1. Title Page

- **Title of the Project**: Uncovering Key Differentially Expressed Genes in Breast Cancer Subtypes Using RNA-Seq

- **Author(s)**: Anisha Kumari , **email**: meetanishakumari2210@gmail.com
- **Organization/Industry**: Mapmygenome

- **GitHub/Repository Link**: Include links to workflows, scripts, or datasets.

2. Abstract

- **Background:** Breast cancer is a heterogeneous disease with distinct molecular subtypes, such as Basal-like and Luminal, which exhibit different clinical outcomes and therapeutic responses. Understanding the transcriptomic variations between these subtypes is critical for identifying novel biomarkers and precision medicine targets.
- **Objective:** The primary aim of this project was to identify differentially expressed genes (DEGs) and enriched metabolic/signaling pathways in Basal-like versus Luminal breast cancer samples to uncover potential therapeutic vulnerabilities.
- **Methods:** RNA-Seq data from 6 samples (3 Basal, 3 Luminal) were processed using a Galaxy-based pipeline. Key steps included quality control (FastQC), alignment (HISAT2), and quantification (featureCounts). Differential expression analysis was performed using **DESeq2**, followed by functional enrichment via **goseq** (GO/KEGG) and pathway visualization using **Pathview**.
- **Results:** Analysis revealed 972 significant DEGs (P_{adj})< 0.05, with 489 genes upregulated and 483 downregulated in the Basal subtype. **KRT16** and **PTCHD1** were identified as the top upregulated markers, while **TFF1** and **ERBB4** were the most significantly downregulated. Pathway analysis highlighted significant enrichment in **Vitamin Digestion and Absorption** (driven by **FOLH1**) and Calcium **Signaling** (driven by **CACNA1B** and **CALM**).

- **Conclusion:** The findings suggest that Basal-like breast cancer achieves aggressive growth through metabolic reprogramming of vitamin uptake and hijacked calcium signaling. These identified genes, particularly **FOLH1(PSMA)**, and **CACNA1B** represent promising candidates for targeted radioligand or inhibitory therapies.

---

3. Introduction

- **Background:** Breast cancer remains a leading cause of oncology-related mortality worldwide. The Basal-like (often Triple-Negative) subtype is particularly aggressive, characterized by high proliferation rates and a lack of traditional hormone receptor targets, unlike the more manageable Luminal subtypes **[1]**.
- **Problem Statement:** While traditional chemotherapy is the standard for Basal-like cancers, resistance is common. There is an urgent industrial and scientific need to identify subtype-specific cell-surface markers and signaling "hubs" that can be targeted with next-generation biologics or small-molecule inhibitors.
- **Objectives:**

1. To execute a complete bioinformatics pipeline to compare the transcriptomic landscapes of Basal-like and Luminal breast cancer.
2. To identify a statistically robust set of Differentially Expressed Genes (DEGs) that define the molecular shift between these subtypes.
3. To perform functional enrichment analysis to pinpoint biological pathways that could serve as potential therapeutic vulnerabilities.

---

4. Materials and Methods

- **Data Sources**
  - Original Dataset and Origins: The raw transcriptomic data was sourced from the NCBI Gene Expression Omnibus (GEO) under accession number GSE52194. The study, titled "mRNA-sequencing of breast cancer subtypes and normal tissue," provides a high-resolution map of the digital transcriptome across distinct breast cancer subtypes.
  - Sample Selection: While the original series contains 17–20 biological replicates, this project utilized a targeted subset of 6 paired-end datasets. To ensure a high-contrast comparison of molecular drivers, 3 Basal-like (TNBC) and 3 Luminal (non-TNBC) samples were selected for the analysis.
  - Data Retrieval and Integrity: Raw sequences were retrieved from the NCBI Sequence Read Archive (SRA) using the fasterq-dump utility. This tool was selected to ensure the high-fidelity extraction of FASTQ files and associated Phred quality scores, which are critical for standardizing downstream quality control (QC).
  - Reference Genome: Alignment was conducted against the Human Reference Genome **(hg38/GRCh38)**. The project utilized pre-indexed genome files within the Galaxy environment to ensure optimized compatibility and reproducible mapping with the HISAT2 and featureCounts algorithms.
- **Tools and Software**
  - Galaxy Platform: All computational tasks were executed on the Galaxy Project web-based ecosystem **[8]**, utilizing its integrated history management to ensure full reproducibility and auditability of the data pipeline.
  - **Bioinformatics Toolkit:**
  - FastQC: Employed for comprehensive quality control, assessing per-base sequence quality, adapter content, and sequence duplication levels.
  - HISAT2: Utilized for high-speed alignment, mapping raw reads to the **hg38/GRCh38** reference genome via a hierarchical indexing strategy **[9]**.
  - featureCounts: Used for highly efficient read summarization to assign mapped reads to specific genomic features (Exons/Genes).

  - DESeq2: A biostatistical framework used for normalization and differential expression analysis, employing a negative binomial distribution model to account for overdispersion in RNA-Seq data **[10]**.
  - goseq & Pathview: Utilized for Gene Ontology (GO) and KEGG pathway enrichment analysis **[11]**, providing high-resolution visualization of molecular shifts on metabolic diagrams **[12]**.

- **Pipeline/Workflow**
  1. **Retrieval:** fasterq-dump → 2. **QC:** FastQC → 3. **Alignment:** HISAT2 → 4. **Quantification:** featureCounts → 5. **DE Analysis:** DESeq2 → 6. **Functional Mapping:** goseq & Pathview.

- **Experimental Design:**
  - Preprocessing and Quality Control: Every dataset underwent stringent QC via FastQC. Only reads maintaining an average Phred score > 30 were progressed to alignment, mitigating the risk of sequencing artifacts biasing the differential expression results.
  - Main Analytical Methods: A "Basal vs. Luminal" contrast was executed in DESeq2. The analysis incorporated size-factor normalization for library depth and dispersion estimates to account for biological variability between replicates.
  - Statistical Filtering: To define the high-confidence set of 972 DEGs, a rigorous threshold was applied: an adjusted p-value (P_{adj})< 0.05 (Benjamini-Hochberg corrected) and a $|\text{Log}_2\text{Fold Change}| \geq 1.0$.
  - Validation and Benchmarking: Internal validation was performed via Principal Component Analysis (PCA) and hierarchical clustering (Heatmaps) to verify subtype-specific sample grouping. The pipeline's accuracy was further benchmarked by recovering established biomarkers **KRT16** (Basal-specific) **[5]** and **TFF1/ERBB4** (Luminal-specific) **[6]**.

