Based promoter prediction for oncogene in cancer

Donia sameh Farouk

Department of Computer and Software Engineering

Ainshams university

Cairo,Egypt

20P3424@eng.asu.edu.eg

Seif Eldin Amr Mostafa

Department of Computer and Software Engineering

Ainshams university

Cairo,Egypt

20P2006@eng.asu.edu.eg

Submitted to

Ashraf AbdelRaouf

Assistan Professor in Computer Science Misr

International University

Cairo, Egypt

ashraf.raouf@miuegypt.edu.eg

***Abstract*— In biomedical research, digital twins—virtual models of biological systems—are transforming personalized medicine by simulating disease behavior and treatment response. Leveraging generative artificial intelligence (AI) and spatial biomedical data, these models enable researchers to test "what if" clinical scenarios, especially in fields such as oncology and computational pathology. In this study, we focus on oncogenes, which are cancer-causing genes often activated by nearby promoter regions. Using a convolutional neural network (CNN), we trained a model to distinguish promoter DNA from non-promoter regions with 86% accuracy. We then scanned the 1000 base pairs upstream of 235 cancer-related genes from the COSMIC Cancer Gene Census in the GRCh38 human genome, applying a 301 bp sliding window. Our model identified promoter-like sequences for each gene, with multiple sites in some cases. To explore therapeutic potential, we simulated CRISPR-Cas9 silencing on these predicted promoters, generating a list of candidate guide RNAs for future validation. This approach demonstrates how combining deep learning with digital twin simulations can accelerate the discovery of precise, patient-specific gene editing targets for early cancer intervention.**

Keywords— Generative AI, Digital twin, Oncogenes, promoter, convolutional neural network, COSMIC Cancer Gene Census, GRCh38 genome, CRISPR-Cas9

I. Introduction

Introduction—Cancer is a leading global health challenge, accounting for roughly 1 in 6 deaths worldwide; in 2022, there were 20 million new cases and 9.7 million deaths [1].

. The number of cases is projected to rise by 77 % by 2050, placing enormous strain on healthcare systems across all regions.

At the same time, significant disparities persist, with low- and middle-income countries facing higher mortality rates due to limited access to early diagnosis and advanced therapies. These trends underscore the urgent need for innovative, patient-centered strategies that can identify and target cancer drivers before tumors progress. Digital twin technology—computerized replicas that integrate an individual's genetic profile, environmental exposures, and molecular data—offers a powerful platform for personalized medicine. By simulating how specific genomic alterations drive tumor growth and screening targeted interventions in silico, digital twins can accelerate discovery and tailor treatments to each patient's unique biology. In this work, we develop a genomic digital twin pipeline that combines deep learning–based promoter prediction and CRISPR-Cas9 silencing simulations to enable early, individualized cancer detection and intervention [2].

II. What Is a Biological Digital Twin?

A digital twin is one such substitute solution. The goal of a digital twin is to replicate a biological twin as closely as possible using computer models or simulations. However, because of the present gaps in our knowledge of biological entities, particularly humans. A one-to-one replica of the biology level concept can be used to immediately transfer the ideas of personalized biology and traditional biology to the medical level. Traditional medicine and customized medicine follow from this. However, for several reasons (such as practical and ethical ones), it is not possible to bring the concept of a biological twin to the level of medicine. One requires a substitute solution as a result like shown in fig1.

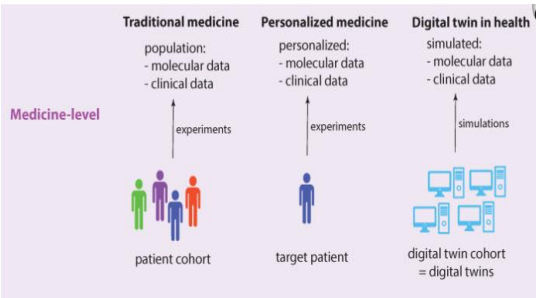

Fig. 1. Digital twin in health

A. Personalized Biology Through Digital Cloning

Assume that we are interested in a single member of this sample, which could be either a single unicellular creature or a single multicellular organism. This would be one animal in the former scenario and one cell in the latter. This organism is referred to as the target organism. Then, by concentrating on the target organism, we might conduct personalized biology experiments that are comparable to personalized medicine research. To obtain additional experimental data, we would want to conduct research on a sample of identical organisms since we are concerned about this organism. Cloning may theoretically create such similar organisms. Because each biological twin is identical to the target organism, this will enable us to produce a cohort of biological twins of our target organism.

