SILENCED BY METHYLATION

2025-2026 CYSF

LOGBOOK

Helli Chalishajar

FAIRVIEW SCHOOL

Problem

Dec 23

Breast cancer outcomes depend heavily on how early the disease is detected, yet current screening & diagnostic pathways can miss early cases or require unnecessary procedures.
↳ Results in a major problem for our healthcare system.

QUESTION:
↳ Can specific DNA methylation patterns be used to detect early-stage breast cancer, & how practical & efficient could this approach be in today's healthcare system?

HYPOTHESIS:

- if specific DNA methylation patterns in key breast cancer-associated genes are identified & analyzed using a ML model, they can be used as biomarkers.
- By training a ML model, a classifier can learn distinct info or a "methylation fingerprint" in hopes of early detection for breast cancer.
- The results can demonstrate how integrating methylation-based screening can revolutionize the healthcare system.
  ↳ Improved diagnostic accuracy, patient care, & much more.

Breast Cancer

Dec 26

- Disease where abnormal cells in the breast grow uncontrollably, forming tumours that can invade nearby tissues and spread (metastize) to other parts of the body
  - Most commonly starts in the milk ducts or lobules.

- Affects both genders

- Key types are named after their origin, or characteristics
  - eg) ductal, Her-2 positive, etc.

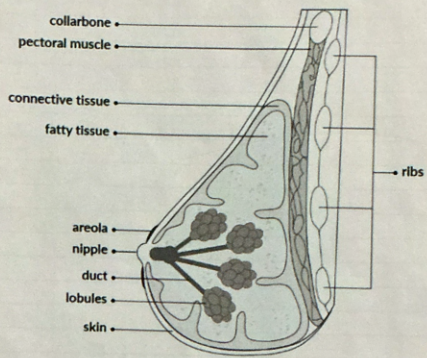
- Exact cause of breast cancer is unknown.
- However, researchers have found factors that increase your risk of getting it, such as:
  - hormones
  - lifestyles
  - environment

  } Breast cancer = combination of your genetic makeup & the world around you.

Specific factors include:
- A family history of breast cancer
- A personal history of breast cancer
- A personal history of breast conditions
- Beginning your period at a younger age
- Beginning menopause after 55
- Being female
- Dense breast tissue
- Alchohol
- obesity
- Radiation exposure
- Increasing age
- Never being pregnant

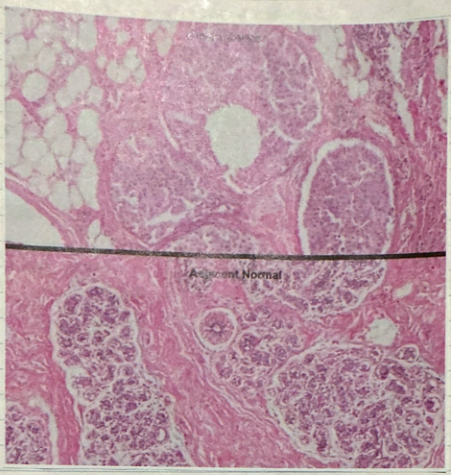
DIAGRAM →

The Breast



© Canadian Cancer Society

Differences in normal & cancerous breast cells



OVERALL APPEARENCE :

· Cancer :
- loss of normal structure : The ducts / lobules are irregular
- Crowded tissue as the cells are tightly packed together
- Tumour cells are infiltrating into the surrounding stroma.

Normal :
- Organized lobules & ducts
- Clear boundaries ; seperated from the stroma
- Orderly layout

CELL DENSITY & NUCLEUS

Cancer
- Many more cells per area
- Large, dark nuclei, indicating high DNA content

- high nuclear - to - cytoplasmic ratio
- Nuclei vary in size and shape

Normal
- Lower cell density
- small, lighter nuclei
- More cytoplasm relative to nucleus.

GLAND & LUMEN APPEARENCE

Cancer
- Misshapen glands
- Irregular lumens
- Cells often pile up, losing the normal single-layer lining

Normal
- Round / oval lumens
- Clear seperation between epithelial cells & surrounding tissue.

Oct 22

Roles / differences in healthy vs cancerous cells

HEALTHY

- Healthy cells grow & divide only when the body needs new cells. They stop reproducing when enough cells are present
  - ↳ process called contact inhibition
- Specialized functions
  - → Normal breast cells mature into distinct cell types with specific functions such as producing milk during lactation.
- Apoptosis:
  - ↳ Healthy cells have a built-in process to self destruct if they are damaged beyond repair or are useless now
    - vital in maintaing a healthy body.
- They stick together & stay in their designated location in the body.

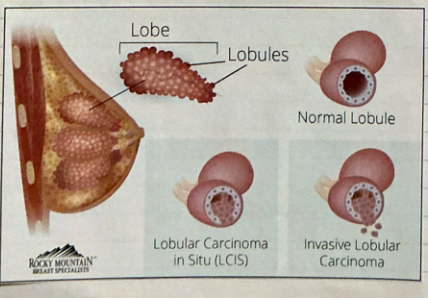
CANCEROUS:

- Cancer cells ignore the body's signals to stop growing and dividing, instead multiplying continuously and forming tumours.
- Divide so rapidly that they do not mature into specialized, functional cells.
- Bypass apoptosis, giving them an extended lifespan, contributing to tumour growth.
- lose the surface molecules that keep them anchored, resulting in metastasis.
- influences nearby normal cells to develop new blood vessels to supply the tumour
  - ↳ Angiogenesis.

Categorizing breast cancer

TYPES:

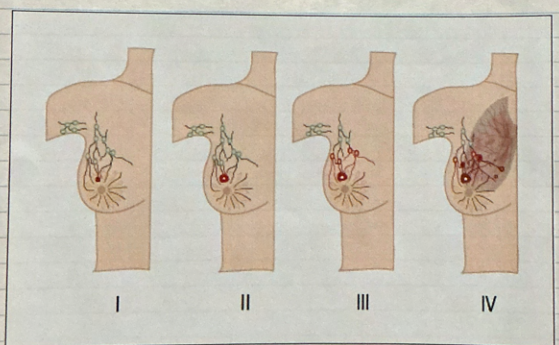
- Ductal carcinoma in situ (DCIS)
  - ↳ non-invasive cancer where abnormal cells are confined to the milk ducts & have not invaded surrounding tissue.
- Lobular carcinoma in situ (LCIS)
  - ↳ A marker of increased breast cancer risk due to abnormal cells in the lobules.
- Invasive ductal carcinoma (IDC)
  - ↳ Most common breast cancer & occurs when cancer cells spread from the ducts into surrounding tissue.
    - 70-80% of cases
    - Easier to detect as a lump.
- Inflammatory breast cancer
  - ↳ rare, aggressive type that blocks lymph vessels & causes redness, swelling, & skin thickening rather than a lump.
- Triple-neg breast cancer
  - ↳ lacks estrogen, progesterone, & HER2 receptors
- HER2-pos. breast cancer
  - ↳ overexpresses the HER2 growth receptor, leading to rapid growth.



