Abstract

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].

In in-silico biomarker search studies that rely on data from high-throughput experiments, pre-selecting 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].

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.

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 1) [5].

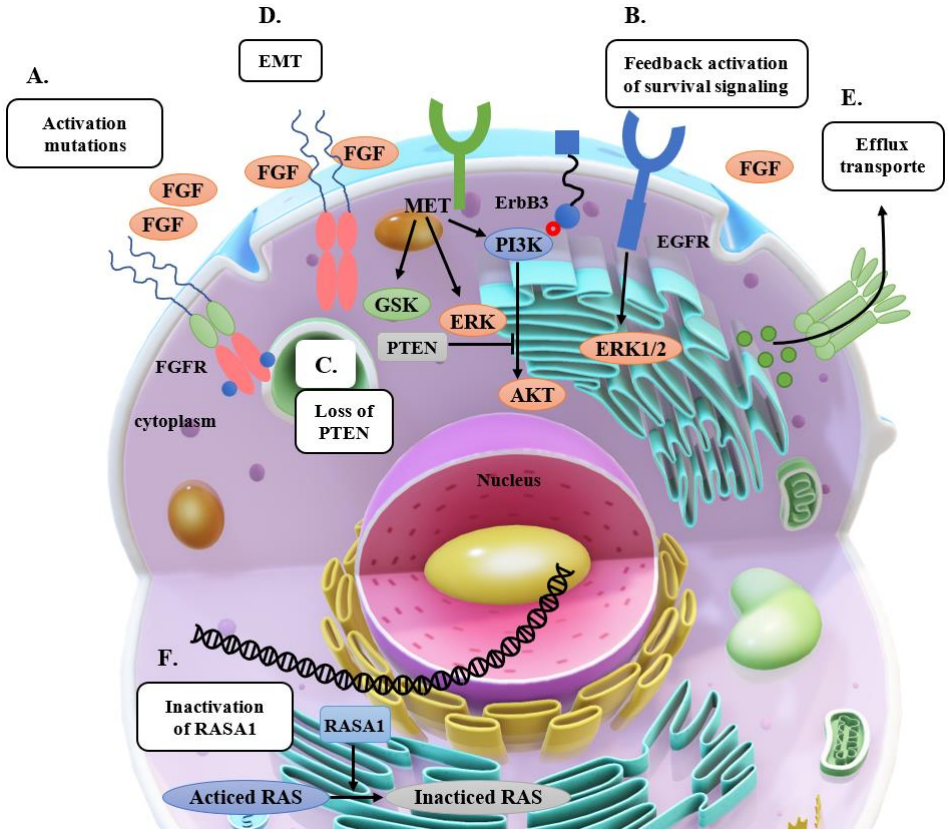

Figure 1. Mechanisms of resistance to FGFR inhibitors: (A) gatekeeper mutations in the FGFR kinase domain, (B) activation of alternate signaling pathways like EGFR, ERBB3, or MET, (C) loss of PTEN leading to increased activation of PI3K-AKT, (D) the epithelial-mesenchymal transition (EMT) may lead to 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 [6].

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 [7], 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.

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 [8]. 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) [7] 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 (Figura 2). These three types of cancer were chosen because FGFR aberrations are most commonly observed in them [9, 10]. 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.