B. Data Generation vs. Data Analysis

The fact that a digital twin produces (biologically realistic) data is a crucial component of this notion. This suggests that a digital twin cannot be used for data analysis or, more generally, for question answering. Rather, a digital twin serves as a substitute for a target patient, producing data that appears to be derived from the target patient. It is crucial to stress this since there is a substantial difference between data and, for example, a machine learning technique for data analysis. A digital twin, then, is a computer simulation that creates data that resembles virtual biological or biomedical trials.

III. Applications and Clinical Integration of Digital Twins in Biomedicine

We believe that there is a conceptual issue with a digital twin that has not yet been adequately addressed, despite all of the advancements made in other sectors. In particular, the term "digital twin" is utilized in a confusing, imprecise, and jumbled manner throughout the literature, concealing rather than elucidating design concepts. We tackle this problem in this study by offering a data-centric machine learning viewpoint on a digital twin. The definitions given should be applicable to the inanimate nature of a mechanical digital twin for engineering, even if the following discussion focuses on a biological digital twin (for medical and health issues).

A. Digital Twins in Biomarker Discovery and Precision Medicine

Using extensive biological datasets, DTs can be developed for biomarker and drug identification to model pharmacological effects and forecast treatment effectiveness. Future randomized clinical trials (RCTs) may benefit from the use of DTs and virtual patient cohorts, which could save costs and improve efficacy. DTs can facilitate the smooth integration of genetic and molecular profiles with imaging data in the field of medical diagnostics, creating virtual disease states that are highly useful for medical research. Clinicians may be able to improve the precision and efficacy of tailored diagnostics by using DT simulation to inform their surgical planning and targeted therapy selection decisions [3].

B. Clinically Validated Use Case: The Artificial Pancreas

The artificial pancreas model used to treat patients with type I diabetes is arguably the best application example that has undergone clinical testing too far. A mathematical model that mimics a target patient's glucose metabolism is called an artificial pancreas. In particular, the simulation model predicts the necessary insulin level by using real-time blood glucose levels obtained from the patient via a sensor to carry out a closed-loop control. A patient-attached pump administers the proper dosage of insulin in the event of a deviation. Although this specific outcome is undoubtedly quite good, it only looks at one level of the human body [4].

IV. Biological Background

Currently, the term "oncogene" has multiple definitions. Four definitions that describe varying quantities of genes are listed on the National Institutes of Health's Genetics Home Reference Guide website.

"A gene mutation that leads to the development of cancer is called an oncogene." Oncogenes, also known as proto-oncogenes, regulate cell division while they are in their normal, unmutated condition [5].

According to table 1, proto-oncogene is a normal gene that plays a role in apoptosis, proliferation, or cell division. A mutant form of a proto-oncogene that causes or contributes to the development of cancer is called an oncogene. Gain-of-function mutations are known as oncogenes because they either directly or indirectly prevent apoptosis or encourage cell proliferation. Point mutations, gene amplification, or chromosomal rearrangements can all produce oncogenes, which can then activate and promote cell division. In humans, around 50 distinct oncogenes have been identified.