Nottingham (Modify - Richardson) grading:
- scored 1 to 3 categories:
  - ↳ tubule formation
  - ↳ nuclear pleomorphism
  - ↳ mitotic rate
- total score → Grade
  - ↳ Grade 1: slow
  - ↳ Grade 2: intermediate
  - ↳ Grade 3: fast

STAGES

- ↳ Stage 0 = non-invasive
- ↳ Stage 1 = Early invasive
- ↳ Stage 2 = locally growing
- ↳ Stage 3 = Locally advanced
- ↳ Stage 4 = Metastatic



Why early detection is vital

Dec 29

- matters because it finds cancer before it spreads
- when a tumour is small & localized, treatments like surgery, radiation, or targeted therapy have a much higher chance of working.
  ↳ As cancer grows, it becomes more difficult

- Clearly shown in 5-year survival statistics:
  Breast Cancer:
  - Stage 1: 100% survive after 5 years
  - Stage 2: 93%
  - Stage 3: 74%
  - Stage 4: 23%

  Colorectal Cancer:
  - Stage 1: 92-95%
  - Stage 2: 85%
  - Stage 3: 11-14%

INTERPRETATION:
- People diagnosed at later stages have much lower chances of living 5 years after diagnosis compared to those diagnosed early.
- Once cancer metastizes, curative surgery becomes unlikely and treatment shifts toward controlling symptoms & slowing progression.

↳ Early detection avoids these scenarios.

Current breast cancer detection methods & limitations

MAMMOGRAPHY:

- Utilizes low dose X-rays to image breast tissue enabling radiologists to detect abnormalities like masses & microcalcifications that can indicate cancer.
- standard screening approach
- shown to reduce breast cancer mortality when implemented in population screening programs.

Accuracy & challenges:

- sensitivity (ability to correctly detect cancer) varies greatly
  - Around ≈ 79% on average, but can be much lower in younger or dense-breasted women.
- specificity (ability to correctly identify non-cancer) also varies, meaning mammograms sometime wrongly suggest cancer when none is present.

False negs.

- Miss about 1 in every 8 breast cancers, especially in women with dense breast cancer.

False pos.

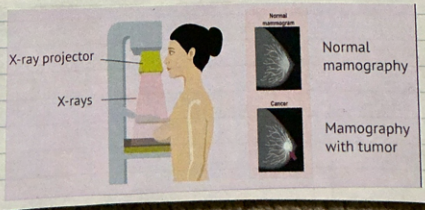
- Up to half of women screened over 10 years recieve a false-pos. mammogram.

Overdiagnosis

- can detect indolent lesions that might never harm the patient but lead to overtreatment.
- Estimate is over 20-50% in screening programs.

Radiation Exposure:

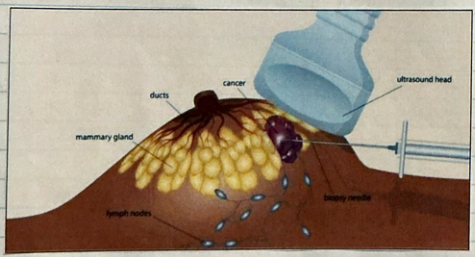
- ionizing radiation
- discomfort from breast compression



BIOPSIES:

- involve removing tissue from a suspicious area to examine it under a microscope.
- Core needle biopsy (CNB) is the most common for breast lesions seen on imaging.

- Biopsies are invasive procedures, often performed with a needle under imaging guidance.
  - ↳ can be painful, anxiety-inducing, & require local anesthesia.
  - ↳ carries risks like bleeding, infections, or scarring
- expensive relative to imaging
- Because imaging can't always distinguish benign from malignant lesions, many biopsies are performed unnecessarily
  - ↳ >70% of biopsies triggered by suspicious imaging turned out to be benign.



Limitations in young women & dense breast tissue

Breast tissue that has more glandular & fibrous components (denser) appears white on X-ray mammograms, the same colour as many cancers.
- Mammography's ability to detect cancer drops significantly in women with dense tissue.
  ↳ sensitivity as low as 30-46%
- Young women are often the ones with denser breast tissue.
- Dense tissue = "masking effect"
  ↳ leads to false negs. & later diagnoses.

FAILURES OF CURRENT METHODS
- Mammographies fail to detect some early tumours, especially in dense breasts
- High false pos. = unecessary biopsies & follow-ups
  ↳ increasing cost, anxiety, and risk
- Imaging relies on physical structure, not biology.
- Some cancers are notoriously hard to see on mammograms.
  ↳ Can be missed until more later
- MRI & supplemental imaging improves detection but is not feasible for routine screening in all populations.

Dec 30

Need for molecular & epigenetic detection

Before visible tumours form, cancer cells undergo biochemical changes, especially in DNA methylation, where methyl groups are added/removed from specific DNA regions.

↳ Because these alterations occur at the molecular level, they can be seen in circulating DNA long before tumours are large enough to be imaged.

Cancer cells shed circulating tumour DNA (ctDNA) into the bloodstream. This fraction of cell-free DNA carries both genetic & epigenetic signatures from the tumour.

In blood, ctDNA methylation patterns have been shown to be highly specific markers of cancer presence, sometimes even when imaging shows no mass.

EVIDENCE

- A multimodal cfDNA analysis approach that combines methylation with fragmentomic features achieved:
  - ↳ Area under the curve (AUC) ≈ 0.9 for discriminating early-stage breast cancer from non-cancer controls.
  - ↳ Specificity: ~93.6%.
  - ↳ Sensitivity: ~62-66% for stage I-II cancers
- Indicates strong ability to detect cancer based on molecular signals where imaging may struggle

When cfDNA methylation markers were combined with imaging results diagnostic accuracy for early cancer improved greatly, boosting combined AUC close to 0.94.

