
Introduction to Bioinformatics

Project Presentation

Group 4

Introduction

- Breast cancer is one of the most common malignant diseases among women, with a high incidence and mortality rate worldwide. Despite advances in treatment, the heterogeneity of breast cancer presents significant challenges in diagnosis, prognosis, and therapy.
- Biomarkers play a crucial role in improving breast cancer detection, predicting patient outcomes, and guiding personalized treatments. Identifying reliable biomarkers helps classify tumors, determine disease progression, and optimize therapeutic strategies.
- Traditional biomarkers such as ER (Estrogen Receptor), PR (Progesterone Receptor), and HER2 are widely used but have limitations:
 - They fail to fully capture tumor heterogeneity.
 - Some subtypes, like triple-negative breast cancer (TNBC), lack targeted biomarkers, making treatment challenging.
 - Many biomarkers provide limited predictive power for treatment response.

Objective

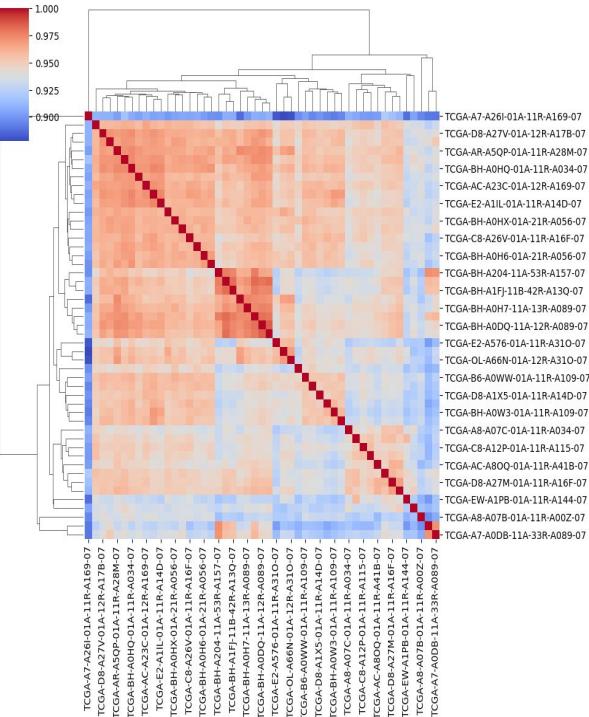
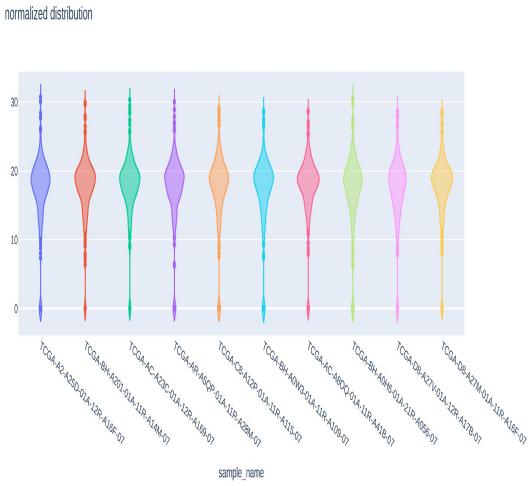
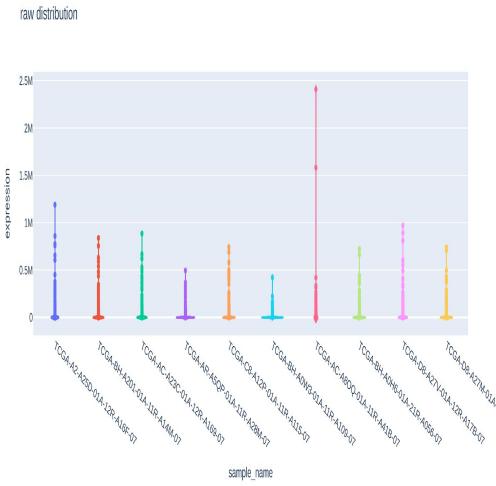
- Identify novel biomarkers for breast cancer using RNA-Seq gene expression analysis and machine learning techniques.
- Select the most relevant genes based on differential expression and feature importance.
- Perform further analysis on selected genes using specialized databases:
 - GEPIA, UALCAN, GeneCards, bc-GenExMiner, cBioPortal, Kaplan-Meier Plotter, and Reactome.