Table1.Major Classes of Cancer Genes

Class	Definition
Oncogene	A mutant form of a proto-oncogene that initiates or participates in the development of cancer by stimulating cell division or inhibiting apoptosis
Tumor suppressor gene	A gene whose product inhibits cell division and proliferation and activates apoptosis

A. Oncogenes and Proto-oncogenes

Proto-oncogenes are essential genes that normally guide a cell’s growth, division, and survival—much like a car’s gas pedal tells the engine when to accelerate. Under typical conditions, they produce proteins such as growth factors, receptor kinases, or intracellular signaling molecules that keep cell renewal on track. However, when a proto-oncogene acquires certain alterations, it can become permanently “stuck down,” transforming into an oncogene that forces the cell to proliferate uncontrollably.

Although they are present in all healthy cells, oncogenes were initially identified in viruses that cause cancer. Strictly speaking, the proto-oncogene is the original, unaltered wild-type allele of an oncogene. The proto-oncogene of the wild type stimulates cell division and proliferation as in fig 2. Cell division is tightly regulated during a multicellular organism's development. An organ or tissue should stop growing and its cells should stop dividing

once it reaches the proper size. It is obvious that changes to the genes that regulate cell division have the potential to be extremely harmful. The oncogenes that cause cancer are these mutated forms [6].

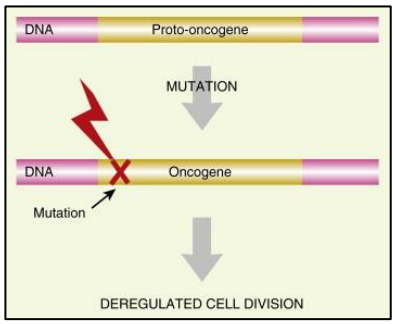

Fig.2. Oncogenes Are Mutant Alleles of Proto-Oncogenes

Key routes to oncogene activation include:

- Inherited or Somatic Mutations: A single–base change in the DNA sequence (for example, in the RAS gene) can lock the protein in an active state. Such variants may be passed down in families or arise spontaneously during cell division.

- Epigenetic Deregulation: Chemical flags—like DNA methylation or histone modifications—can switch a proto-oncogene on without touching its underlying code, boosting its expression inappropriately.

- Chromosomal Rearrangements: Pieces of chromosomes can reshuffle during mitosis, placing a strong regulatory element from one gene next to a proto-oncogene. This new "on" switch drives continuous gene expression (as seen in the BCR–ABL fusion).

- Gene Duplication (Amplification): Extra copies of a proto-oncogene flood the cell with its protein product, analogous to flooring the accelerator pedal.

B. *Tumor Suppressor Genes*

Tumor suppressors serve as the cell's brakes, enforcing checkpoints that pause division, repair DNA mistakes, or initiate programmed cell death (apoptosis) when damage is too severe. When these genes malfunction, the brakes fail:

- Germline Mutations: In conditions like Li-Fraumeni syndrome, individuals inherit a defective TP53 allele in every cell. Loss of the second TP53 copy during life unleashes unchecked growth, dramatically raising cancer risk.
- Acquired (Somatic) Mutations: More commonly, tumor suppressors pick up damaging mutations over time that disable one or both gene copies, erasing critical safety checks.

- Epigenetic Silencing: Just as for oncogenes, chemical tags can lock a tumor-suppressor promoter in an "off" position, silencing its expression without altering the DNA letters themselves [6].

C. Promoters and Their Role in Gene Regulation

What Is a Promoter? A promoter is a DNA segment immediately upstream of the transcription start site. It contains motifs like the TATA-box, CpG islands, and binding sites for transcription factors.

Why Upstream Regions Matter? These sequences dictate when, where, and how much a gene is transcribed. Small changes—mutations, deletions, or epigenetic marks (e.g., DNA methylation)—can dramatically boost or silence gene output. By targeting Promoter by accurately mapping promoters of overactive oncogenes, we gain a precise "knob" to turn down harmful gene expression at the DNA level, rather than blocking proteins further downstream [7].