↳ Molecular cfDNA methylation doesn't depend on tissue density as it detects circulating markers.

Basics of Genetics vs Epigenetics

DNA:

- DNA is the molecule that carries the hereditary information of nearly all living organisms. It consists of a long sequence of four nucleotide bases:
  - Adenine
  - Thymine
  - Cytosine
  - Guanine
  - ↳ The order of these bases forms the genetic code that dictates biological traits and functions.
- Genes are specific stretches of this sequence that contains instructions for building proteins, which are the molecules that perform most cellular functions in the body.
  - ↳ When a gene is "read" by the cell, its DNA sequence is transcribed into RNA & then translated into a protein.
    - Proteins do most work in the cell (signalling, growth, repair)

Process:
- DNA with coding sequence is transcribed to mRNA
- mRNA is read in triplets (codons) that match specific amino acids
- A chain of amino acids form a protein

  - ↳ Only 1-2% of the human genome is coding DNA

KEY FUNCTIONS OF DNA & GENES:

- Information storage: DNA stores instructions for all cellular processes.
- Protein coding: Genes specify the amino acid sequence of proteins.
- Regulation: Regulatory sequences control gene expression patterns

→

Genetic mutations are permanent changes in the DNA sequence.

↳ For example, substitutions, deletions, or insertions of base pairs. These changes can permanently alter the structure of a gene & its protein product.

- The actual letters (A, T, C, G) are altered.
- Mutations can occur due to:
  - Errors during DNA replication
  - Exposure to radiation
  - Inherited changes from parents
- Why it's permanent?
  - The cell copies that change everytime it divides
  - It cannot simply be "undone" by the cell.

EPIGENETIC CHANGES

In contrast, epigenetic changes do not alter the DNA sequence itself.

↳ They affect how the gene is expressed; whether it's "on" (active) or "off" (silenced).

Operates "above the code" of DNA, they modify DNA or chromatin structure through chemical tags rather than altering the order of bases.

↳ These tags include methyl groups added to cytosine bases or chemical groups added to histone proteins.

↳ Heavily methylated gene = harder to read = silenced

↳ Unmethylated gene = easier to read = active

Cells have enzymes that add/remove epigenetic marks:

- DNA methyltransferases (+ methyl groups)
- Demethylases (- them)

Allows cells to
- respond to environmental signals
- Adapt to stress, diet, disease
- Change gene expression during development.

They serve as indicators of biological conditions, including disease states like cancer. These markers are useful for:

- Early detection
- Prognosis
- Monitoring treatment response
- Tracking disease progression

Types of epigenetic biomarkers:

- DNA methylation patterns
- Histone modifications
- Non-coding RNAs influencing gene regulation.

Jan 1

DNA Methylation

An epigenetic mechanism that modifies DNA without changing the underlying genetic code.
- ↳ A methyl group ($-CH_3$) is added to a cytosine base in the DNA molecule.
- ↳ This usually happens at CpG sites
  - locations where a C is right next to a G linked by a phosphate bond

When the methyl group attaches to the fifth carbon position on cytosine, this creates a chemical base called 5-methylcytosine (5mC)
- ↳ This change is purely chemical & does not alter the DNA letters = epigenetic

DNA methylation patterns help determine whether a gene is active or inactive, in several important ways:

1) Blocking transcription factors
   - Transcription factors are proteins that bind to specific DNA sequences to start gene expression. When a region is methylated, the methyl groups can physically block these proteins from binding, so the gene isn't turned on.
2) Recruiting Repression Proteins
   - Methyl groups attract special proteins that recognize methylated DNA. These proteins then recruit other enzymes that tighten up the local DNA structure (chromatin), making it harder for cellular machinery to access the gene.

CpG Islands

- short stretches of DNA rich in CpG sites, & they are often found near gene promoters (the control regions where gene expression begins).
  - ↳ In normal cells, these CpG islands are usually unmethylated, allowing important genes to be expressed.
  - ↳ In many cancers, these islands became abnormally methylated, especially around tumor suppressor genes, turning them off when they shouldn't be.

BIOMARKERS

Measurable indicators of a biological state or condition.

- DNA methylation is useful as a biomarker because:

① It is stable:

- Methyl groups remain in place & are preserved even after cells die, so methylation patterns can be detected in blood, saliva, or tissue samples.

② Reflects gene activity states

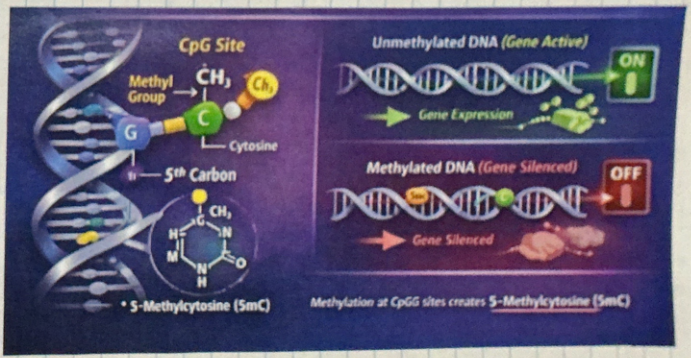
- Because methylation directly influences whether genes are turned on or off, abnormal methylation often correlates with disease processes such as cancer.

NORMAL METHYLATION PATTERNS:

- Normal DNA methylation patterns are carefully organized chemical marks that help cells function properly.
- In healthy cells, methylation is often found on repetitive DNA sequences and non-essential regions to keep them silenced, which protects the genome from instability & unnecessary gene activity.
- At the same time, important genes, especially tumor suppressor genes, have low methylation at their promoter regions so they can stay active & do their job.

ABNORMAL METHYLATION PATTERNS:

- Abnormal DNA methylation occurs when the normal balance of methyl groups on DNA is disrupted, causing genes to be regulated incorrectly.
- Instead of helping the cell control which genes should be active or silent, abnormal methylation can shut off important protective genes or allow harmful genes to become active.
- This loss of proper gene regulation interferes with normal cell growth & division and plays a major role in the development of cancer.