5. Results

- **Key Findings**:
    - Technical Quality: Initial quality control via FastQC confirmed high-confidence base calling (Phred > 30) across all 12 sequencing files.
    - Alignment Efficiency: Mapping to the **hg38** reference genome using HISAT2 was highly consistent, with overall alignment rates ranging from 87.11% to 89.83%, ensuring that the downstream analysis is based on a robust set of mapped reads.
    - Global Separation: Principal Component Analysis (PCA) demonstrated a clear separation between groups, with PC1 accounting for 51% of the total variance, validating the biological distinction between Basal and Luminal subtypes.
    - Differential Expression: The differential expression analysis (DEA) successfully differentiated the transcriptomic landscapes of the two breast cancer subtypes. Out of 28,395 genes analyzed, 972 met the statistical threshold for significance ($P_{adj} < 0.05$). The symmetry of the results 489 upregulated and 483 downregulated genes indicates a high-quality normalization process and a robust biological signal.
    - The complete list of all 972 differentially expressed genes, including their Fold Change and P-values, is provided in the Appendix at the end of this report.
    - Functional Pathways: Enrichment analysis identified significant alterations in structural integrity markers and specialized metabolic routes **[4]**. Specifically, KEGG mapping highlighted the Vitamin digestion and absorption and Calcium signaling pathways as key metabolic features of the Basal phenotype**[3, 4]**.

- **Tables**:

  - **Table 1: HISAT2 Alignment and Mapping Statistics**

    Summary of sequence alignment results for Basal and Luminal breast cancer subtypes. The data demonstrates uniform mapping efficiency, confirming successful integration with the **hg38** reference genome.

Sample ID and Subtype	**Total Input Reads**	**Overall Mapping Rate(%)**
Basal_Sample_1	44.4 million	89.83%
Basal_Sample_2	18.4 million	87.83%
Basal_Sample_3	53.9 million	87.11%
Luminal_Sample_1	64.3 million	88.87%
Luminal_Sample_2	51.5million	88.52%
Luminal_Sample_3	55.7 million	88.10%

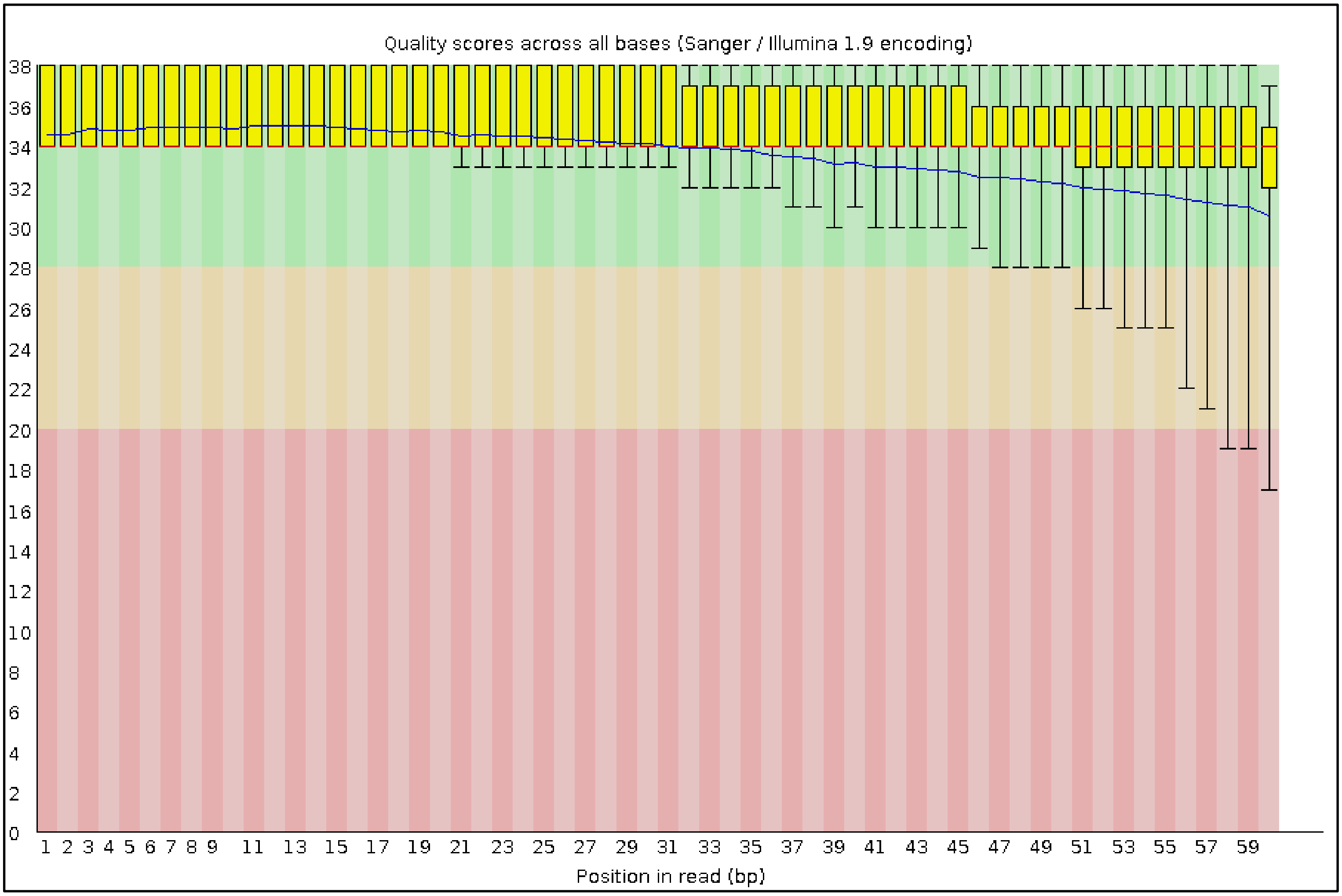
- **Table 2: Representative Differentially Expressed Genes(Basal vs. Luminal)**

  Summary of key genes showing significant fold changes and ($P_{adj} < 0.05$), categorized by regulation status and biological significance.