Analyzing RNA-seq Data

- Collecting RNA-seq Data from TCGA
- Preprocessing of RNA-seq Data
- Differential Expression Analysis
- Identifying Key Genes Using Random Forest Algorithm

Preprocessing

- Raw RNA-Seq data requires preprocessing to remove biases.
 - UQ-FPKM
 - Log2 Transform



UQ-FPKM

- Reducing Sensitivity to Highly Expressed Genes
- Avoiding Zero-Inflation Bias in Low-Expression Genes
- Improved Cross-Sample Comparability

Algorithm 1 UQ-FPKM Normalization

Require: Gene expression matrix E of size (Samples \times Genes)

Ensure: Normalized expression matrix E'

```
1: Step 1: Remove Genes with Zero Expression in All Samples
2:  $E_{filtered} \leftarrow$  Remove rows where all values are zero
3: Step 2: Compute Upper Quartile (75th percentile) for Each Sample
4: for each sample  $s$  in  $E_{filtered}$  do
5:    $V_s \leftarrow$  Non-zero values in sample  $s$ 
6:    $UQ_s \leftarrow$  75th percentile of  $V_s$ 
7: end for
8: Step 3: Compute Normalization Factor
9:  $MeanUQ \leftarrow$  Mean of all  $UQ_s$  values
10: Step 4: Normalize Each Gene Expression Value
11: for each sample  $s$  in  $E_{filtered}$  do
12:    $NF_s \leftarrow UQ_s / MeanUQ$ 
13:   for each gene  $g$  in sample  $s$  do
14:      $E'_{s,g} \leftarrow E_{filtered_{s,g}} / NF_s$ 
15:   end for
16: end for
17: Return  $E'$ 
```

Differential Expression Analysis

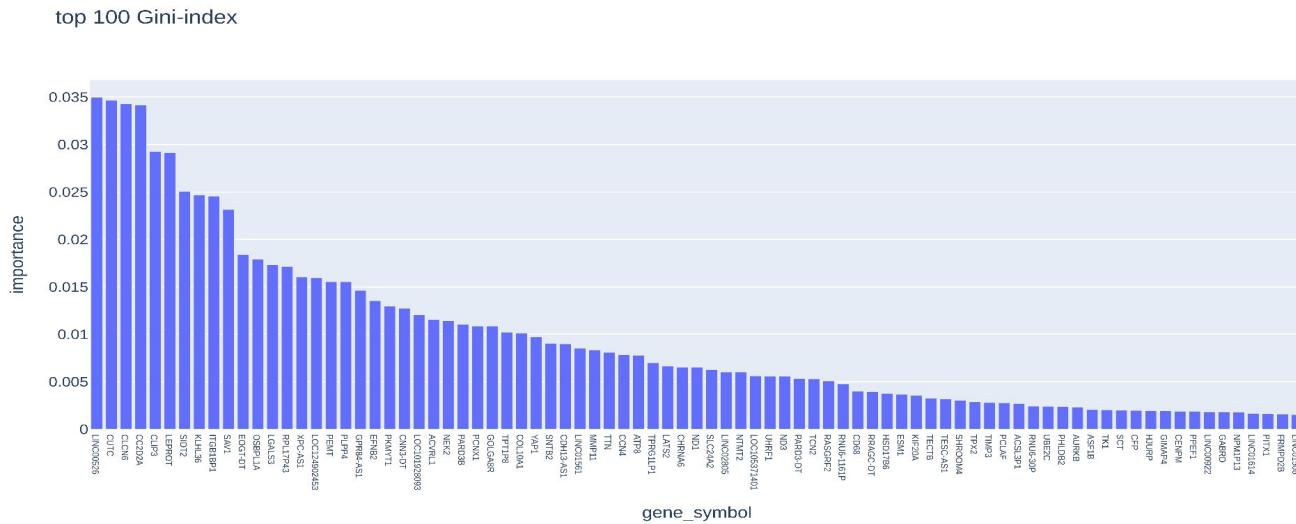
- Identifies genes with significant expression differences between cancerous and healthy tissues.
- $|\log \text{fc}| > 1 \text{ & } |\log \text{fc}| < -1$
- False Discovery Rate < 0.01