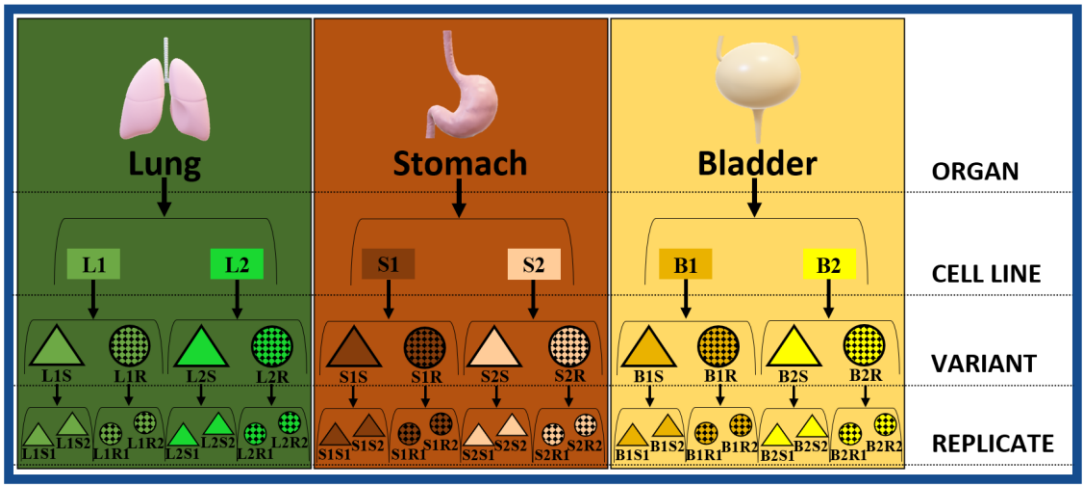

Figure 2. The RNA-seq experimental setup utilized 6 cell lines (L1: NCI-H1581, L2: NCI-H1703, S1: SNU-16, S2: KATO III, B1: RT-112, and B2: UM-UC-14), each in two variants (wild type sensitive (S: L1S, L2S, S1S, S2S, B1S, B2S) to the FGFR inhibitor and resistant cell line (R: L1R, L2R, S1R, S2R, B1R, B2R)), and two biological replicates per variant, resulting in a total of 24 experimental samples.

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 1, Figure 2). 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 1). 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 2). To mimic intra-patient diversity, two biological replicates were used for each cell line (Figure 2).

Table 1. 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

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.

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 (Figure 2).

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.

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. For example, many results did not have a consistent direction of change but rather were random as shown example in Figure 3.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 2), had a difference too small to reach the detectable threshold technics used in daily diagnostic clinical practice (Figure 3.B). Furthermore, some results, despite having a sufficiently large difference, had a small minimal fold change, and as a consequence, the biological effect could be undetectable for such a detected difference [11, 12] (Figure 3.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.