CpG Site
Methyl Group
CH_3
CH_3
C
G
Cytosine
5th Carbon
CH_3
* 5-Methylcytosine (5mC)
Unmethylated DNA (Gene Active)
ON
Gene Expression
Methylated DNA (Gene Silenced)
OFF
Gene Silenced
Methylation at CpGG sites creates 5-Methylcytosine (5mC)

Jan 3

DNA Methylation in Cancer

HOW ABNORMAL DNA METHYLATION CONTRIBUTES TO CANCER

- Normal DNA methylation helps cells regulate which genes are active & which are silent.
- In cancer, this carefully balanced system becomes disrupted, & 2 major types of abnormal methylation are consistently observed:
  - ↳ Hypermethylation
  - ↳ Hypomethylation

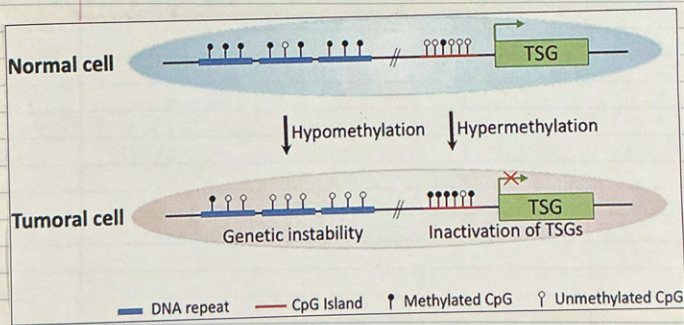
1. HYPERMETHYLATION OF TUMOUR SUPPRESSOR GENES

- Increase in methyl groups on DNA at specific sites, most notably within CpG islands near gene promoters.
  - ↳ In normal cells, many promoter regions are unmethylated so that necessary genes can be expressed.
  - ↳ In cancer cells, these regions acquire excess methylation, repressing gene activity.
- Frequently occurs at tumour suppressor genes
- When these protective genes are switched off, cells can proliferate abnormally & avoid normal checkpoints that prevent tumour growth.
- Cancer cells can show a pattern called CpG Island Methylator Phenotype, where many CpG islands across the genome become hypermethylated.
  - ↳ Hypermethylation has been found at genes like:
    - BRCA1 (breast cancer)
    - MLH1 (colorectal cancer)
    - p16/INK4a (many cancers)

    where it blocks transcription factor binding & recruits chromatin machinery, effectively silencing the gene.

2. HYPOMETHYLATION

- The loss of DNA methylation, especially across large stretches of the genome rather than at specific gene promoters.
  - ↳ Cancer cells often have global hypomethylation, meaning much of their genome loses methylation compared to normal cells.
- Can reactivate sequences that should be silent, such as repetitive elements or transposable elements.
  - ↳ These pieces of DNA can move around the genome & cause mutations or break chromosomes.
- Loss of methylation near certain oncogenes or regulatory genes can increase their activity, promoting uncontrolled growth.
- Typically occurs near repititive sequences & large blocks outside gene promoters.
- Tumours often show reduced methylation in LINE-1 elements, which correlates with higher cancer grade and progression.



- Cancer cells typically show both:
  - ↳ localized hypermethylation at specific promoter CpG islands = silencing tumor suppressors
  - ↳ Global hypomethylation across large genomic regions = promoting instability & activation of sequences that can drive malignacy.

- DNA methylation changes can occur before mutations accumulate before they alter gene regulation quickly and without changing the DNA code itself.
  - ↳ Methylation acts like an off/on switch, & switches are much easier/faster to flip than rewriting DNA instructions.

- Methylation alterations may reflect early disruptions in cell identity & growth control.

- Because hypomethylation broadly affects the genome, when too many genes turn on at once, the DNA becomes unstable, increasing the chance of errors that push cells toward cancer.

DNA Methylation Patterns as Biomarkers

Jan 4

DNA methylation patterns serve as great biomarkers for several reasons:

- DNA methylation biomarkers often reflect unique patterns associated with specific diseases, allowing them to distinguish between healthy tissue, benign conditions, & distinct cancer types.
- Methylation patterns differ not only between tumour & normal tissue but also between cancers of different organs, enabling classification of tumour origin & diagnoses.
  - ↳ eg) panels of differentially methylated regions have been used to differentiate breast, colon, liver, lung, pancreas, & stomach cancers with high accuracy
    - sensitivity: ~96.4%; specificity: ~97.7%
- DNA methylation is detectable in multiple biological samples, including tissue biopsies and cfDNA.
- A study using cfDNA methylation profiles across 6689 participants accurately detected multiple cancer types with 99.3% specificity & up to ~67% sensitivity in stages I-III.
- Liquid biopsy methylation tests can detect tumour-associated epigenetic changes before clinical symptoms arise, allowing earlier detection & easier repeated testing without surgery.

Comparison to Genetic Mutation Biomarkers

Genetic Mutation Biomarkers:
- changes in the DNA sequence itself.
  - ↳ eg) mutations in oncogenes or tumour suppresser genes
- Definitive signs of genomic alteration & can directly point to causal mechanisms of disease.
- some mutations occur late in disease progression or may not be unique to cancer types, & mutation tests may miss early stages when changes are subtle or present at low frequency.

DNA Methylation Biomarkers:
- Methylation reflects changes in gene regulation & cell state, not just structural changes in DNA.
  - ↳ provides insight into how a disease is affecting cells.
- Abnormal methylation often appears earlier in cancer, sometimes before structural mutations or tumour formation.
  - ↳ valuable for early detection
- Methylation marks are chemically stable & detectable in blood & bodily fluids, whereas RNA or protein biomarkers may degrade more easily.

- While genetic mutations show permanent DNA damage, DNA methylation biomarkers reveal early & reversible changes in gene activity, making them especially powerful for early cancer detection.

Breast cancer specific methylation patterns Jan 9

COMMONLY METHYLATED GENES IN BREAST CANCER

- Breast cancer cells show hundreds of CpG sites that are methylated differently compared to normal breast tissue.
- In 1 large study of TCGA breast tumour samples vs. matched normal tissue, 368 CpG sites were found to be significantly different, mostly hypermethylated in tumours & some hypo.
  - ↳ Hypermethylated:
    - RASSF1A - tumour suppressor
    - BRCA1 - DNA repair guardian
    - APC - cell growth regulator
    - GSTP1 - Detox & protective gene
    - ETS1, ABO, EBF1, NFYC,
      - ↳ part of a panel of genes whose CpG sites show significantly higher methylation in breast cancer & help distinguish tumour vs. adjacent tissue.