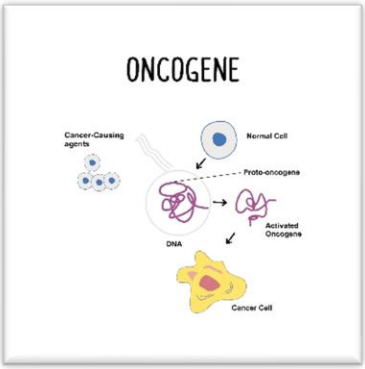

Fig.3. Cancer and Oncogenes

D. Key Drivers of Cancer Development

1. Oncogene Overactivation

When proto-oncogenes pick up mutations, extra copies, or strong "on" switches in their promoters, they turn into oncogenes that flood the cell with growth signals as in fig 3. This constant "go" drive—whether from a mutated RAS locked in the active state, amplified MYC copies, or fusion proteins like BCR–ABL—pushes cells past every checkpoint toward relentless division.

2. Tumor Suppressor Silencing

Tumor suppressors (the cell's brakes) such as TP53, RB1, and BRCA1/2 normally pause the cycle, repair DNA damage, or trigger programmed death. When these genes are inactivated—by inherited or acquired mutations, or by epigenetic tags that shut off their promoters—the brakes fail. Damaged or dangerous cells escape repair and keep multiplying.

3. Genomic Instability

With growth accelerators jammed and brakes gone, cells replicate rapidly, but their DNA repair machinery can't keep up. Errors accumulate—point mutations, chromosomal breaks, copy-number changes—creating a turbulent genome that spawns new mutations, fuels tumor heterogeneity, and accelerates progression [8].

V. Solution

The advancement of gene editing techniques has made it possible to directly target and modify gene sequences in

nearly all dangerous oncogenes. This technology has immense promises for usage in a wide range of sectors, from applied biomedicine and biotechnology to basic research. The shift from basic research to advancements in gene edition in clinical practice has been greatly hastened by the most recent developments in programmed nucleases, such as ZFNs, TALENs, and CRISPR/Cas. We turned our ideas into practice with a clear, step-by-step workflow that finds identify promoter sequences with a neural network, check the existence of these promoters upstream oncogene, flags which oncogenes are overexpressed, and runs CRISPR silencing simulations on those high-risk targets.

A. Data Preparation

1) Oncogene List:

Loaded COSMIC_CancerGeneCensus_v101_GRCh38.tsv and filtered 235 genes labeled "oncogene" with its needed data, sample of the extracted oncogenes:

	gene	chr	start	end
0	A1CF	10	50799409	50885675
1	ABL1	9	130713946	130887675
2	ABL2	1	179099327	179229684
3	ACKR3	2	236567787	236582358
4	ACVR1	2	157736444	157875862

Fig.4. Promoter Prediction per Chromosome

2) Upstream Sequence Extraction:

Pulled the 1 000 bp directly upstream of each gene's transcription start site from the GRCh38 primary assembly FASTA (Homo_sapiens.GRCh38.dna.primary_assembly.fa), using Ensembl/GENCODE coordinates.

3) Simulated Expression Profiles

- Built a toy dataset of expression values to mimicnormal vs. overexpressed oncogenes (e.g., sampling from log-normal distributions).
- Tagged each oncogene as "overexpressed" if its simulated expression exceeded the 90th percentile of the baseline distribution.

B. PROMOTER PREDICTION MODEL:

1) Hybrid CNN+GRU Architecture

Model summary is shown as follows

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 298, 2)	458
max_pooling1d (MaxPooling1D)	(None, 99, 2)	0
dropout (Dropout)	(None, 99, 2)	0
conv1d_1 (Conv1D)	(None, 99, 14)	770
bidirectional (Bidirectional)	(None, 256)	110,592
dropout_1 (Dropout)	(None, 256)	0
dense (Dense)	(None, 128)	32,896
dense_1 (Dense)	(None, 64)	8,256
dense_2 (Dense)	(None, 64)	4,160
dense_3 (Dense)	(None, 16)	1,040
dense_4 (Dense)	(None, 1)	17

