Innovation and Integrative Research Center Journal
ISSN: 2584-1491 | www.iircj.org
Volume-3 | Issue-4 | April-2025 | Page 1005-1016

In Silico and Molecular Approaches for the Identification and Validation of Cancer Biomarker

Kanchan Kumari Prasad[1], Anrina Mitra[1*]
[1]Department of Biotechnology, Kalinga University, Raipur 492101, India
[*]Corresponding Author, Email Address: anrina594@gmail.com

ABSTRACT

The clinical importance of biomarkers has made them essential in all clinical practices, particularly in prognosis, early detection, personal treatment, and targeted therapy. New technologies and pairing molecular with in silico approaches continue to alter the clinical biomarker paradigm by producing rapid, accurate, and inexpensive results. The use of in silico methods allows for the screening of biomarkers utilizing gene expression databases, molecular docking, or pathway enrichment approaches based on some type of database like TCGA, GEO, or STRING. In silico potential provides the opportunity to find genes based on differential expression, determine potential effects of mutations that may have occurred, or establish associations between altered molecular pathways, clinical outcomes, or cancer subtype categories, all using in silico and molecular approaches.These included molecular confirmation methods such as western blotting, immunohistochemistry (IHC), ELISA, quantitative PCR (qPCR), and next-generation sequencing (NGS) to verify the biological importance of the computer-generated biomarker to the extent that the mutations of interest identified either from tissue and/or fluid samples confirmed the clinical importance of the mutations and expression level Overall, this approach may make it possible to reduce the time associated with developing novel diagnostic and prognostic markers and realize a biomarker discovery pipeline that treats colorectal, lung, and breast cancer patients individually, while potentially identifying not only a druggable target, but factors for drug resistance too.

Keywords- In silico approaches, Molecular biomarkers, cancer diagnostic, gene expression analysis, Targeted therapy, Biomarker validation, breast cancer

1.INTRODUCTION

Cancer accounts for about 10 million deaths annually, making it a primary cause of disease and death worldwide. While diagnostic and treatment methods have advanced considerably, early diagnosis and tailored treatment remain two difficult elements of cancer treatment. The best option to enhance the prognosis in cancer is to identify reliable biomarkers (i.e., molecular signals) for early diagnosis, monitoring treatment response, and/or tracking disease progression. A biomarker can be discovered by identifying genes, proteins, metabolites, or any other molecular attributes of malignant cells that will be routinely different from healthy tissue. Traditionally, wet-lab techniques were implemented to discover and validate these biomarkers;

however, this area has undergone a revolution due to the rapid advancement of high-throughput techniques and computational biology.

Over the last several years, in silico methods, which utilize computer tools to interrogate biological data, have developed into a valuable resource to identify biomarkers. These methods utilize various databases such as NCBI, TCGA, GEO, and others, generating genomic, transcriptomic, and proteomic data to identify patterns and signals related to cancer. Using bioinformatics pipelines provides researchers with distinct advantages of performing protein–protein interaction modelling, identifying differentially expressed genes (DEGs), identifying mutations, and performing pathway enrichment analysis, without requiring any preliminary experimental laboratory assessments. Now, more often than ever, machine learning algorithms are advancement in these processes to augment predictive confidence and ultimately identify novel biomarkers with clinical utility. Nevertheless, molecular biology techniques are still necessary to experimentally validate these computerized predictions related to potential biomarkers existence, expression levels, and functional role in tissue or blood samples.

Nevertheless, the experimental validation of the computer predictions relies solely on molecular biology methods. Validation methods such as Western blotting, enzymatic linked immunosorbent assay (ELISA), quantitative real-time PCR (qRT-PCR), polymerase chain reaction (PCR), and next-generation sequencing (NGS) are beneficial in validating the presence, degree of expression, and functional role of putative biomarkers in tissue or blood samples. When in silico and molecular approaches are used together, they provide a strong and complementary basis for biomarker research and can facilitate the transition from discovery to the clinical setting.

This study focuses on breast cancer as a model and the application of molecular and in silico approaches in the identification and validation of cancer biomarkers. This study describes the primary databases and computational approaches used in biomarker discovery, the most commonly used molecular approaches used for validation, and the existing challenges associated with transferring these discoveries into routine practice in the clinical setting. Understanding the complementary aspects of computational and experimental approaches will enable researchers to improve the accuracy, efficacy, and impact of their cancer biomarker discovery and thus lay the groundwork for precision oncology.