Volcano Plot with Highlighted Genes



Identifying Key Genes Using Random Forest Algorithm

- distinguishing between cancerous and healthy samples based on gene expression.
- The Gini Index measures each gene's contribution to classification accuracy.
- Top-100 genes are selected for further analysis.

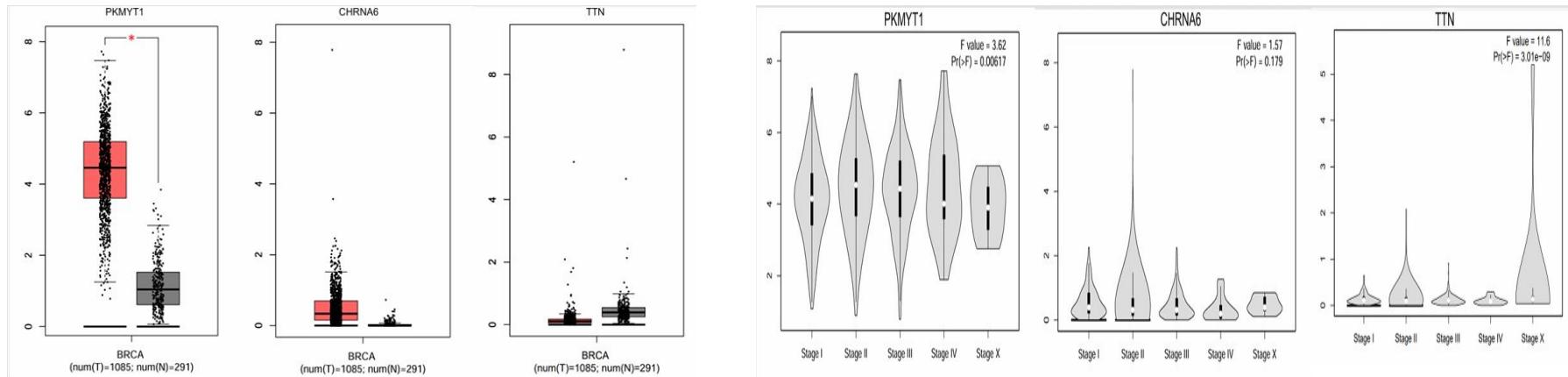


Beyond Expression – Functional and Clinical Insights

- Differential expression alone is not enough to confirm a gene as a biomarker.
- Further analysis is required to:
 - Validate clinical relevance (association with survival, treatment response).
 - Understand biological function (pathways and molecular interactions).
 - Assess genomic alterations (mutations, copy number variations).
- We use multiple databases to ensure biomarker reliability.

GEPIA & UALCAN – Further Expression Analysis

- Provide gene expression data from TCGA and GTEx for both normal and cancerous tissues.
- Allow various comparisons based on:
 - Age, gender, nodal status, and cancer stages.
 - Tumor vs. normal expression levels.
 - Survival analysis.



GeneCards - Prediction of subcellular localization

- Aggregates gene function, protein interactions, and disease associations.
- Used to analyze subcellular localization of selected genes.
- Identifies potential therapeutic targets based on cellular positioning.
 - Membrane proteins: Good drug targets (e.g., antibodies).
 - Nuclear/cytoplasmic proteins: May influence cell signaling and gene regulation.