- 3 specific CpG sites, with official names cg12374721, cg18081940, & cg04475027, are found in or near the genes PRAC2, TDRD10, & TMEM132C. The way these sites are methylated is very different in cancer vs. normal tissue, & that difference:
  - ↳ Helps detect cancer
  - ↳ Predict outcomes like survival.

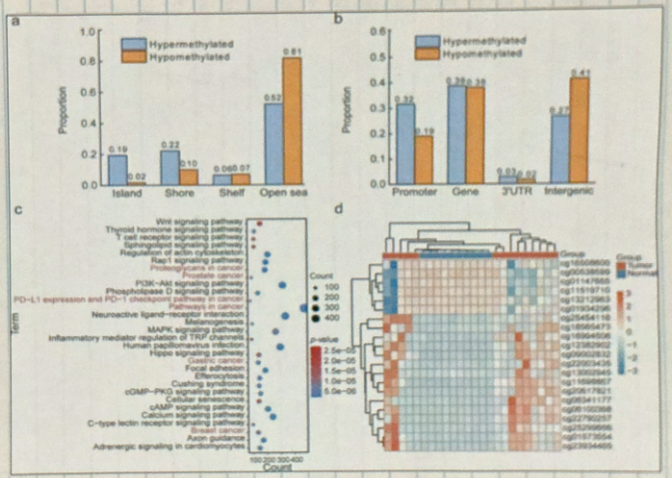
DIFFERENCES BETWEEN NORMAL & TUMOUR TISSUE METHYLATION:

- Cancer tissues show many more methylated changes compared with healthy breast tissue.
- The same TCGA study identified large methylation differences between tumour & normal tissue

↳ Promoter regions (the on/off switches just before a gene) are often hypermethylated in tumour tissue.
  - Especially true for promoters of genes involved in growth control, DNA repair, & cell adhesion.

↳ Hypomethylation in tumours tends to occur outside of promoter regions, including in gene bodies or intergenic regions.

- Normal breast tissue generally has lower average methylation at many CpG sites associated with regulatory functions, especially in regions that need to remain active for normal cell function
  ↳ vise versa for tumour tissue.


a
Hypermethylated
Hypomethylated
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
0.19
0.02
0.22
0.10
0.06
0.07
0.52
0.81
Island
Shore
Shelf
Open sea
b
Hypermethylated
Hypomethylated
Proportion
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.32
0.19
0.39
0.38
0.03
0.02
0.27
0.41
Promoter
Gene
3'UTR
Intergenic
c
Term
Wnt signaling pathway
Thyroid hormone signaling pathway
T cell receptor signaling pathway
Sphingolipid signaling pathway
Regulation of actin cytoskeleton
Rap1 signaling pathway
Proteoglycans in cancer
Prostate cancer
PI3K−Akt signaling pathway
Phospholipase D signaling pathway
PD−L1 expression and PD−1 checkpoint pathway in cancer
Pathways in cancer
Neuroactive ligand−receptor interaction
Melanogenesis
MAPK signaling pathway
Inflammatory mediator regulation of TRP channels
Human papillomavirus infection
Hippo signaling pathway
Gastric cancer
Focal adhesion
Efferocytosis
Cushing syndrome
cGMP−PKG signaling pathway
Cellular senescence
cAMP signaling pathway
Calcium signaling pathway
C−type lectin receptor signaling pathway
Breast cancer
Axon guidance
Adrenergic signaling in cardiomyocytes
100 200 300 400
Count
Count
100
200
300
400
p.value
2.5e−05
2.0e−05
1.5e−05
1.0e−05
5.0e−06
d
Group
cg16606606
cg00638599
cg01147855
cg11519710
cg13212983
cg01934296
cg25454116
cg18665473
cg16994506
cg15282902
cg09002832
cg22933435
cg13063945
cg11699857
cg03617621
cg08341177
cg06100398
cg22790257
cg15290656
cg01573554
cg23934465
Group
Tumor
Normal
p
1
0
−1
−2
−3

PANEL A: where methylation changes happen relative to CpG islands

- Blue = hypermethylated
- Orange = hypomethylated

Locations:
↳ Island (gene start)
↳ Shore (near island)
↳ Shelf (farther away)
↳ Open sea (far from genes)

- Tumours are more hypermethylated at CpG islands & shores
  ↳ Cancer is likely turning OFF protective genes
- Tumours are more hypomethylated in the open sea
  ↳ Quiet regions
  ↳ Losing methylation here cause genomic instability

PANEL B: where methylation changes occur relative to genes

Locations:
↳ Promoter = gene "on/off switch"
↳ Gene body = middle of gene
↳ 3' UTR = Tail end
↳ Intergenic = between genes

- Promoters are more hypermethylated = gene silencing
- Intergenic regions are more hypomethylated = loss of stability.

PANEL C: which biological pathways are affected

- This dot plot shows which gene pathways are enhanced for methylation changes.

- Bigger dots = more genes affected
- Darker colour = stronger statistical significance

Major pathways affected:
↳ Cancer signalling pathways
↳ Immune signalling
↳ Cell growth & survival
↳ Hormone signalling
↳ Inflammation-related pathways.

PANEL D: Heatmap distinguishing tumour vs. normal tissue.

Rows = specific CpG sites
Columns = samples
Red = higher methylation
Blue = lower methylation

↳ Tumour samples cluster together
↳ Normal samples cluster together
↳ Clear colour difference between the two groups.

Jan 10

Differences in Methylation by breast cancer subtype

1) Luminal Tumours (A & B)
   - ↳ Tend to have overall higher methylation frequencies.
   - ↳ Many studies find luminal tumours are most methylated overall, especially luminal B

2) Basal-like tumours
   - ↳ lower methylation levels than luminal tumours at CpG sites
   - ↳ Often resemble progenitor or stem-like cell methylation patterns, with less promoter CpG island methylation than luminal tumours.

3) HER2-Enriched Tumours
   - ↳ Shows distinct CpG methylation profiles, with many cancer-related genes hypermethylated.

Prediction Models / Machine Learning Jan 11

A prediction model is a tool that uses patterns learned from data to estimate outcomes for new, unseen samples.