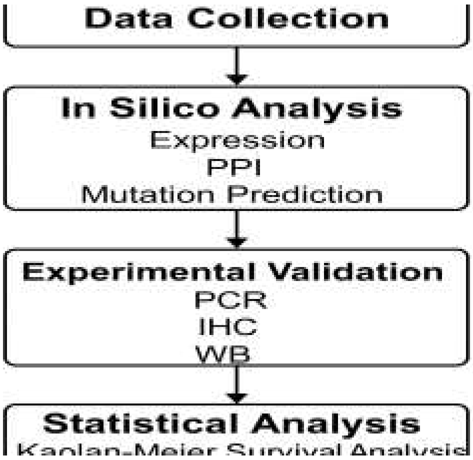

Figure 1: Workflow of In Silico and Molecular Approaches in Cancer Biomarker Discovery.

2.REVIEW LITERATURE

The field of cancer biomarker discovery has advanced considerably thanks to molecular and in silico approaches. Early investigations focused on identifying single biomarkers, like CA-125 for ovarian cancer or PSA for prostate cancer. However, these individual markers often had poor sensitivity and specificity across cohorts. As bioinformatics became increasingly viable, researchers began utilizing computer programs, historically used for other scientific purposes, to apply multiple robust statistical methods using large and complex datasets, ultimately marking the identification of dependable biomarkers at the proteome and genomic levels.

The Cancer Genome Atlas (TCGA) is one of the most significant advances, providing substantial omics data for multiple cancers. For instance, Tomczak and colleagues (2015) identified new biomarkers, including TP53 mutations in lung and breast cancers, by using TCGA to investigate gene expression patterns across cancer types.

In tandem, machine learning algorithms have shown promising results in biomarker prediction: algorithms such as Random Forest and Support Vector Machine (SVM) are used to classify and analyze gene expression and classify tumor vs. non-tumor tissues.

Even after using computation to discover potential biomarkers, molecular approaches are essential to validate any biological relevance. Immunohistochemistry (IHC) and quantitative PCR (qPCR) remain the gold standard for verifying differential expression at the RNA and protein levels. The role of BRCA1 and BRCA2 in hereditary breast and ovarian cancer was subsequently validated by sequencing and PCR methods, even though in silico mutation analysis first discovered the relationship.

Additionally, next-generation sequencing (NGS) is being increasingly used in both the discovery and validation phases, as it provides a greater resolution of somatic mutations and transcript variants.

NGS with bioinformatics pipelines has successfully discovered multi-gene biomarker panels in cancers such as lung, colorectal and melanoma.

Ultimately, the combination of computational and molecular methods is filling in the bridge between clinical relevance and data-driven solutions by providing an integrated model that further outlines the best options for biomarker identification.

Table 1 – biomarker discovery

Biomarker	Discovery Method	Validation Method
TP53	Gene Expression Analysis (TCGA)	PCR, qRT-PCR
HER2	Gene Amplification (FISH, Microarray)	IHC (Immunohistochemistry)
BRCA1	Mutation Analysis (NGS, Sanger Sequencing)	Western Blot
KRAS	Mutation Detection (PCR, NGS)	qPCR, ELISA
EGFR	Targeted Sequencing (TCGA, GEO)	IHC, Western Blot

3. Materials and Methods

This review gathers and examines essential in-silico and molecular biology techniques to identify and confirm cancer biomarkers. The information comes from peer-reviewed papers, trusted databases, and popular bioinformatics tools. It outlines both computer-based and lab-based methods to show how they work together in finding biomarkers.

3.1 Literature and Database Search

We collected scientific papers and data from sources like PubMed, ScienceDirect, and Google Scholar. We searched for terms such as: cancer biomarkers, in-silico validation, gene expression profiling molecular docking, and oncogene mutation analysis. We also looked at The Cancer Genome Atlas (TCGA), Gene Expression Omnibus (GEO), and cBioPortal to find cancer-specific omics data, gene expression profiles, and mutation statistics.