Localization for PKMYT1 Gene

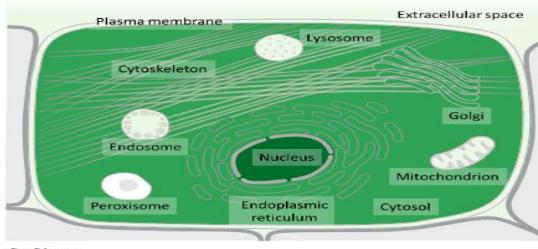
Subsections: Cellular Components

Subcellular locations from UniProtKB/Swiss-Prot for PKMYT1 Gene

Protein: PMYT1_HUMAN

> Endoplasmic reticulum membrane; Peripheral membrane protein {ECO:0000269 PubMed:9001210}
> Golgi apparatus membrane; Peripheral membrane protein {ECO:0000269 PubMed:9001210}

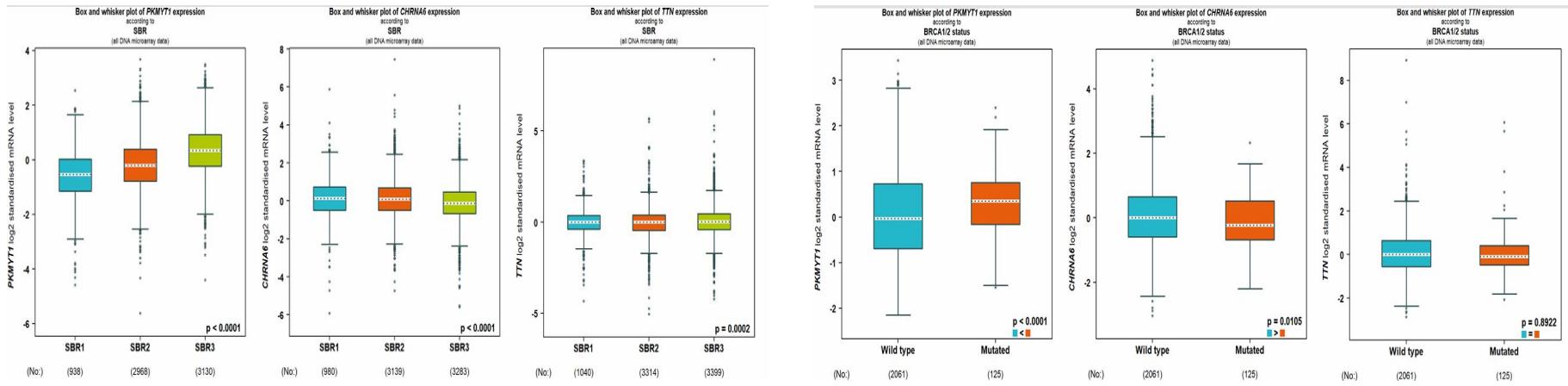
Subcellular locations from COMPARTMENTS ?



Compartment	Confidence
nucleus	5
golgi apparatus	4
cytosol	4
endoplasmic reticulum	4
cytoskeleton	2
lysosome	1
endosome	1
mitochondrion	1
extracellular	1
plasma membrane	1