Fig.5. Model Architecture Summary

2) Windowing & Encoding

- One-hot encoded each 1 000 bp region and slid a 301window across it

3) Training & Performance

- Trained on balanced promoter vs. non-promoter data.
- Achieved 86 % test accuracy
- Input: One-hot encoded DNA
- Output: Promoter likelihood

C. Genome-wide Promoter Scanning

Geneticists have been describing the location of genes (and genetic elements) using a variety of methods ever since genetic tests were developed. The cytogenetic location, which characterizes genes based on bands in dyed chromosomes, was one of the first methods to be employed. Nonetheless, scientists have been able to pinpoint the precise positions of genes and other components thanks to the development of high-resolution genomic tools. The valuable beginning and ending positions of a gene (or genetic element) on a chromosome are described by this molecular location, sometimes referred to as genomic coordinates [9].

We concentrated on finding actual human oncogenes and examining their upstream DNA sequences to find putative promoter areas to gain a better understanding of cancer gene regulation. Oncogenes are genes that can cause cancer and unchecked cell division when they are mutated or overexpressed. Upstream of genes, promoters are essential regulatory DNA regions that govern transcription. To do this, we used the GRCh38 version of the human reference genome (.fa file) to check 1000 base pairs upstream of each known oncogene. Because it is likely to contain promoter elements that affect gene expression, this area was chosen. To guarantee precise mapping, we employed genomic coordinates obtained from the Genome Reference Consortium and genome annotation data from GENCODE. Using a sliding window technique with 30 base pair (bp) windows, our model was applied over these upstream sequences to detect promoter activity, allowing for the fine-grained detection of patterns resembling promoters. This approach aids in identifying the specific sites in cancer-related genes that might act as regulatory "switches" [10]-[11].

1) Sample Results:

- Ran the trained model on every 301 bp window for the 235 oncogenes.
- Found promoter signatures in all genes, with up to 7 hits in one gene.

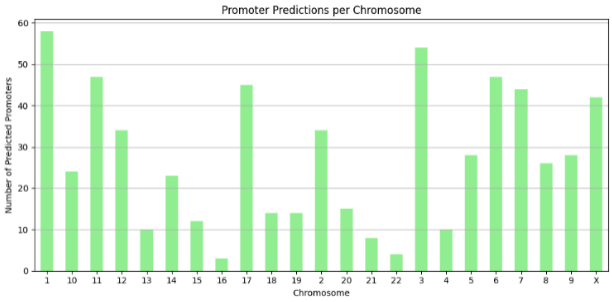

Fig.5. Promoter Prediction per Chromosome

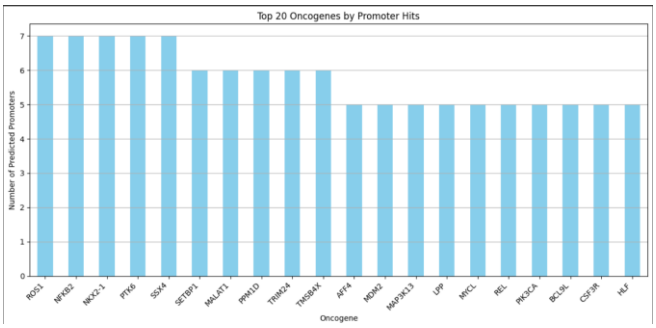

Fig.5. Top 20 Oncogenes by Promoters Hits

D. Simulate CRISPR silencing for upstream promoters

CRISPR/Cas is the most recent gene editing tool to be created. As an "immune system," it was first found in bacteria and archaea to defend these species against viral infections such as phages. The utilization of a guide RNA (gRNA) that attaches to the DNA target site and a nuclease called CRISPR-associated caspase protein (Cas) that cleaves DNA strands complementary to the CRISPR system's RNAg is an intriguing aspect of CRISPR systems. This gene editing technique can be used to target specific mutations, combat cancer, and increase adaptive immunity. It has been demonstrated that pancreatic, prostate, colon, and breast cancers can be treated with this gene editing approach to reduce tumor size, migratory capacity, and medication resistance.