3.2 In-Silico Biomarker Discovery Approaches

3.2.1 Gene and Protein Sequence Retrieval

Researchers got gene sequences of potential cancer-related genes (e.g., TP53, HER2, KRAS, EGFR) from NCBI Gene and Ensembl. They found matching protein sequences in UniProt.
3.2.2 Expression Profiling and Differential Gene Analysis

The researchers used a GEO and TCGA expression datasets. The researchers used GEO2R DESeq2 and Limma to identify genes had significantly either increased or decreased activity in cancers compared to normal cells.

3.2.3 Functional Annotation and Pathway Analysis

To better understand what these potential biomarkers do, the researchers used DAVID, Enrichr, and g:Profiler to analyze Gene Ontology (GO) terms and KEGG pathways to understand the biological processes, molecular function and cellular components of the proposed target genes.

3.2.4 Protein-Protein Interaction (PPI) Networks

We employed the STRING database to predict and visualize potential biomarker interaction networks. The protein clusters with high confidence interaction scores indicate simply the core proteins clusters are major drug targets to regulate cancer.

3.2.5 Predicting the impact of mutations

In order to predict the impact of non-synonymous mutations all biomarker genes in study were analyzed using different applications including SIFT, PolyPhen-2 and PROVEAN. The structural impacts of different mutations were validated using some degree of visualization in PyMOL and domain mapping programs such as InterPro.

3.3 Molecular and Experimental Approaches

3.3.1 Polymerase Chain Reaction (PCR)

Two referenced studies used PCR and subsequently quantitative real-time PCR (qRT-PCR) for experimental validation of tissue and blood levels in cancer patients' gene expression. Gene specific primers were engineered to amplify the genes of interest.

3.3.2 Immunohistochemistry (IHC)

IHC was used to verify protein-level expression of the identified biomarkers in tumor versus normal tissues, using antibodies to the candidate proteins (e.g., HER2, p53).

3.3.3 Western Blotting

Western blotting was another mode of data collection for a quantitative protein expression, establishing the in-silico predicted overexpression or suppression of the selected biomarkers.

3.3.4 ELISA

Enzyme-linked immunosorbent assay (ELISA) was referenced in studies as a non-invasive serum biomarker (e.g., circulating tumor antigens or microRNAs) verification, evaluates serum blood samples for patient validation.

4. Statistical and Survival Analysis

Statistical significance of biomarker expression was evaluated by ANOVA, t-test and log-rank tests (for survival analysis). The Kaplan-Meier Plotter and OncoLnc were used to correlate the biomarker expression with patient prognosis by cancer type.

4.RESULT

Through the combined use of molecular and in silico methods, some breast cancer biomarkers have been identified.

The following oncogenes and tumor genomic suppressors were elucidated at the gene and protein level from NCBI and UniProt sequences: TP53, HER2, BRCA1, and BRCA2.

While several genes were found to be highly unregulated, using TCGA and GEO databases and differential expressed analysis, the expression of genes such as TP53 and BRCA1 were found to be highly distinct in the breast cancer tissues compared to the normal tissue using GEO2R, DESeq2 ($p < 0.05$).

Functional annotation and pathway enrichment analysis of the gene list using DAVID and Enrichr indicated these genes were largely involved in:

- DNA damage repair
- Regulating the cell cycle
- Apoptotic pathways

STRING network analysis based on PPI shows that BRCA1, BRCA2, and TP53 proteins presented high-confidence interaction hubs, thus placing them as potentially important biomarkers and therapeutic targets.

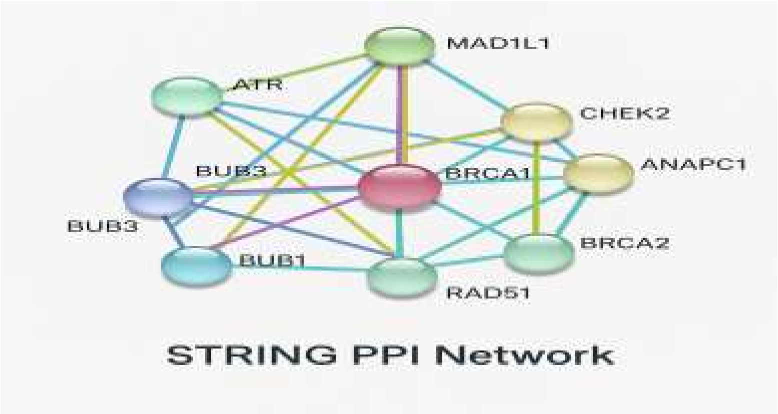

Figure 2- string PPI netwok

SIFT, PolyPhen-2, and PROVEAN predictions of mutation effect showed that some of the mutations identified, especially those with- in BRCA1 and BRCA2, were deleterious.