Gene Symbol	Log2 Fold Change	Adjusted p-value	Regulation	Significance
KRT16	+6.76	$< 10^{-18}$	Upregulated	Top Basal-Type Biomarker
PTCHD1	+5.81	$< 10^{-13}$	Upregulated	Novel Basal Hedgehog Signaling Target
FOLH1	+1.74	0.02	Upregulated	Vitamin Metabolism Target(Basal)
TFF1	-6.87	$< 10^{-19}$	Downregulated	Top Luminal-type Biomarker
ERBB4	-6.24	$< 10^{-15}$	Downregulated	Critical Luminal Growth Factor Receptor
CACNA1B	+1.87	0.01	Upregulated	Calcium Signaling Target

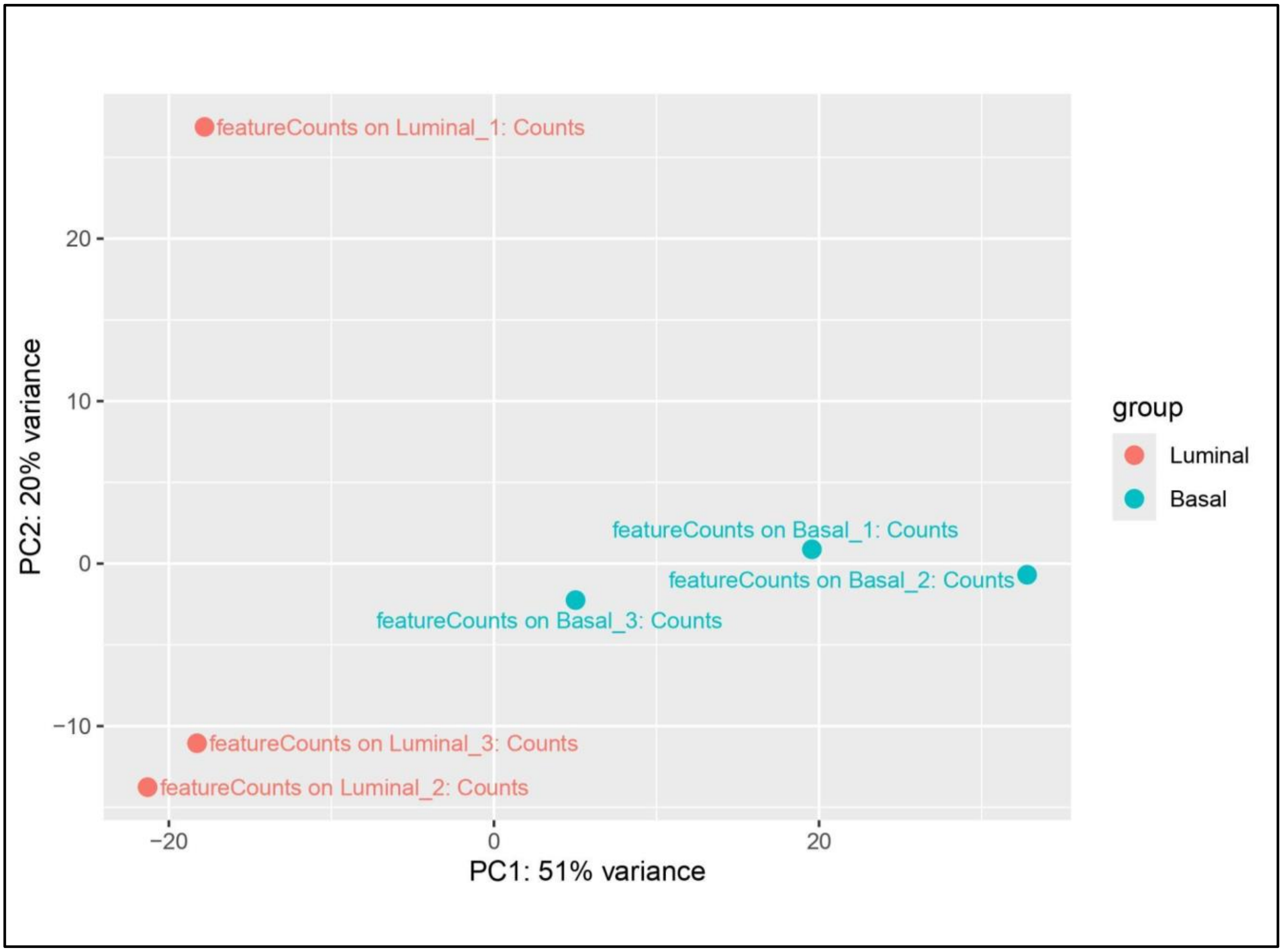
Figures/Graphs:

- Fastqc Report



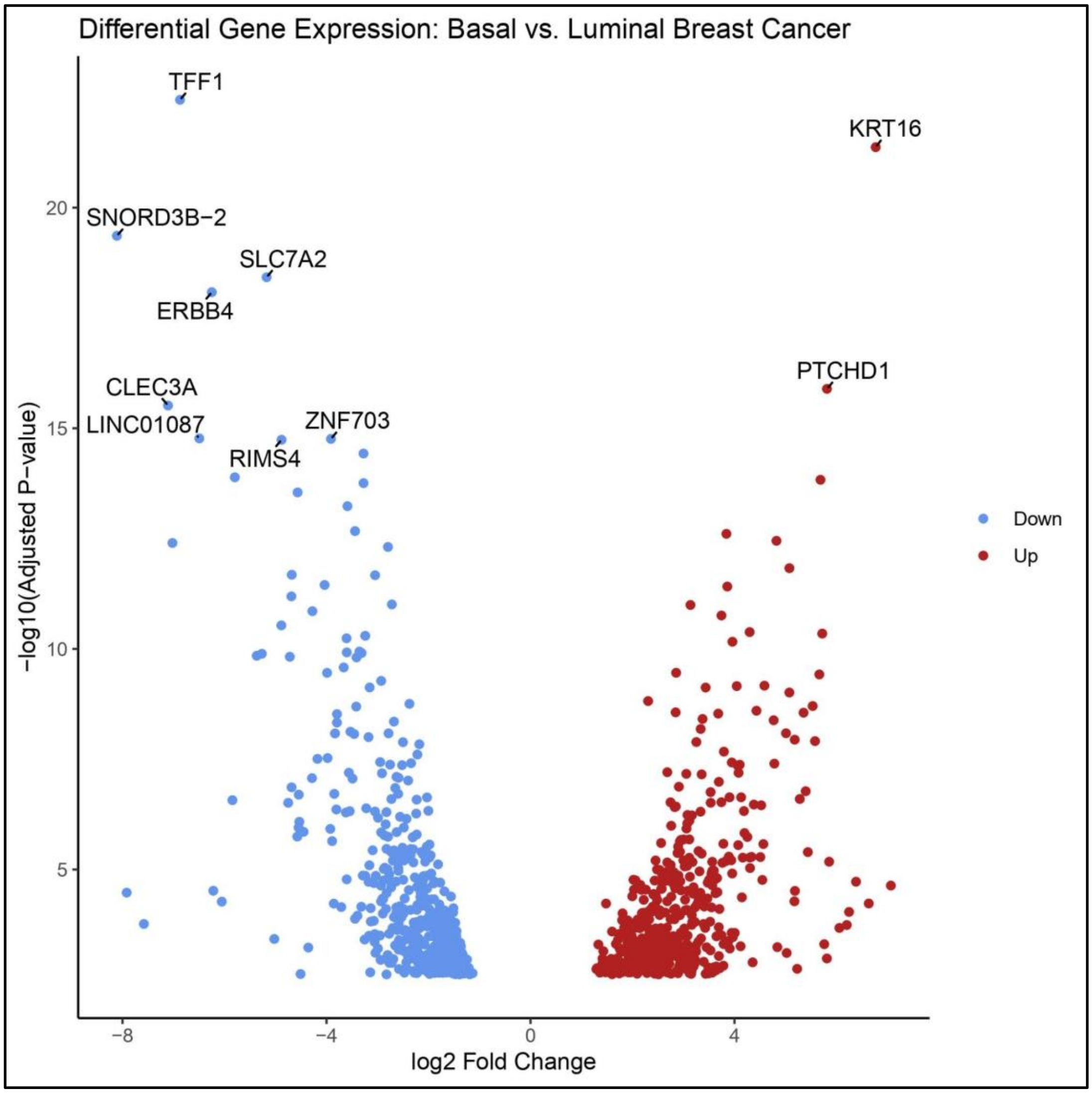
- **Figure 1: Representative Sequence Quality Profile(FastQC)** : The per-base quality scores remain consistently in the "green zone" (Phred > 30), indicating high-accuracy base calling. This technical validation ensures that the genetic variations identified in subsequent steps are biological rather than technical artifacts.

- PCA Plot



- **Figure 2: Principal Component Analysis (PCA) :** The PCA plot shows a clear molecular separation between Basal (teal) and Luminal (red) samples. Principal Component 1 (PC1) accounts for 51% of the total variance, confirming that the biological subtype is the primary driver of gene expression differences in this study.