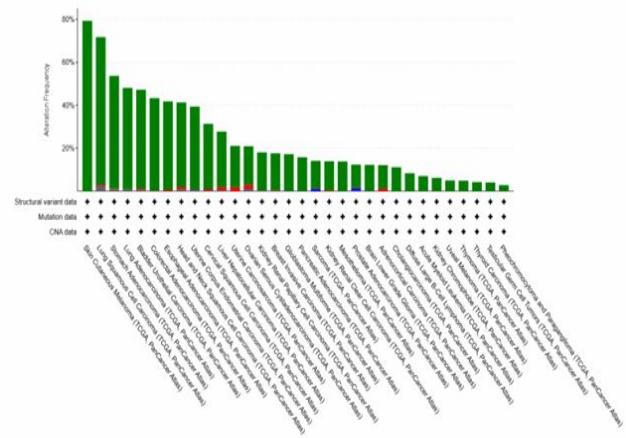
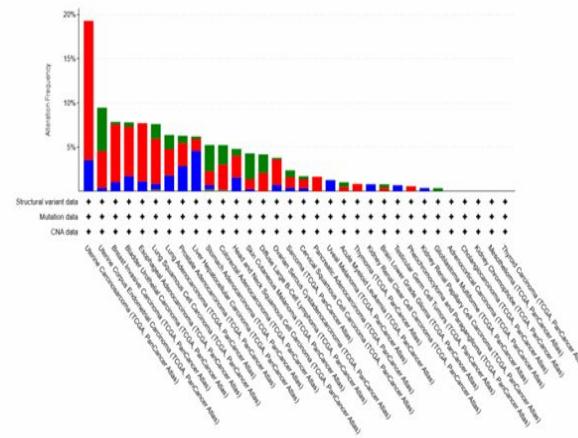
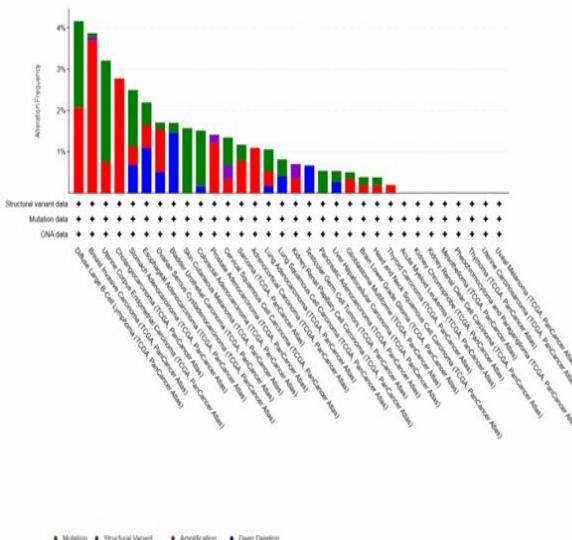
bc-GenExMiner - Clinico-pathological Parameters

- Focuses on breast cancer-specific gene expression.
- Provides analysis based on:
 - SBR (Scarff-Bloom-Richardson) grading system, assessing tumor differentiation.
 - BRCA1/2 mutation status, crucial for hereditary breast cancer risk.
- Helps identify genes linked to tumor aggressiveness and prognosis.



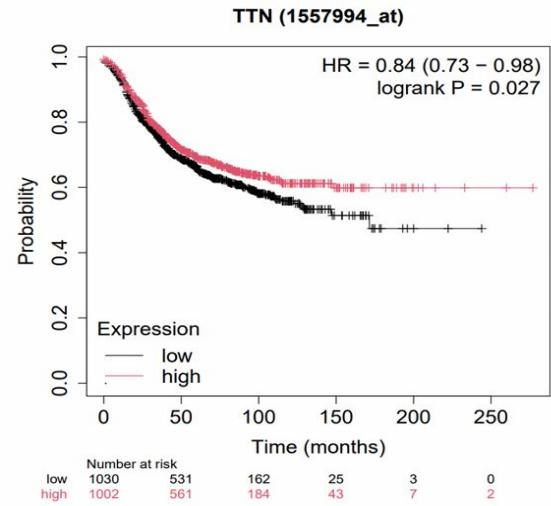
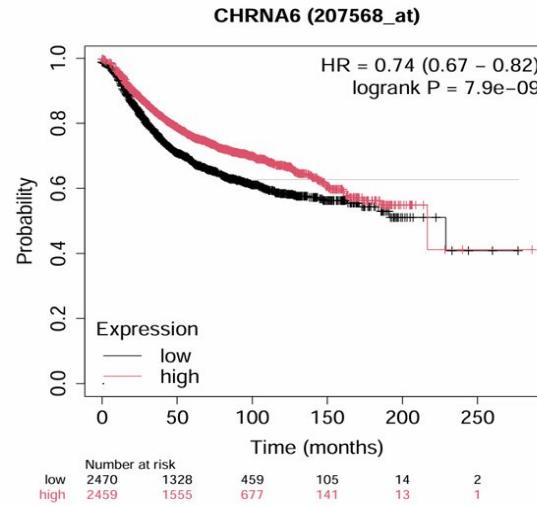
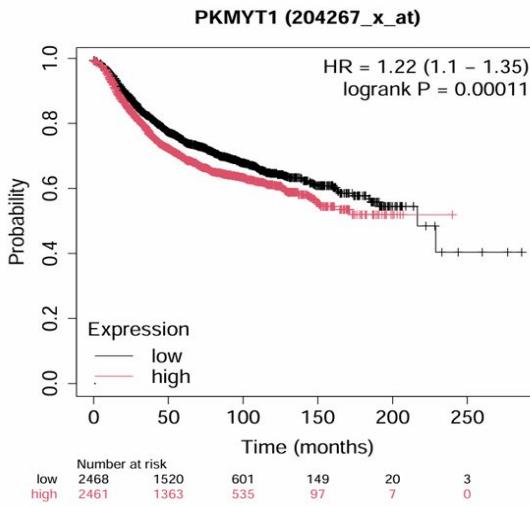
cBioPortal – Genetic Alterations & Somatic Mutations