In ML, a prediction model learns relationships between input features (eg. DNA methylation measurements) & output targets (eg. normal vs. cancer) by analyzing many sample cases.

↳ The goal is for the model to make accurate predictions on new data.

TRAINING DATA:

↳ used to teach the model what patterns correspond to which outcomes.

↳ Each sample in the training data includes both the features & correct labels.

↳ The model adjusts its parameters during training so that its predictions increasingly match the correct labels.

TESTING:

↳ seperate portion of the data that the model never saw during training.

↳ Used only after training to evaluate how well the model can generalize.

CLASSIFICATION:

↳ Type of supervised ML where the model predicts one of a set of discrete classes for each data point

- The classifier learns the relationship between features & the class label.

MY MODEL

Dataset selection

Jan 14

- Before beginning any coding, I explored multiple publicly available biomedical data repositories, including the GEO, NCBI datasets, World Health organization datasets, & other cancer genomics databases.

- I specifically searched for datasets that met several key criteria:
  - 1) Genome-wide DNA methylation data rather than gene expression data.
  - 2) A clear distinction between tumour & normal breast tissue.
  - 3) A sufficient number of biological samples for ML analysis
  - 4) Raw / minimally processed data that I could clean & analyze myself.

  - Because abnormal methylation patterns are known to play a role in cancer development, I specifically wanted a dataset measuring genome-wide CpG methylation.

After evaluating several datasets, I selected GEO dataset GSE66695 because it met all my criteria.
- contains methylation data from 120 breast tissue samples (80 tumour, 40 normal) measured using the Illumina HumanMethylation450 BeadChip.
  - This platform interrogates over 450 000 CpG sites across the genome, providing high-resolution epigenetic coverage.
- well suited for ML because it contains a large number of features & a clearly defined binary classification problem (Tumor vs. normal).

Platform Change

Jan 16

I initially attempted my analysis using Google Colab because it provides a free & accessible Python environment. However, the methylation dataset contained nearly half a million CpG sites, resulting in a very large matrix with tens of millions of data points.

While working in Colab, I encountered repeated issues as colab:

- ↳ Has limited session duration
- ↳ Has RAM limitations
- ↳ Can disconnect during large file parsing

Because parsing & cleaning large biological datasets requires stable memory & processing time, these interruptions made it difficult to reliably load & process the series matrix file.

I therfore transitioned to Kaggle, which provided a more stable execution environment & better support for large input files.

↳ This allowed me to complete the preprocessing without technical interruptions.

Parsing the series matrix file Jan 16

The raw data from GEO was provided as a series matrix file.
↳ A structured text file that contains both metadata (info on the study & samples) & the actual experimental measurements.

This file contained:
↳ study-level metadata
↳ Sample level metadata
↳ large table of methylation beta values

Parsing refers to the process of programmatically identifying & extracting the correct portion of a structured file.

Instead of manually coping the methylation table, I wrote code to automatically detect the markers !series_matrix_table_begin and !series_matrix_table_end.
↳ Ensured I extracted only the actual methylation matrix & avoided accidently including metadata in my dataset.

This approch is important because it makes the process reproducible & reduces the risk of human error.

After parsing, I obtained a matrix where:
↳ each row = A CpG site
↳ each column = A biological sample

At this stage, the dataset contained 485,577 CpG sites across 120 samples.

Data Cleaning & Quality Control Jan 16

- Biological datasets often contain missing values due to technical noise, probe failure, or experimental variability.
  - ↳ Missing values can cause ML algorithms to fail or introduce bias.

First, I calculated the fraction of missing values for each CpG site. I removed CpGs with more than 10% missing data because excessive missingness may indicate unreliable measurement.
  - ↳ Improves overall data quality.

For the remaining CpGs, I performed median imputation.
- ↳ Imputation is the process of replacing missing values with estimated values.
- ↳ I chose the median methylation value across samples for each CpG because the median is robust to outliers.
- ↳ In methylation data, tumour samples can sometimes have extreme values, & using the mean could distort the distribution.

After cleaning, the dataset retained ≈ 485 200 CpG sites, preserving nearly full genome-wide coverage.

Data transformation for ML Jan 16

The original matrix was structured as CpGs × samples. However, ML models require data in the format samples × features, where:

↳ Each row = 1 patient sample
↳ Each column = 1 feature (CpG site)

I therefore transposed the matrix
↳ flipping rows & columns

This step is critical because ML algorithms interpret rows as independent observations.

Label extraction & Alignment

Jan 17

To train a supervised ML model, I needed labels indicating whether each sample was tumour or normal.
↳ These labels were stored in the meta-data section of the series matrix file

I programmatically extracted the sample IDs (GSM identifiers) & the corresponding group labels (Tumour or Normal).

I then cleaned the labels by removing extra quotation marks & whitespace to ensure consistency.

Label alignment is extremely important because if sample labels do not match the correct columns in the methylation matrix, the model may learn incorrect relationships.
- I verified that:
  ↳ Number of labels matched number of samples
  ↳ There were 80 tumour & 40 normal
  ↳ The order of labels = exactly to the order of sample columns.

Finally, I converted tumour samples to 1 & normal samples to 0, creating a binary classification target variable.

High-Dimensional Feature Reduction

Jan 17

After cleaning & transposing the dataset, each sample contained approximately 485,200 CpG methylation features. However, I only had 120 biological samples.

↳ This creates a "high-dimensional" problem, where the number of features greatly exceeds the number of samples.

- In this situation, ML models can easily memorize noise rather than learn real biological patterns (overfitting).

↳ To prevent this, I implemented a 2-stage feature reduction process:

① VARIANCE THRESHOLD FILTERING

- I applied a VarianceThreshold filter with a threshold of 1e-5

- Variance measures how much a feature changes across samples. If a CpG site has extremely low variance, it means its methylation level is almost identical in all samples, regardless of whether the tissue is tumor or normal.
  ↳ Such CpGs cannot help distinguish between class & add unnecessary noise.

The threshold value of 1e-5 (0.00001) was chosen because methylation beta values range from 0 to 1. A variance < 0.00001 indicates almost no meaningful variation across samples.
↳ Setting the threshold this low ensures that I only remove CpGs that are truly uninformitive, while retaining biologically meaningful variability.

→

② ANOVA F-TEST

: After removing near-constant CpGs, I applied an ANOVA F-test using selectKBest to identify CpGs that differ most strongly between tumour and normal groups.