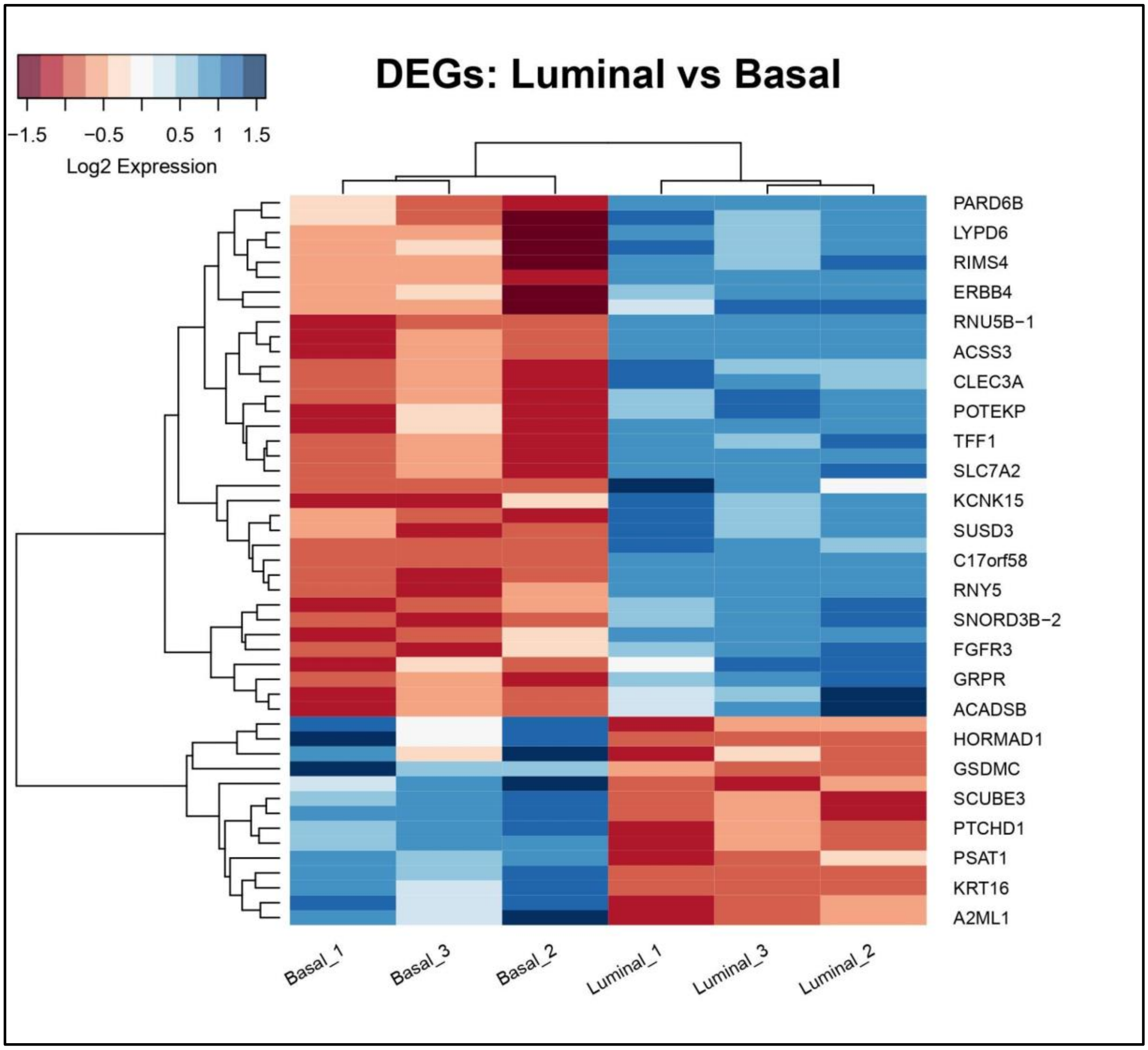
- Volcano Plot



- **Figure 3: Volcano Plot of Differential Expression**: This visualization displays the global distribution of the 972 DEGs. The X-axis (Log2 Fold Change) reveals the extreme upregulation of **KRT16** and **PTCHD1** in Basal samples, while the Y-axis (-

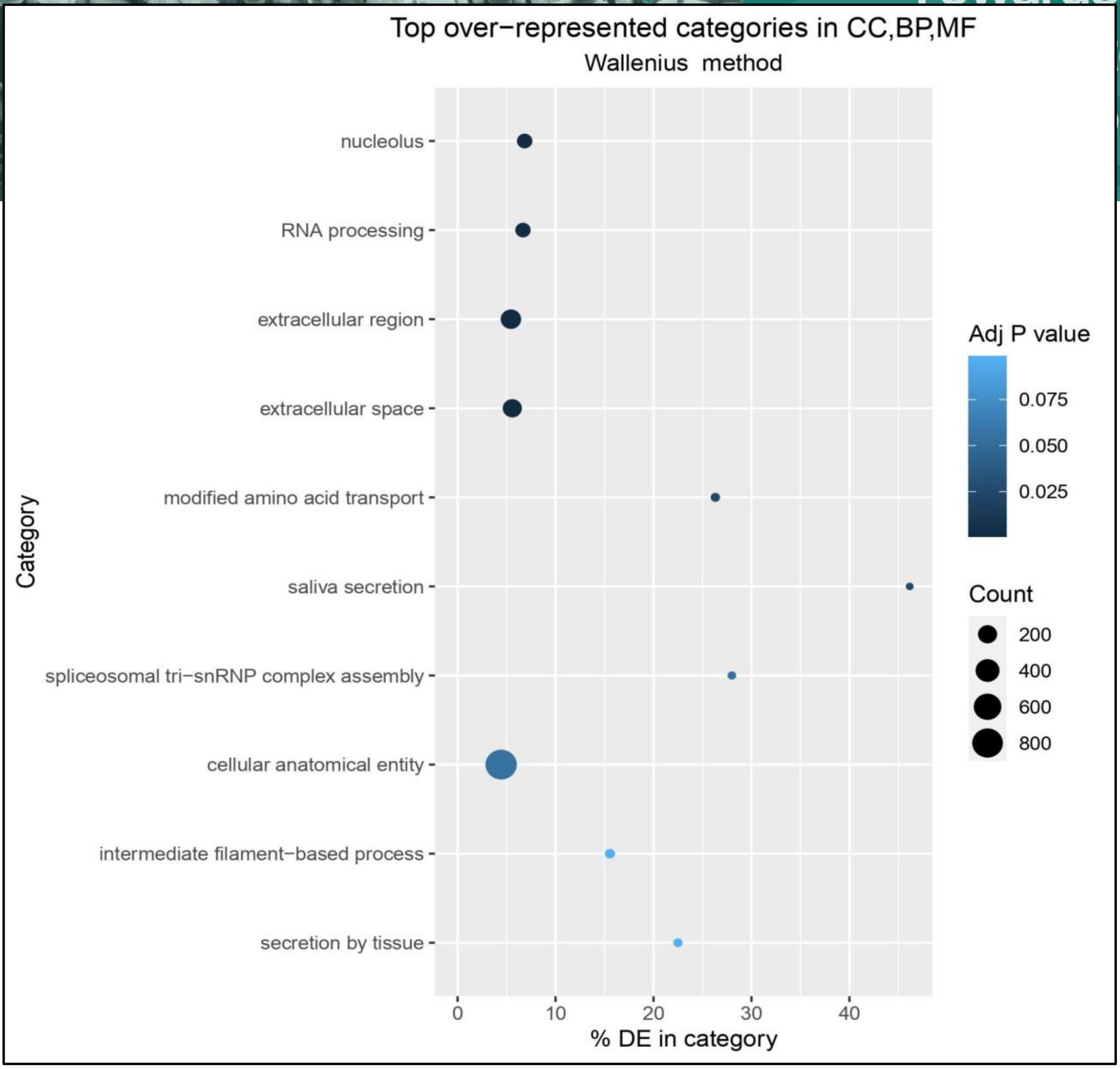
log10 P-value) confirms the high statistical confidence in the downregulation of **TFF1** and **ERBB4**.