Numerous CRISPR-based approaches have been put forth for the treatment of cancer. Inactivating the genes that promote tumor growth is one strategy. To stop tumor growth, for instance, it has been suggested to use CRISPR to inactivate the oncogene MYC as in fig4. Many forms of cancer are known to have overactive MYC genes, and deactivating these genes may delay or even reverse the disease's growth. Increasing the immune system's reaction to cancer cells is an additional strategy. For instance, scientists have knocked off or reduced the expression of the PD-1 protein on T cells using CRISPR-based gene editing. Furthermore, genetic mutations that cause cancer, such those resulting from BRCA1 and BRCA2 mutations, can be repaired using CRISPR-based gene editing. For instance, research has indicated that CRISPR-Cas9 has the potential to be used in cancer treatment by correcting BRCA1 abnormalities in human cells.

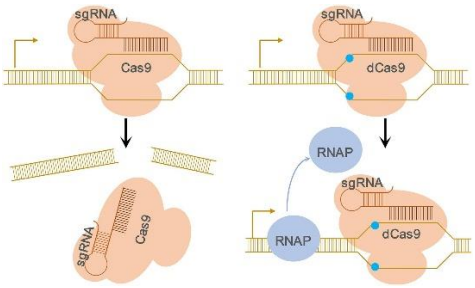

Fig.6. CRISPR Silencing

Finding oncogenes that are essential to the genesis of cancer is the first step in the mechanism of CRISPR-based techniques for oncogene inactivation. Oncogenes are frequently linked to mutations or aberrant gene amplifications that cause their corresponding proteins to be overexpressed, which results in unchecked cell division and growth. Following the identification of the target oncogene, scientists create a gRNA that selectively identifies and attaches to the amplified or mutant oncogene region. Inactivating the genes that promote tumor growth is one strategy. In animal models of lymphoma, for instance, it has been demonstrated that deactivating the MYC oncogene inhibits the formation of tumors. The foundation of this approach is the idea that oncogenes, which stimulate cell growth and proliferation, are overexpressed in cancer cells due to genetic alterations. By deactivating these oncogenes, cancer cell proliferation can be halted [12].

Our Silencing criteria works as follows:

1) Selecting High-Risk Targets
   - From the 235 genes with predicted promoters, we kept only those marked as “overexpressed” in our simulated profiles. This focuses our efforts on the oncogenes most likely driving tumor growth.
2) Building the Genomic Twin
   - For each overexpressed gene, we built a minimal “twin” of its 301 bp promoter windows. This virtual snippet is the substrate for our editing simulations.
3) Guide RNA (gRNA) Design
   - Scanned each window for “NGG” PAM motifs.
   - Designed gRNAs to flank core promoter elements, scoring them by on-target efficiency and off-target risk.
4) Silencing Mechanism & Simulation
   - Cas9 cuts DNA three bases upstream of the PAM.
   - NHEJ repair introduces small indels. When indels land in critical motifs (e.g. TATA box), they disrupt transcription factor binding.
   - Outputs are saved in excel sheet as follows:

	gene	chr	strand	gRNA_target_start	gRNA_target_end	gRNA_sequence	promoter_probability	simulated_effect
0	AR	X	+	67543261	67543281	CAAAGCAAACGTTTACAGAG	0.718869	Silencing AR by targeting predicted promoter
1	AR	X	+	67543565	67543585	CCTCAGTTTGTAGGAGACTC	0.505562	Silencing AR by targeting predicted promoter
2	CDK4	12	+	57746949	57746969	TCGGCATAGGTATTAGTCAC	0.668768	Silencing CDK4 by targeting predicted promoter
3	CDK4	12	+	57747046	57747066	GATGTGTGGAGAAAAGTTTC	0.690286	Silencing CDK4 by targeting predicted promoter
4	CDK4	12	+	57747399	57747419	CTTGTCACATCCTCTAAAAT	0.533227	Silencing CDK4 by targeting predicted promoter