The ANOVA F-Test compares:
↳ The difference in average methylation between tumour & normal samples.
↳ The variability within each group.

For each CpG, the F-statistic measures how clearly that CpG seperates the two groups. If tumour samples have consistently higher (or lower) methylation than normal samples, the CpG will recieve a high F-score.
↳ If both groups overlap heavily, the F-score will be low.

I calculated the F-score for every CpG & ranked them from highest to lowest. I then selected the top 2000 CpGs with the strongest seperations.
↳ This reduced the dataset from ≈ 485,200 CpGs to 2,000 highly informative CpGs.

Model Training Jan 18

After reducing the feature space to 2000 CpGs, I trained a logistic regression classifer to predict whether a sample was tumour or normal.

LOGISTIC REGRESSION: Supervised ML algorithm used for binary classification problems. It works by assigning a weight (coefficien) to each feature. The model multiplies each CpG methylation value by its corresponding weight, sums them together, & converts the result into a probability between 0 & 1 using a sigmoid function.

↳ Positive coefficient: higher methylation increases the likelihood of tumour classification
↳ Negative coefficient: higher methylation is associated with normal tissue.
↳ Coefficent of 0: CpG does not influence model

L1 REGULARIZATION:
- To prevent overfitting & improve interpretability
- Adds a penalty proportional to the the absolute value of the coefficients.
  ↳ forces weak coefficients to become exactly zero, effectively removing unimportant CpGs from the model.

Although the model began with 2000 CpGs after feature selection, L1 regularization reduced this further to only 87 CpGs with non-zero coefficients.

This makes my model:
↳ Simpler
↳ More interpretable
↳ Less likely to memorize noise
↳ More biologically meaningful.

Regularization strength ($C = 0.2$)

- In logistic regression, the parameter C controls how strong the regularization penality is. Smaller values of C correspond to stronger regularization.

↳ I set $C = 0.2$, because it applies stronger regularization than the default value of 1.0.

- Encourages model to be more selective & only retain CpGs with strong predictive power.

BALANCED CLASS WEIGHTS

- Because the dataset contains 80 tumour samples & 40 normal samples, the classes are imbalanced. Without correction, the model can favour the tumour class simply because it appears more frequently.

- By setting class_weight = "balanced," the algorithm adjusts the importance of each class during training.

↳ This ensures that mistakes on normal samples are penalized appropriately, leading to fairer & more reliable classification.

Model Evaluation on held-out test set Jan 18

After training the logistic regression model, I evaluated its performance on a held out test set containing 25% of the data. This resulted in 30 test samples.
- ↳ 20 tumour
- ↳ 10 normal

The model achieved the following results
- ↳ ROC AUC = 1.00
- ↳ PR AUC = 1.00
- ↳ Accuracy = 100%

The confusion matrix showed:
- ↳ 10 normal samples correctly classified as normal
- ↳ 20 tumour samples correctly classified as tumour
- ↳ 0 false pos.
- ↳ 0 false neg.

ROC AUC:
- ↳ Measures how well the model seperates tumour from normal samples across all possible probability thresholds.

PR AUC:
- Precision - Recall AUC
  - ↳ Precision: Of all the samples predicted as tumour, how many were actually tumour.
  - ↳ Recall: Of the true tumour samples, how many were correctly detected.

These results indicate extremely strong seperation between tumour & normal methylation patterns in the held-out data. However, perfect performance can sometimes occur due to chance or favourable splitting... additional validation steps were necessary to confirm generalizability.

Stratified 5-Fold Cross Validation Jan 19

To test whether the model was overfitting or memorizing the training data, I performed stratified 5-fold cross-validation.

How it works:
1. Splitting the dataset into 5 subsets
2. Training the model on 4 subsets
3. Testing on the remaining subset
4. Repeating this process 5 times so each subset serves as the test set once.

Stratified means that each fold maintained the same tumor-to-normal ratio as the full dataset.

Importantly, the entire pipeline, including variance filtering, ANOVA feature selection, scaling, & logistic regression was placed inside the cross-validation loop.
↳ Prevents data leakage, meaning no info from the test fold influenced feature selection.

RESULTS:
ROC AUC scores across the 5 folds:
↳ 1.00
↳ 1.00
↳ 0.9766
↳ 1.00
↳ 1.00

Mean ROC AUC:
0.9993 ± 0.0094

PR AUC scores:
↳ 1.00
↳ 1.00
↳ 0.9883
↳ 1.00
↳ 1.00

Mean PR AUC:
0.9977 ± 0.0047

Even when the data was repeatedly split in different ways, the model consistently achieved near-perfect performance.

The slight variation shows that performance is not artificially locked at 1.00, which suggests realistic generalization rather than pure memorization.

Top 20 CpG feature extraction + Heatmap

Jan 22

TOP 20 CpG Feature Extraction:

- After cross-validation, I wanted to show which CpG sites were the most important for my trained model, so I extracted the strongest features based on the logistic regression coefficients.
- Since feature selection happens in multiple steps, I first recovered the correct CpG names that actually made it into the final model.
- I then extracted the logistic regression coefficients & ranked CpGs by the absolute value of their coefficients, since a larger absolute value means that CpG has a stronger effect on the model's decision.
- Using this ranking, I created a small table containing the top 20 CpGs, along with their coefficients, so I could report the most influential methylation markers.
- Next, I loaded the illumina 450k annotation file so I could attach gene names & region info to each CpG.

HEATMAP VISUALIZATION:

- To build my heatmap, I extracted the beta values for just the top 20 CpGs from my cleaned methylation matrix & reordered the samples so normals appeared first & tumours appeared second.
- I then z-scored each CpG across samples, so all CpGs were on the same scale for visualization
- Finally, I plotted the heatmap using imshow() with a colourbar, showing the standarized methylation patterns across samples for the top 20 CpGs.

Permutation Testing Jan 22

Although cross-validation showed strong performance, extremely high AUC values can sometimes occur if a model accidently memorizes patterns specific to the dataset.

To confirm that my model was learning real biological signal rather than random structure, I performed permutation testing.

How it works:
1) Randomly shuffles the tumour/normal labels
2) Keeps methylation data the exact same
3) Retrains entire pipeline
4) Measures cross-validated ROC AUC

↳ By shuffling labels, the true biological relationship between methylation patterns & cancer status is destroyed.