- Heatmap



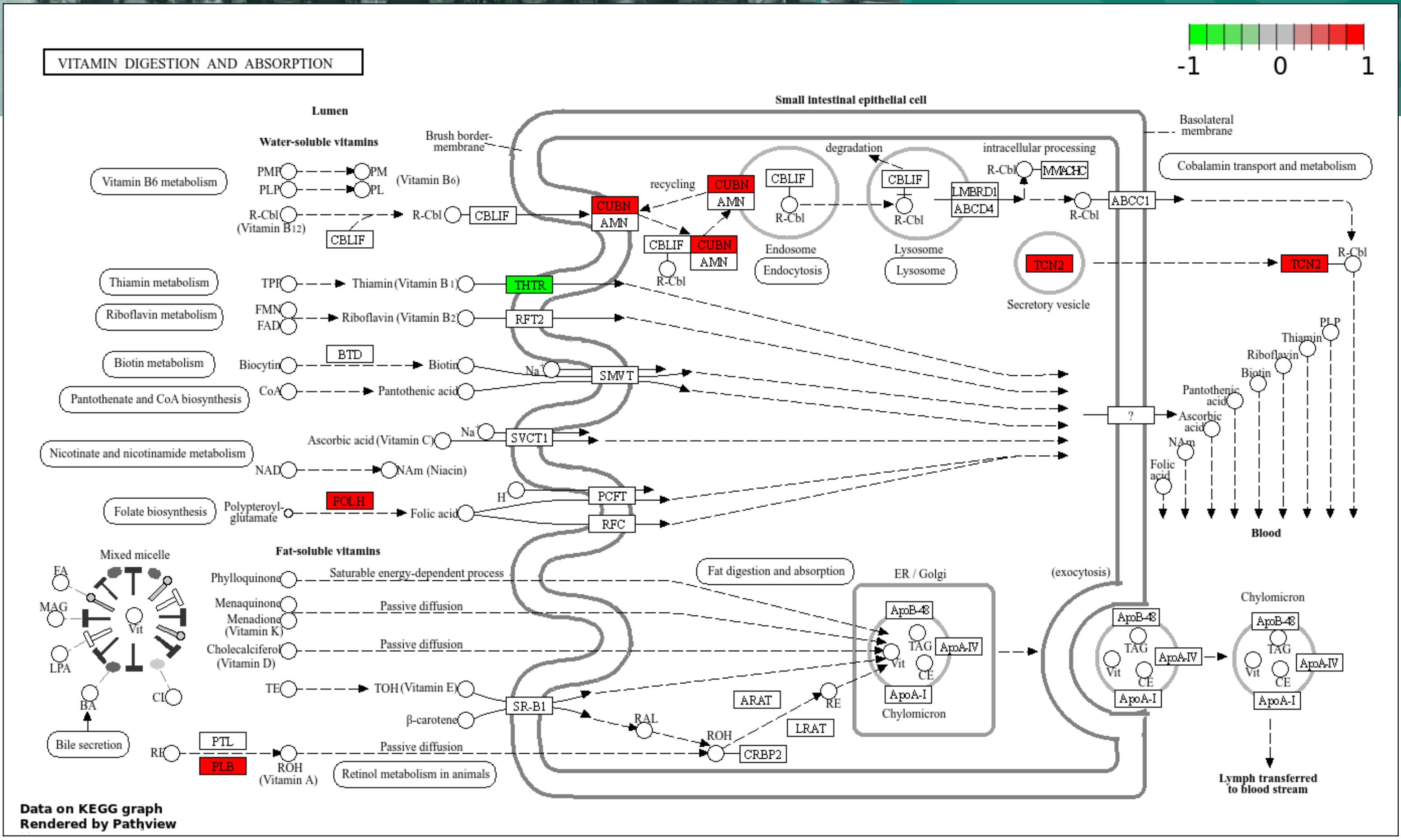
- **Figure 4: Hierarchical Clustering Heatmap**: The heatmap shows a perfect binary split between the 6 samples. The 3 Basal replicates(Samples 1-3) cluster together with high expression of **PTCHD1**, while the 3 Luminal replicates (Samples 4-6) show high expression of hormone-responsive genes like **ERBB4** and **TFF1**, validating the experimental design.

- GO Plot
- **Figure 5: Gene Ontology (GO) Enrichment Analysis:** This dot plot illustrates significantly over-represented biological processes. The enrichment in structural