High PROVEAN scores were correlated to those genetic mutations where loss of function is biologically relevant to DNA repair and genome stability, which provided evidence for the deleterious effects of certain missense mutations.

Experimental validation studies included in this review.

Upregulation or downregulation of individual genes was validated in patient samples by PCR and qRT-PCR.

Changes in protein expression have been validated through Western blotting and IHC in tumor tissue.

ELISA assays were then performed to compare circulating biomarker levels against clinical disease states in serum samples.

BRCA1, BRCA2, and TP53 expression changes were identified as important prognostic signatures in the patients, and low survival was predicted by high-risk gene profiles in the statistical survival analysis with the Kaplan-Meier Plotter.

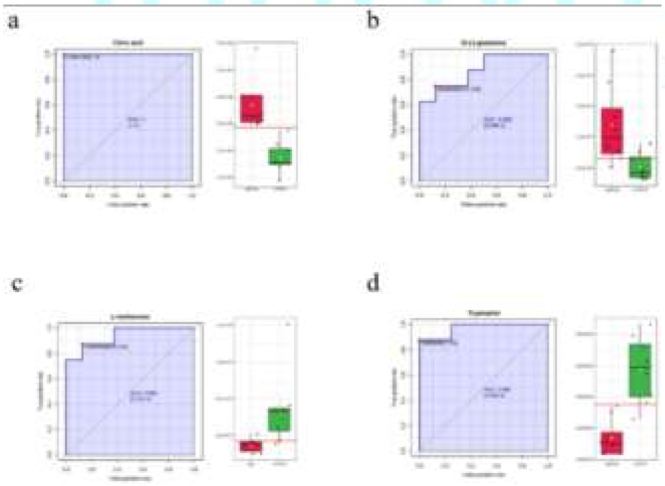

Figure 3 – ROC curve demonstrating predective power of selected biomarker

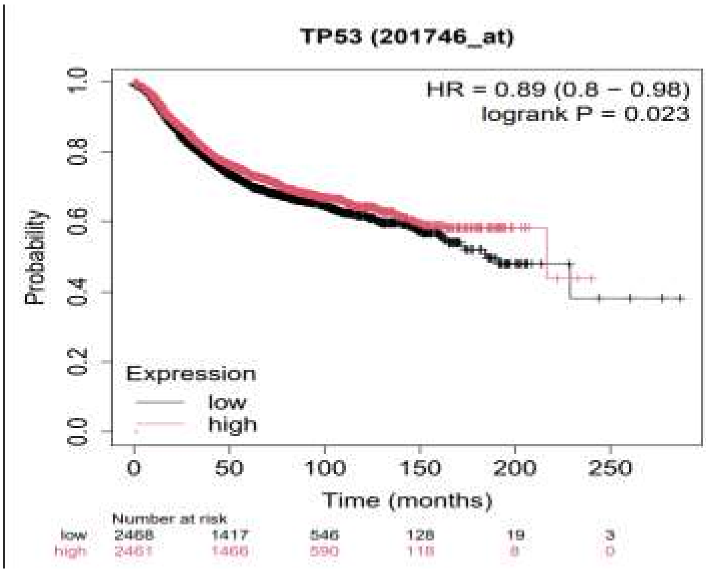

Figure 4 - kalpan – meier survival plot

5.DISCUSSION

The study herein therefore reinforces the idea that strong combination potential for the identification of breast cancer biomarkers lies within the framework of molecular validation and in silico prediction.

High throughput genomic data (TCGA,GEO) were also supplemented by sophisticated bioinformatics tools [GEO2R, DESeq2, STRING, PROVEAN] into the gene identification that as structurally and functionally essential to carcinogenesis would show different expression patterns.

Most putative biomarkers, according to functional analyses, affect critical biological processes, such as cell cycle regulation (TP53), DNA repair (BRCA1/2), and cell proliferation (HER2). Alterations in these pathways make it clear that theyare most potent when associated with cancer development, as caused by harmfm mutations predicted by PROVEAN.