- Provides insights into genetic alterations in selected genes.
 - Analyzes mutations, copy number variations, and deep deletions.
 - Helps determine whether genomic changes drive cancer progression.



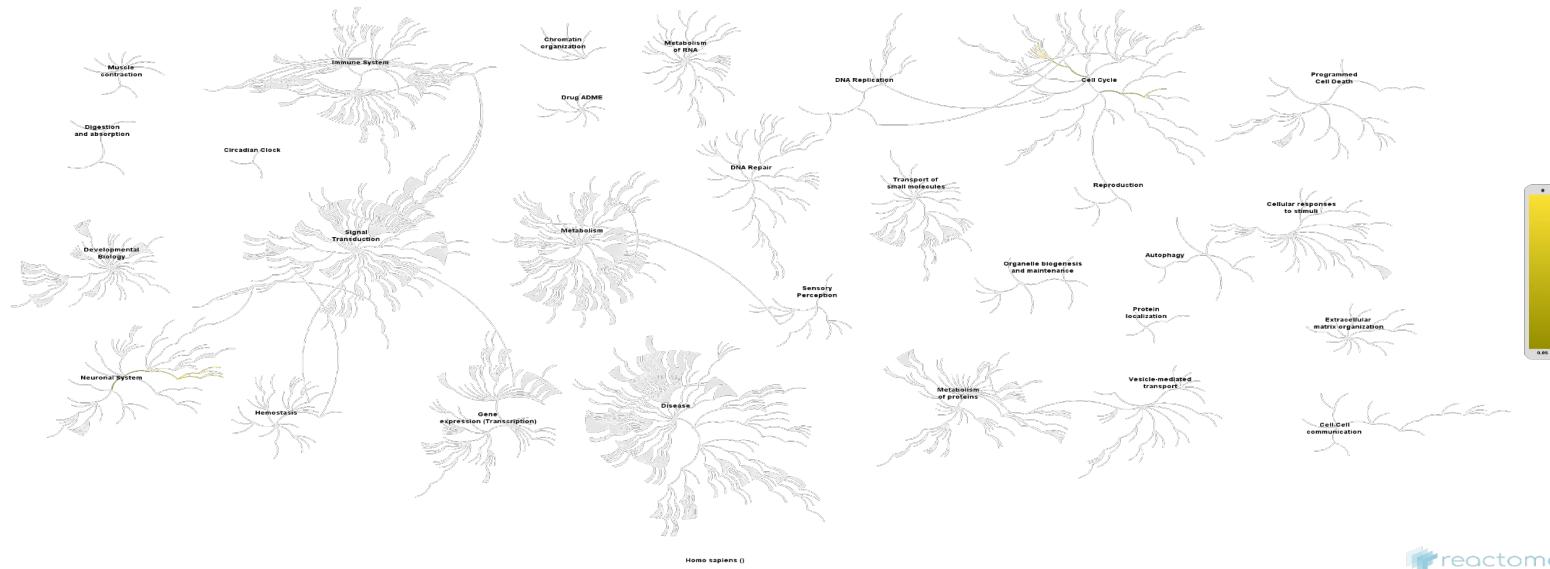
Kaplan-Meier Plotter – Survival Analysis