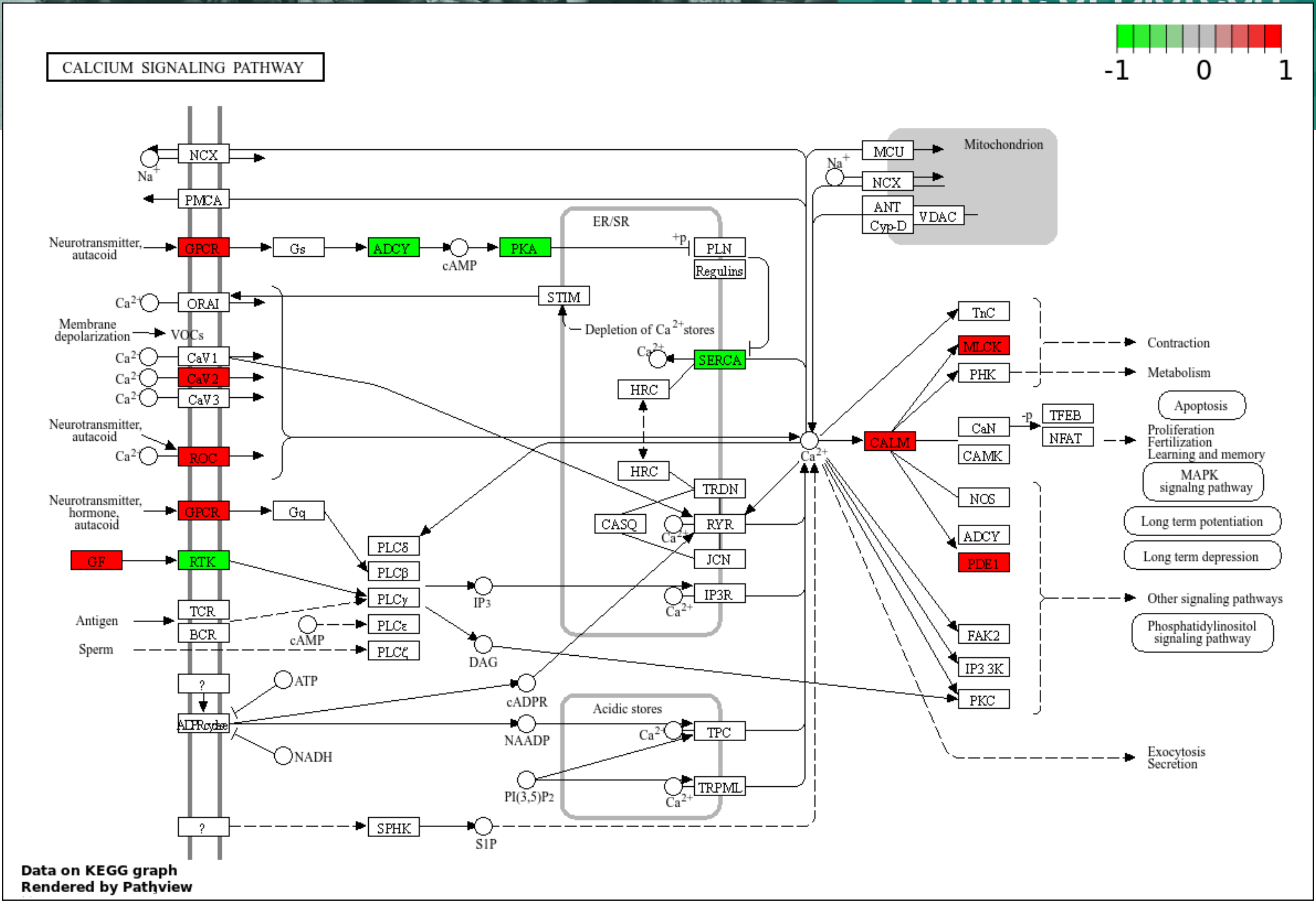


and RNA-processing categories highlights the unique molecular framework of the Basal subtype.

- **KEGG Pathway Visualizations (Pathview):**



- **Figure 6: Vitamin Digestion and Absorption (04977)**: The Pathview map indicates a concentrated upregulation (Red) in the **FOLH1** node. This suggests the Basal subtype has enhanced its capacity to capture extracellular folate.



Data on KEGG graph
Rendered by Pathview

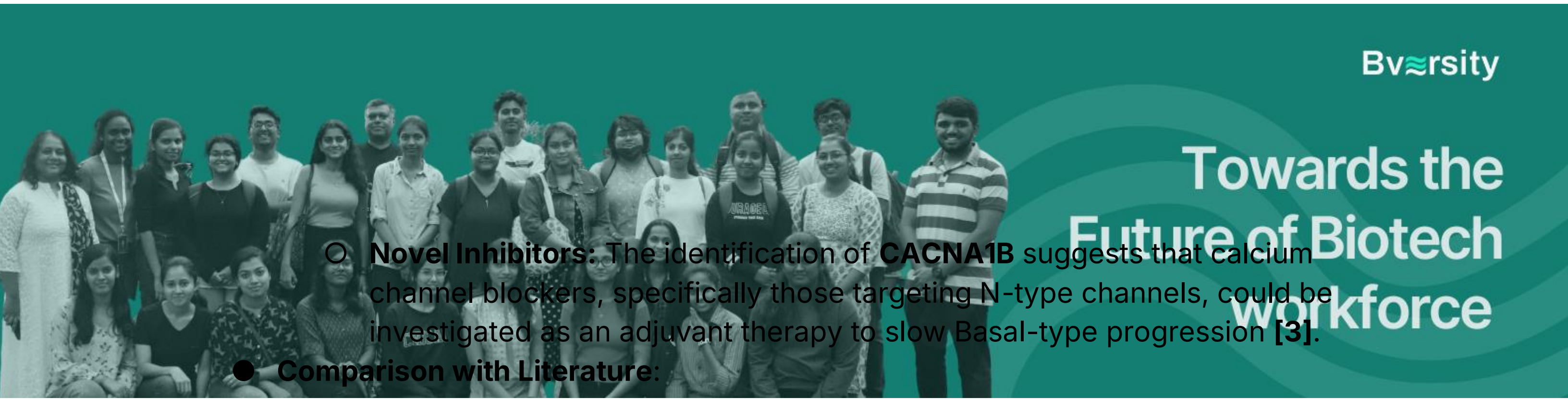
- **Calcium Signaling Pathway (04020):** Mapping reveals critical upregulation in the **Cav2 (CACNA1B)** and **CALM** nodes. This shift signifies a hijacked signaling mechanism used by the cancer cell to drive motility and survival signals.

6. Discussion

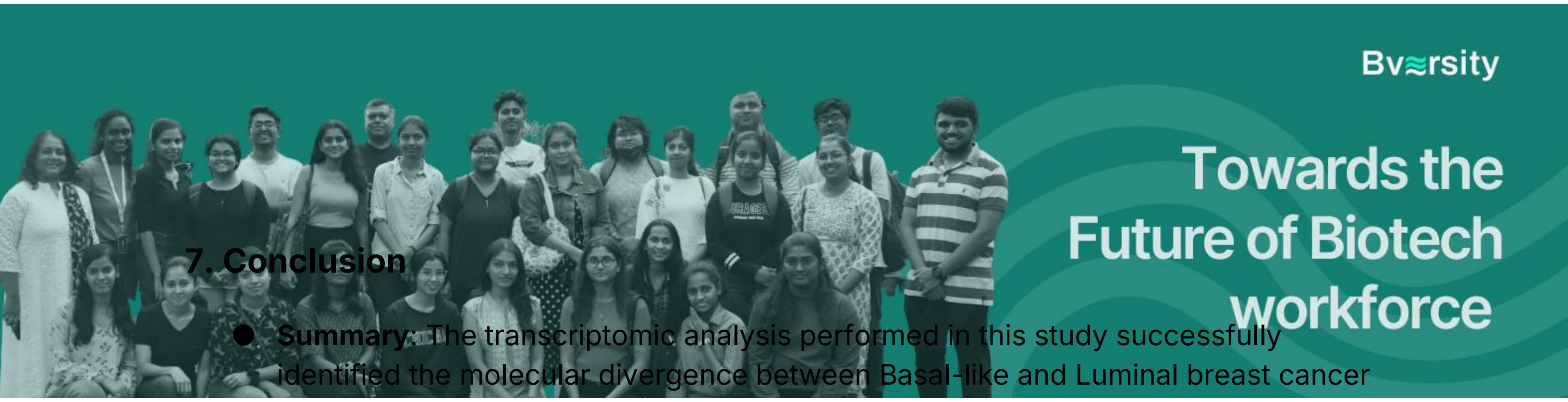
- **Interpretation**:
  - The transcriptomic profile of the Basal-like samples reveals a highly aggressive phenotype compared to the Luminal subtype. The massive upregulation of **KRT16** (a basal cytokeratin) and **PTCHD1** confirms the structural identity and signature marker of these cells **[5]**.This shift is further supported by the GO enrichment in intermediate filament-based processes, indicating a fundamental change in the cellular cytoskeleton. Conversely, the significant downregulation of **TFF1** and **ERBB4** highlights the loss of estrogen-responsive regulatory mechanisms that typically characterize more differentiated Luminal tumors **[6]**.