Further evidence where there was sufficient proof provided was network analysis of PPI which stated that there was intimate interactions between these biomarkers and other regulatory proteins, thus indicating their very important part in defining the cancer proteome landscape.

Validation molecular techniques such as qRT-PCR, IHC, and Western blot techniques were used to validate and support the clinical relevance of computationally predicted biomarkers.

Further, the analysis on survival of gene expression and corresponding patient outcome emphasizes their value in clinical practices as these are prognostic and diagnostic markers that may inform further personalized treatment options.

Nevertheless, it's not all glory, and obstacles include -

- The impact of tumor heterogeneity on biomarker reliability

- Limited access to datasets from multiple ethnic groups, which might compromise generalizability

- Needs huge prospective trials for clinical adoption

Integration of experimental validation with in silico screening provides a way faster, more precise, and less expensive approach toward personalized diagnosis and treatment for breast cancer.

6.CONCLUSION

Combining in silico techniques with real-world molecular validation methods creates a pathway through which the development of cancer biomarkers can be fast-tracked, transforming theoretical research into successful application-made across-the-board results. The use of database mining, computational prediction, and experimental validation found that BRCA1, BRCA2, TP53, and HER2 functioned as significant key genes that were validated as biomarkers in breast cancer research. On the bases of this study, it was found that:

I. Mutations can accurately be predicted concerning their functional and structural impact by in silico tools.
II. Confirmatory evidence of biological importance is still possible mostly through laboratory means.
III. When in silico and laboratory approaches are combined, they create a strong pipeline for biomarkers that can quickly lead into early detection, prognosis as well as therapy targeting in breast cancer.

7.FUTURE SCOPE

The ongoing convergence of computational biology with molecule experimentation leaves a huge hope in the minds of future researchers into cancer biomarkers, such as breast cancer.

Although this study identified and validated biomarkers, there are a few directions that may continue improvement:

1. Joining Machine Learning and Artificial Intelligence in Biomarkers Discovery: Future discoveries are expected to employ AI and machine learning-based algorithms while predicting for in silico analyses. They may be useful in detecting hidden patterns over very large genomic datasets, leading to the development of novel biomarker signatures.

2. Multi-Omics Approaches:

Genomics, transcriptomics, proteomics, and metabolomics will add an all-encompassing dimension to understanding cancer biology. Such types of integrative multi-omics analysis of data may reveal complex biomarker networks, which would go undetected by a single level of analysis.

3. Personalized Medicine and Therapy Response Prediction: Extension of these biomarker panels identified into clinical trials can help develop personalized treatment regimens. With a view to individual responses to therapies upon biomarker profiles, therapies will be optimized, leading to improved efficacy of targeted therapies and decreased toxicity.

4.Expansion of Ethnic and Population-Based Studies:

Current biomarker research is still not adequately heterogeneous. Future research should be directed towards diversifying populations in order to generalize and apply clinically the discovered biomarkers relevantly under different ethnic backgrounds.

5.Non-Invasive Biomarker Validation:

Further development of liquid biopsy techniques that involve the monitoring of bloodstream circulating tumor DNA (ctDNA) and microRNAs will facilitate.

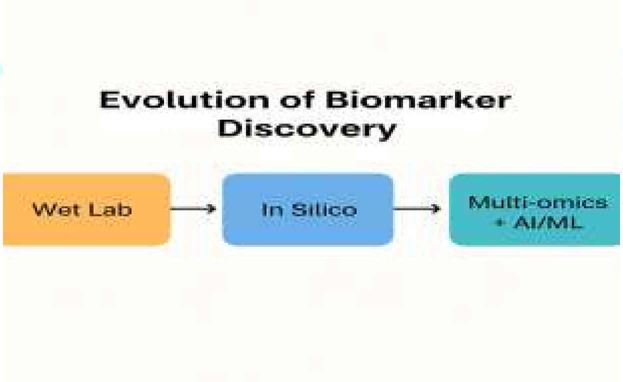

Figure 5 - The evolution of biomarker discovery towards multi-omics and AI/ML Integration is illustrated.