- Examines the association between gene expression and patient survival.
- Uses Kaplan-Meier survival curves to compare outcomes based on gene expression levels.
- Identifies potential prognostic biomarkers that influence disease progression.

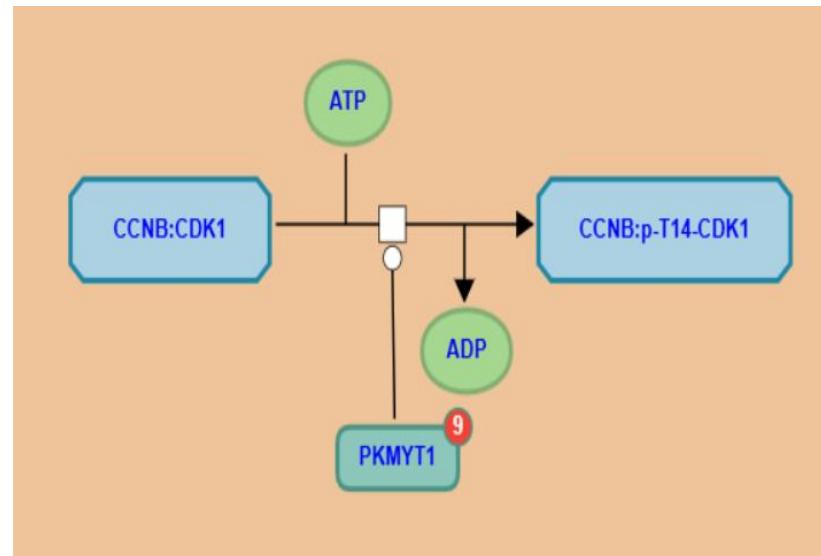
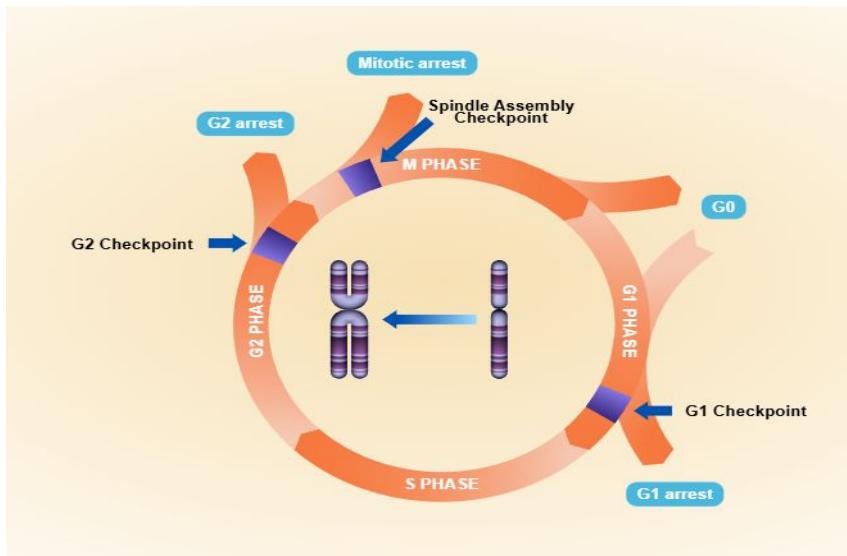


Reactome – Understanding Pathways

- Understanding pathways helps identify key regulatory genes involved in tumor development.
- Reactome Maps selected genes to known signaling and metabolic pathways.
- It also Helps determine whether a gene plays a critical role in cancer biology.



Cont.





THANK YOU FOR YOUR ATTENTION!