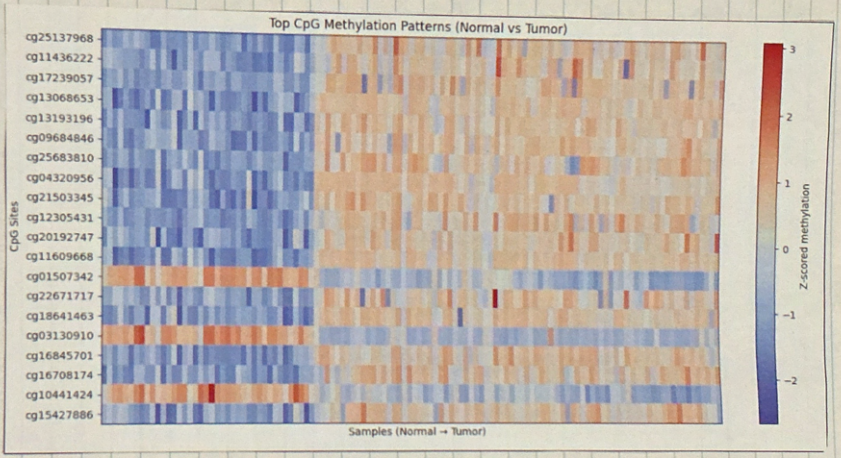
- If the model still performs well after shuffling, it would indicate overfitting or data leakage.
- If performance drops towards 0.5 (random guessing), it confirms that the original model was learning real signal.

RESULTS:

↳ After performing 25 permutations, the ROC AUC scores dropped dramatically toward chance levels.
↳ The permutation ROC AUC mean was approx:
~0.5 (with small variation)

* This confirms that when the biological relationship is removed, the model cannot classify samples effectively.

	CpG	UCSC_RefGene_Name	RegionType	Tumor_Methylation	Coefficient	Cancer_Related_Gene
0	cg25137968	HIST2H2BA	Promoter	Hypermethylated in Tumor	0.146332	Yes
1	cg11438222	HIST1H3E	Promoter	Hypermethylated in Tumor	0.108071	Yes
2	cg17239057	NaN	Intergenic	Hypermethylated in Tumor	0.107967	Unknown / Intergenic
3	cg13068653	TAGLN;TAGLN	Gene body / Other	Hypermethylated in Tumor	0.103593	Yes
4	cg13193196	IMMP2L	Gene body / Other	Hypermethylated in Tumor	0.101761	Yes
5	cg09684846	FAM18A	Promoter	Hypermethylated in Tumor	0.083109	Possibly / Unknown
6	cg25683810	NaN	Intergenic	Hypermethylated in Tumor	0.082379	Unknown / Intergenic
7	cg04320956	HAS3;HAS3	Gene body / Other	Hypermethylated in Tumor	0.079678	Yes
8	cg21503345	NaN	Intergenic	Hypermethylated in Tumor	0.077985	Unknown / Intergenic
9	cg12305431	HOXA3;HOXA3	Promoter	Hypermethylated in Tumor	0.077141	Yes
10	cg20192747	NaN	Intergenic	Hypermethylated in Tumor	0.075405	Unknown / Intergenic
11	cg11609668	TAGLN;TAGLN	Gene body / Other	Hypermethylated in Tumor	0.070063	Yes
12	cg01507342	PITPNC1;PITPNC1	Gene body / Other	Hypomethylated in Tumor	-0.066476	Yes
13	cg22671717	NaN	Intergenic	Hypermethylated in Tumor	0.065312	Unknown / Intergenic
14	cg18641463	MAD1L1;MAD1L1;MAD1L1	Gene body / Other	Hypermethylated in Tumor	0.061756	Yes
15	cg03130910	NaN	Intergenic	Hypomethylated in Tumor	-0.059055	Unknown / Intergenic
16	cg16845701	NaN	Intergenic	Hypermethylated in Tumor	0.057336	Unknown / Intergenic
17	cg16708174	RARRES1;RARRES1	Gene body / Other	Hypermethylated in Tumor	0.056524	Yes
18	cg10441424	NaN	Intergenic	Hypomethylated in Tumor	-0.051610	Unknown / Intergenic
19	cg15427886	NaN	Intergenic	Hypermethylated in Tumor	0.051339	Unknown / Intergenic


Top CpG Methylation Patterns (Normal vs Tumor)
cg25137968
cg11436222
cg17239057
cg13068653
cg13193196
cg09684846
cg25683810
cg04320956
cg21503345
cg12305431
cg20192747
cg11609668
cg01507342
cg22671717
cg18641463
cg03130910
cg16845701
cg16708174
cg10441424
cg15427886
CpG Sites
Samples (Normal → Tumor)
Z-scored methylation
3
2
1
0
−1
−2

Biological Interpretation / results Jan 23

several of the top CpGs were located near genes known to be involved in cell cycle regulation, chromatin structure, & cancer-related pathways.

Notably, CpGs were associated with genes such as:
- ↳ HOXA3
- ↳ TAGLN
- ↳ MAD1L1
- ↳ RARRES1
- ↳ HIST1H3G
- ↳ HIST2H2BA
- ↳ HAS3

- Many of these genes have documented roles in cellular differentiation, proliferation, tumour suppression, or chromatin organization. The presence of these genes among the strongest predictive CpGs suggests that my model is capturing cancer-associated methylation changes.

several top CpGs were located in promoter regions. Promoter methylation is particularly important because hypermethylation in promoter regions is often associated with gene silencing.

Other CpGs were located in gene body regions. Gene body methylation can influence transcriptional regulation & chromatin accessibility, & abnormal patterns have also been linked to cancer progression.

Most of the top 20 CpGs were hypermethylated in tumour samples, meaning tumour tissue showed higher methylation levels compared to normal tissue at those sites.

A smaller subset showed hypomethylation in tumour tissue.

Cancer is known to exhibit both:
- ↳ Global hypomethylation (increases genomic instability)
- ↳ Local promoter hypermethylation (silences tumour suppressor genes).

The heatmap showed clear clustering of tumour samples together & normal samples together, with strong colour differences between the two groups.
- ↳ This just further proved my quantitative results & strengthened the results that tumour & normal tissue have distinct methylation signatures.