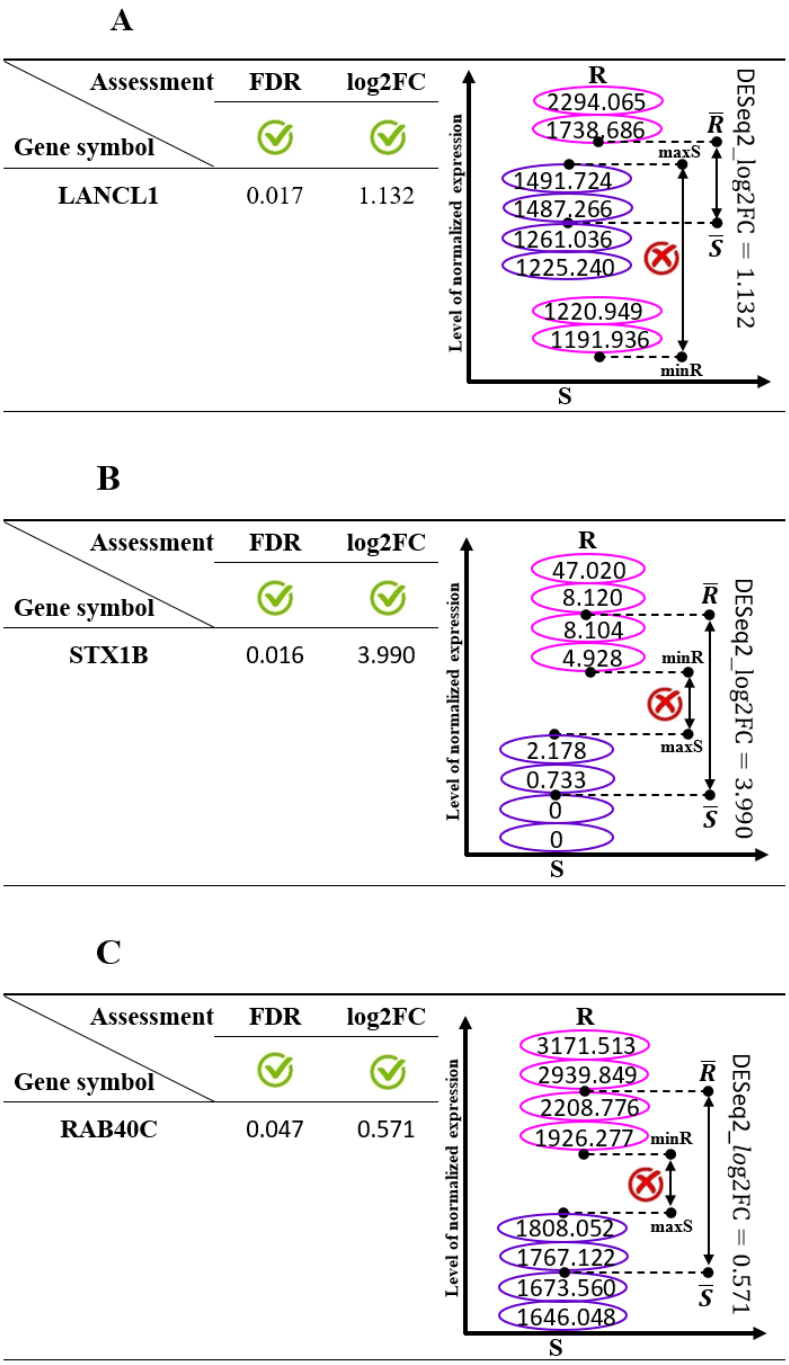

Figure 3. Examples of candidate biomarkers identified by the standard RNA-seq data analysis pipeline with significant q value (FDR) < 0.050, and proper fold change (log2FC) value > 0.500, but lacking the characteristics that a biomarker should possess: (A) LANCL1 is an example with no consistent direction 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.

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 (Figure 4), 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.

