotr Jarzemski, Piotr Radziszewski, Marcin Slojewski, Artur Antoniewicz, Andrzej Paradysz, Piotr L. Chlosta. *Changing Patterns of Urologic Emergency Visits and Admissions during the COVID-19 Pandemic: A Retrospective, Multicenter, Nationwide Study*. Archives of Medical Science. 2020. 17(5):1262–1276. punktacja MEiN: 70. IF: 3.318.
11. Dorota Butkiewicz, Agnieszka Gdowicz-Kłosok, Małgorzata Krześniak, Tomasz Rutkowski, Aleksandra Krzywon, Alexander Jorge Cortez, Iwona Domińczyk and Krzysztof Składowski. *Association of Genetic Variants in ANGPT/TEK and VEGF/VEGFR with*

Progression and Survival in Head and Neck Squamous Cell Carcinoma Treated with Radiotherapy or Radiochemotherapy. Cancers. 2020. 12(6):1506. punktacja MEiN: 140. IF: 6.639.

12. Katarzyna Aleksandra Kujawa, Ewa Zembala-Nożyńska, Alexander Jorge Cortez, Tomasz Kujawa, Jolanta Kupryjańczyk and Katarzyna Marta Lisowska. *Fibronectin and periostin as prognostic markers in ovarian cancer*. Cells. 2020. 9(1):149. punktacja MEiN: 140. IF: 6.600.
13. Patrycja Tudrej, Katarzyna Aleksandra Kujawa, Alexander Jorge Cortez, Katarzyna Marta Lisowska. *Characteristics of in vivo model systems for ovarian cancer studies*. Diagnostics. 2019. 9(3):120. punktacja MEiN: 70. IF: 3.110.
14. Patrycja Tudrej, Katarzyna Aleksandra Kujawa, Alexander Jorge Cortez, Katarzyna Marta Lisowska. *Characteristics of in vitro model systems for ovarian cancer studies*. Oncol Clin Pract. 2019. 15(5):246-259.punktacja MEiN:20.IF:0.
15. Damian Robert Sojka, Agnieszka Gogler-Pigłowska, Natalia Vydra, Alexander Jorge Cortez, Piotr Teodor Filipczak, Zdzisław Krawczyk & Dorota Ścieglińska. *Functional redundancy of HSPA1, HSPA2 and other HSPA proteins in non-small cell lung carcinoma (NSCLC); an implication for NSCLC treatment*. Sci Rep. 2019. 9(1):14394. punktacja MEiN: 140. IF: 4.122.
16. Patrycja Tudrej, Magdalena Olbryt, Ewa Zembala-Nożyńska, Katarzyna Kujawa, Alexander Jorge Cortez, Anna Fiszer-Kierzkowska, Wojciech Pigłowski, Barbara Nikiel, Magdalena Głowala-Kosińska, Aleksandra Bartkowska-Chrobok, Andrzej Smagur, Wojciech Fidyk, Katarzyna Marta Lisowska. *Establishment and Characterization of the Novel High-Grade Serous Ovarian Cancer Cell Line OVPA8*. Int. J. Mol. Sci. 2018. 19(7):2080. punktacja MEiN: 140. IF: 4.183.
17. Alexander Jorge Cortez, Patrycja Tudrej, Katarzyna Aleksandra Kujawa, Katarzyna Marta Lisowska. *Advances in ovarian cancer therapy*. Cancer Chemother Pharmacol. 2018. 81(1):17-38. punktacja MEiN: 100. IF: 3.008.
18. Katarzyna Marta Lisowska, Magdalena Olbryt, Sebastian Student, Katarzyna Aleksandra Kujawa, Alexander Jorge Cortez, Krzysztof Simek, Agnieszka Dansonka-Mieszkowska, Iwona Krystyna Rzepecka, Patrycja Tudrej, Jolanta Kupryjańczyk. *Unsupervised analysis reveals two molecular subgroups of serous ovarian cancer with distinct gene expression*

profiles and survival. J Cancer Res Clin Oncol. 2016. 142(6):1239-52. punktacja MEiN: 100. IF: 3.735.

19. Katarzyna Lisowska, Alexander Jorge Cortez. *Kontrowersje wokół długoterminowych badań Gillesa-Erica Seraliniego nad bezpieczeństwem zdrowotnym kukurydzy GMO*. Studia Ecologiae et Bioethicae. 2014. 12(3):33-54. punkty MEiN: 20. IF: 0.

I am the author and co-author of 66 conference proceedings, including 31 national and 35 international. In 20, I am the first author, and I have presented at scientific conference the results of my work as an oral presentation three times.

Participation in research projects

COI Grant for Young Scientist awarded by Head of Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch entitled. "Molecular and biochemical prediction of radioresistance in patients with prostate cancer" (supervisor: A. J. Cortez).

Harmonia 2012/04/M/NZ2/00133: Biological interactions of ovarian cancer cells with omental derived adipose stem cells (O-ASC) (supervisor: Katarzyna Lisowska, funded under grant NCN) – project contractor.

Strategmed II/266766/17/NCBR/2015: clinical multicenter project CELONKO: Development of modern biomarkers and development of an innovative FGFR kinases inhibitor. (Supervisor: Katarzyna Lisowska, funded under the STRATEGMED grant awarded by NCBiR). Project leader: pharmaceutical company CelonPharma – project contractor.

Preludium 2016/21/N/NZ5/01917: To investigate the relationship between the chaperone protein HSPA2 and the products of normal and mutant TP53 gene in non-small cell lung cancer. (supervisor: Damian Sojka, funded under grant Preludium 11 NCN) – project contractor.

ABM 2019/ABM/01/00044: PALG ALL7 "OVERALL" „Efficacy and safety of obinutuzumab versus rituximab in combination with chemotherapy for adult patients with newly diagnosed CD20-positive acute lymphoblastic leukemia" (supervisor: Sebastian Giebel, funded under grant ABM) – member of Independent Data Monitoring Committee (IDMC).

ABM 2019/ABM/01/00043: CLARA „Total Body Irradiation and Cladribine Before Allogeneic Hematopoietic Cell Transplantation in Patients With AML (Acute Myeloid Leukemia)

and Myelodysplastic Syndromes" (supervisor: Sebastian Giebel, funded under grant ABM) – member of Independent Data Monitoring Committee (IDMC).

ABM 2019/ABM/01/00066: NIVONASO „Phase II study evaluating the efficacy of Nivolumab in the treatment of patients with nasopharyngeal cancer who progressed during or after Platinum-based chemotherapy" (supervisor: Tomasz Rutkowski, funded under grant ABM) – member of Independent Data Monitoring Committee (IDMC).

ABM 2020/ABM/03/00014-00: „Centrum Wsparcia Badań Klinicznych Narodowego Instytutu Onkologii Oddziału w Gliwicach" (supervisor: Magdalena Markowska, funded under grant ABM) – project contractor.

Lider 2021 0242/L-12/2020: „Technology to transform micro and nanofibers obtained in an electrostatic field into innovative objects with defined diameters and properties" (Supervisor: dr inż. Andrzej Hudecki, funded under NCBiR grant) – project contractor.