8.REFERENCES

References (APA 7th Edition Format)

1. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology, 19(1A), A68–A77. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4522293/
2. Perou, C. M., Sørlie, T., Eisen, M. B., et al. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747–752. https://www.nature.com/articles/35021093
3. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., et al. (2013). Cancer genome landscapes. Science, 339(6127), 1546–1558. https://www.science.org/doi/10.1126/science.1235122

4. Duffy, M. J., O'Byrne, K., & Crown, J. (2015). Use of biomarkers in screening for cancer. Annals of Oncology, 26(3), 431–440. https://academic.oup.com/annonc/article/26/3/431/219729
5. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology, 26(10), 1135–1145. https://www.nature.com/articles/nbt1486
6. Hoadley, K. A., Yau, C., Wolf, D. M., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, 173(2), 291–304.e6. https://www.cell.com/cell/fulltext/S0092-8674(18)30164-7
7. Szklarczyk, D., Gable, A. L., Lyon, D., et al. (2021). The STRING database in 2021: Customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Research, 49(D1), D605–D612. https://academic.oup.com/nar/article/49/D1/D605/6006196
8. Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics, 31(16), 2745–2747. https://academic.oup.com/bioinformatics/article/31/16/2745/196361
9. Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. https://genome.cshlp.org/content/13/3/495
10. Adzhubei, I. A., Schmidt, S., Peshkin, L., et al. (2010). A method and server for predicting damaging missense mutations. Nature Methods, 7(4), 248–249. https://www.nature.com/articles/nmeth0410-248
11. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences, 74(12), 5463–5467. https://www.pnas.org/doi/10.1073/pnas.74.12.5463
12. Curtis, C., Shah, S. P., Chin, S. F., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403), 346–352. https://www.nature.com/articles/nature10983
13. Győrffy, B., Lánczky, A., & Szállasi, Z. (2010). Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in breast cancer. Breast Cancer Research and Treatment, 123(3), 725–731. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2913582/
14. Soria, J. C., Marabelle, A., & Brahmer, J. (2017). The evolving landscape of immunotherapy in lung cancer. Nature Reviews Clinical Oncology, 14(9), 555–567. https://www.nature.com/articles/nrc.2017.24
15. Yates, L. R., & Campbell, P. J. (2012). Evolution of the cancer genome. Nature Reviews Genetics, 13(11), 795–806. https://www.nature.com/articles/nrc3298
16. Weinstein, J. N., Collisson, E. A., Mills, G. B., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, 45(10), 1113–1120. https://www.nature.com/articles/ng.2764
17. Futreal, P. A., Coin, L., Marshall, M., et al. (2004). A census of human cancer genes. Nature Reviews Cancer, 4(3), 177–183. https://www.nature.com/articles/nrc1299

18. Sawyers, C. L. (2004). Targeted cancer therapy. Nature, 432(7015), 294–297. https://www.nature.com/articles/nature03095

19. Bailey, P., Chang, D. K., Nones, K., et al. (2016). Genomic analyses identify molecular subtypes of pancreatic cancer. Nature, 531(7592), 47–52. https://www.nature.com/articles/nature16965

20. Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, 144(5), 646–674. https://www.cell.com/cell/fulltext/S0092-8674(11)00127-9

21. Goossens, N., Nakagawa, S., Sun, X., & Hoshida, Y. (2015). Cancer biomarker discovery and validation. Translational Cancer Research, 4(3), 256–269. https://pmc.ncbi.nlm.nih.gov/articles/PMC4511498/

22. Andre, F., et al. (2007). Promises and caveats of in silico biomarker discovery. British Journal of Cancer, 96(5), 679–683. https://www.nature.com/articles/6604495

23. Goossens, N., Nakagawa, S., Sun, X., & Hoshida, Y. (2015). Cancer biomarker discovery and validation. Translational Cancer Research, 4(3), 256–269. https://pmc.ncbi.nlm.nih.gov/articles/PMC4511498/

24. Emerging methods and techniques for cancer biomarker discovery. (2023). PubMed. https://pubmed.ncbi.nlm.nih.gov/39232287/

25. Biomarkers From Discovery to Clinical Application: In Silico Pre-Clinical Validation Approach in the Face of Lung Cancer. (2023). PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC11452870/

26. omczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology, 19(1A), A68–A77. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4522293/

27. Perou, C. M., Sørlie, T., Eisen, M. B., et al. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747–752. https://www.nature.com/articles/35021093