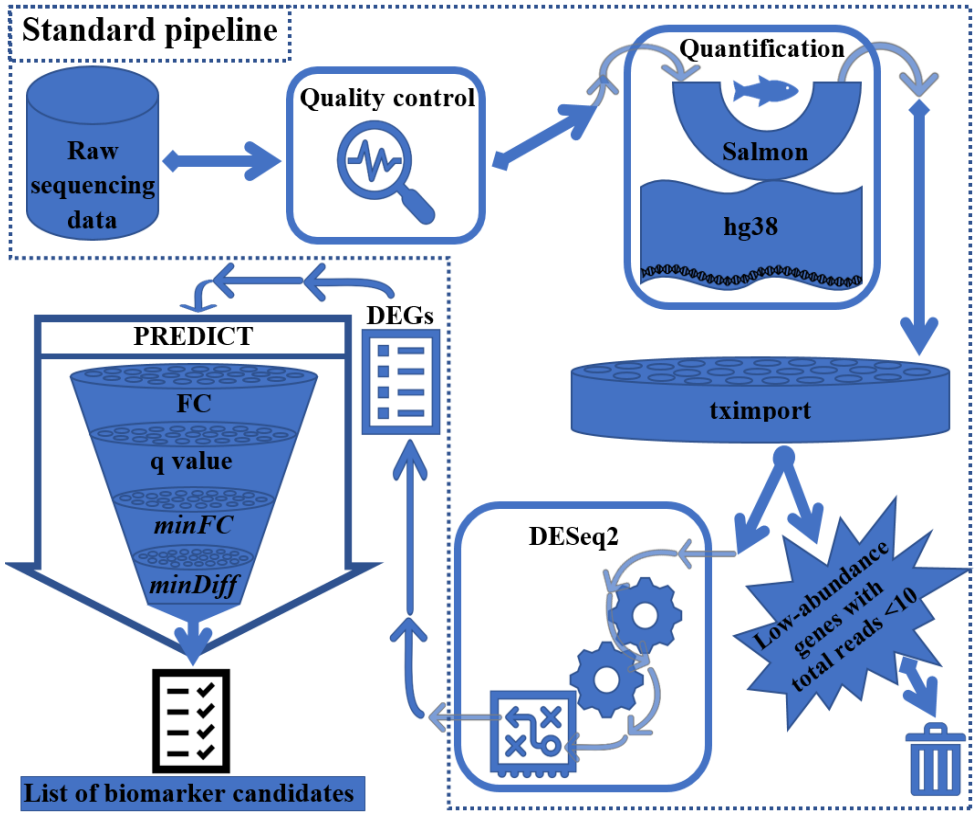

Figure 4. Scheme of the standard pipeline for RNA-seq data analysis, and scheme of Pipeline for Rapid Evaluation and Discovery of Important biomarker CandidaTes (PREDICT).

The PREDICT pipeline includes two measures, namely the minimal Fold Change (*minFC*), and the minimal Difference (*minDiff*) (Figure 5. A, and B, respectively).

minFC (minimal Fold Change) (Figure 5.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:

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

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 [11, 12]. Genes with the log2*minFC* value below the threshold are filtered out.

minDiff (minimal Difference) (Figure 5.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 2)) for proteins that are already recognized as specific to a given organ, in our case, the lung, stomach, and bladder), was evaluated. Information about such proteins was obtained from the Human Protein Atlas [13-15], and in most cases additionally evaluated in GeneCards®: The Human Gene Database [16]. 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. Genes with the *minDiff* value below the threshold are filtered out.

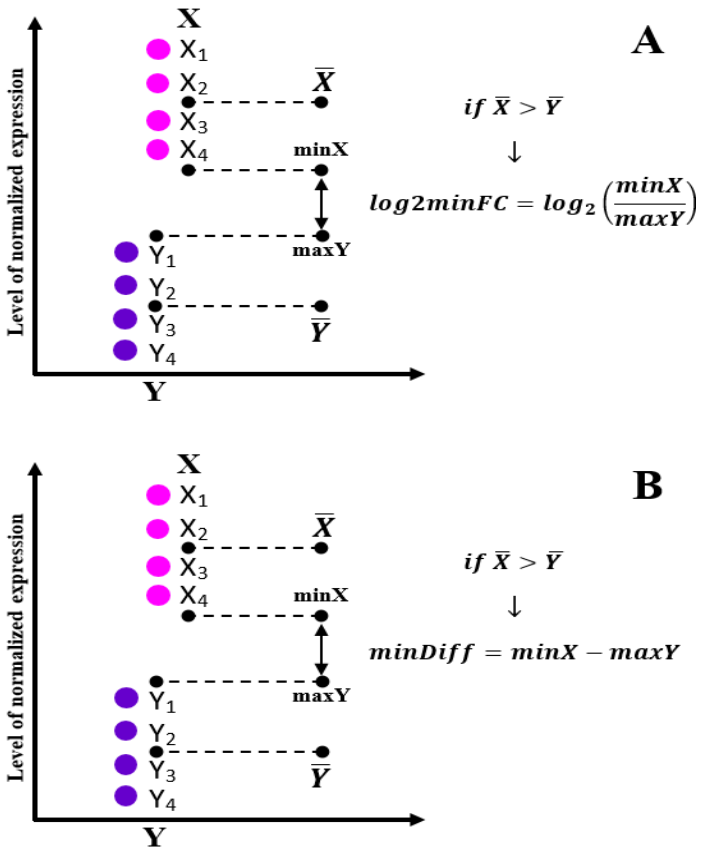

Figure 5. Scheme of the (A) *minFC*, and (B) *minDiff* measures.