Fig.7. CRISPR Silencing Overexpressed Oncogenes Output

VI. Conclusion

To identify potential promoter areas that regulate cancer gene activity, we used the GRCh38 human reference genome to map 1000 base pairs upstream of known oncogenes. We were able to perform a thorough examination of these areas by employing a sliding 301 bp window, which improved our comprehension of how specific genes are activated in cancer. Early cancer identification and individualized treatment are supported by this approach.
The concept of biological digital twins—virtual representations that closely resemble a patient's biology—was also investigated. These twins forecast a patient's potential response to therapies using patient data and cutting-edge technologies like artificial intelligence and machine learning. Digital twins have the potential to significantly enhance clinical judgments and personalize healthcare as they develop.

REFERENCES

[1] "Global cancer burden growing, amidst mounting need for services," *Who.int*. Available: https://www.who.int/news/item/01-02-2024-global-cancer-burden-growing--amidst-mounting-need-for-services. [Accessed: 29-May-2025].

[2] "World cancer day 2024," *World Health Organization - Regional Office for the Eastern Mediterranean*. Available: https://www.emro.who.int/media/news/world-cancer-day-2024.html. [Accessed: 29-May-2025]

[3] J. Wu and V. H. Koelzer,"Towards generative digital twins in biomedical research," Computational and Structural Biotechnology Journal, https://www.sciencedirect.com/science/article/pii/S2001037024003192 (accessed May 28, 2025).

[4] F. Emmert-Streib, and O. Yli-Harja, (2022) What is a digital twin? experimental design for a data-centric machine learning perspective in health, International journal of molecular sciences. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC9653941/ (Accessed: 27 May 2025).

[5] Mahendran Botlagunta, "Nutraceuticals-loaded chitosan nanoparticles for chemoprevention and cancer fatigue," *Elsevier eBooks*, https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/oncogene [Accessed: 28-May-2025].

[6] "Oncogenes, tumor suppressor genes, and DNA repair genes," *Cancer.org*.Available:https://www.cancer.org/cancer/understanding-cancer/genes-and-cancer/oncogenes-tumor-suppressor-genes.html. [Accessed: 29-May-2025].

[7] N. Q. K. Le, E. K. Y. Yapp, N. Nagasundaram, and H.-Y. Yeh, "Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous FastText N-grams," *Front. Bioeng. Biotechnol.*, vol. 7, p. 305, 2019, https://www.frontiersin.org/journals/bioengineering-and-biotechnology/articles/10.3389/fbioe.2019.00305/full

[8] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," Cell, vol. 144, no. 5, pp. 646–674, 2011, https://www.cell.com/fulltext/S0092-8674%2811%2900127-9

[9] P. Samarakoon, "Genomic coordinates to gene lists and vice versa — Annotating gene coordinates and gene lists," *Into the genomics*, 25-Dec-2018. Available: https://medium.com/intothegenomics/annotate-genes-and-genomic-coordinates-ecdad47d0c8e. [Accessed: 27-May-2025].

[10] A. Frankish *et al.*, "GENCODE reference annotation for the human and mouse genomes," *Nucleic Acids Research*, vol. 47, no. D1, pp. D766–D773, https://academic.oup.com/nar/article/47/D1/D766/5144133?login=false

[11] "Human genome overview - genome reference consortium," *Nih.gov*. Available: https://www.ncbi.nlm.nih.gov/grc/human. [Accessed: 27-May-2025].

[12] M. Montaño-Samaniego, D. M. Bravo-Estupiñan, O. Méndez-Guerrero, E. Alarcón-Hernández, and M. Ibáñez-Hernández, "Strategies for targeting gene therapy in cancer cells with tumor-specific promoters," Front. Oncol., vol. 10, p. 605380, 2020 https://pmc.ncbi.nlm.nih.gov/articles/PMC7768042/ [Accessed: 27-May-2025].