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.

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 (Figure 6).

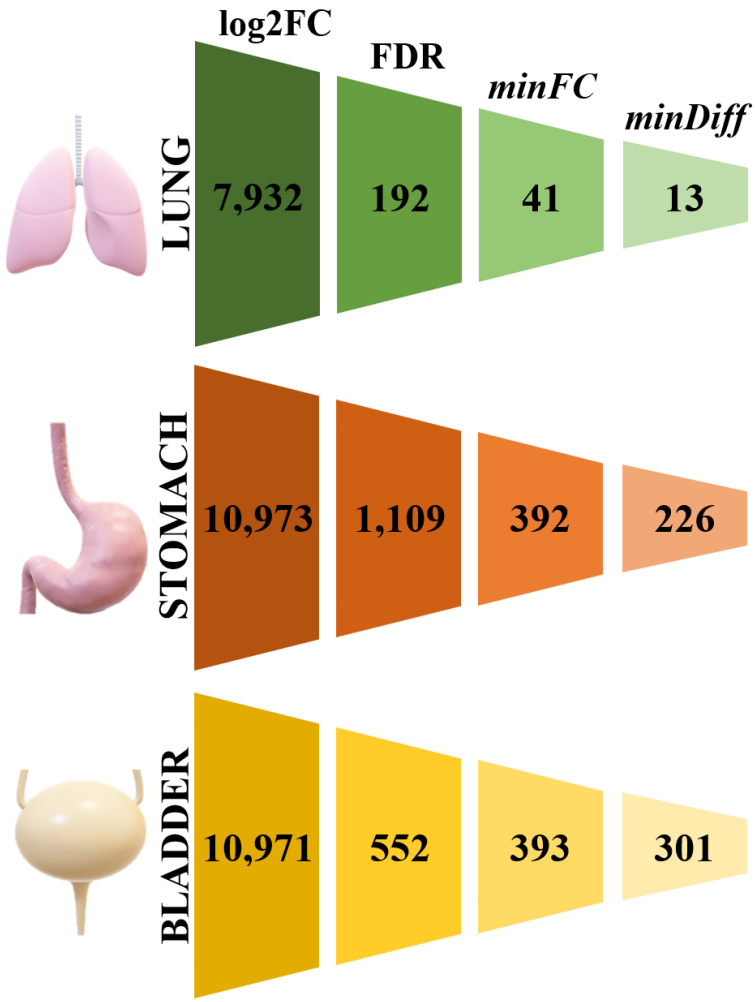

Figure 6. Number of identified DEGs obtained by sequentially applying measures thresholds: log2 fold change (log2FC) < 0.500, q value (FDR) < 0.050, log2*minFC* > 0.100, and *minDiff* > 100.

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.

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. Zarczynska, I., et al., *p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer.* Cells, 2021. **10**(12).
6. Zhou, Y., et al., *FGF/FGFR signaling pathway involved resistance in various cancer types.* J Cancer, 2020. **11**(8): p. 2000-2007.
7. 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.
8. Mroz, E.A. and J.W. Rocco, *The challenges of tumor genetic diversity.* Cancer, 2017. **123**(6): p. 917-927.
9. 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.
10. Wu, Y.M., et al., *Identification of targetable FGFR gene fusions in diverse cancers.* Cancer Discov, 2013. **3**(6): p. 636-47.
11. 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.
12. 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.
13. The Human Protein Atlas, *The lung-specific proteome.* 2023.
14. The Human Protein Atlas, *The urinary bladder-specific proteome.* 2023.
15. The Human Protein Atlas, *The stomach-specific proteome.* 2023.
16. GeneCards®: The Human Gene Database, 2023.