    The most critical finding is the dual-layer adaptation in the Basal subtype:

  - **1. Metabolic Advantage:** The Pathview analysis of the Vitamin Digestion and Absorption pathway shows a specific "hit" at **FOLH1**. This indicates that Basal cells are transcriptionally programmed to increase folate uptake, a necessary precursor for rapid DNA synthesis and cell division **[2, 4]**.
  - **2. Signaling Advantage:** The Calcium Signaling Pathview shows a clear activation of the **Cav2 (CACNA1B)** and **CALM** (Calmodulin) axis. In a cancerous context, this upregulation facilitates a "second messenger" environment that promotes cell motility and survival, explaining the traditionally higher metastatic potential of Basal-like tumors **[3]**.
- **Applications**:
  - The findings from this project have direct industrial and clinical relevance:
  - **Precision Diagnostics:** The inverse expression ratio of **KRT16/TFF1** serves as a high-confidence transcriptomic signature for subtype classification.
  - **Targeted Therapy: FOLH1** (also known as **PSMA**) is a proven target for radioligand therapy. This project suggests that Basal-like breast cancer patients could potentially be screened for **FOLH1**-targeted imaging and treatment **[2]**.



  - **Novel Inhibitors:** The identification of **CACNA1B** suggests that calcium channel blockers, specifically those targeting N-type channels, could be investigated as an adjuvant therapy to slow Basal-type progression **[3]**.
- **Comparison with Literature**:
  - The results align with established oncological standards:
  - Biomarkers: The identification of **KRT16**, **PTCHD1**, **TFF1** and **ERBB4** as the topmost DEGs is consistent with the "gold standard" molecular definitions of Basal and Luminal breast cancers **[1].** Specifically, **PTCHD1** upregulation in the Basal group aligns with its known role in the Hedgehog pathway, a driver of cancer stemness **[7]**.
  - Metabolism: Recent literature identifies metabolic reprogramming (such as the folate pathway) as a hallmark of Triple-Negative/Basal-like cancers, which our **goseq** and **Pathview** results independently confirmed **[4]**.
- **Limitations**:
  - Sample Size: While the 6 samples provided a statistically significant signal ($P_{adj} < 0.05$), a larger cohort would be required to ensure these targets **(FOLH1/CACNA1B)** remain consistent across diverse patient populations.
  - In Vitro Validation: This study is purely computational (silico). While the transcriptomic evidence is strong, the protein-level activity of **FOLH1** and **CACNA1B** in these specific samples would need to be validated via immunohistochemistry (IHC) or Western Blotting.

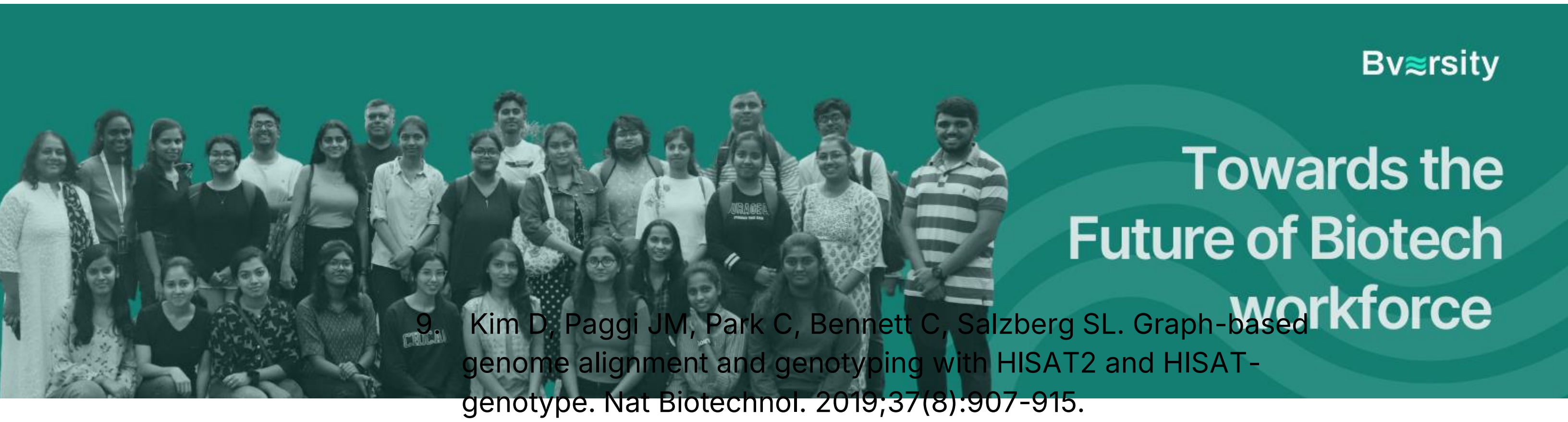


7. Conclusion

- **Summary**: The transcriptomic analysis performed in this study successfully identified the molecular divergence between Basal-like and Luminal breast cancer subtypes. By processing over 28,000 genes through a robust RNA-Seq pipeline, 972 significant DEGs were isolated, highlighting a profound shift in the cellular programming of aggressive subtypes.
- The identification of **KRT16** and **TFF1** validated the accuracy of the pipeline, while the functional enrichment analysis provided new insights into the "Basal-like" survival strategy. The project concludes that:
    - Metabolic Reprogramming: Basal-like cells significantly upregulate **FOLH1**, suggesting a dependency on extracellular folate for rapid proliferation.
    - Signaling Hijacking: The upregulation of the Calcium Signaling hub (**CACNA1B** and **CALM**) indicates a specialized mechanism for cellular motility and invasion.
    - Clinical Impact: **FOLH1** and **CACNA1B** represent high-potential therapeutic targets. Specifically, **FOLH1** serves as a candidate for **PSMA**-based radioligand therapy, while **CACNA1B** offers a target for N-type calcium channel inhibitors.
- **Future Directions**:
    - Pathway Cross-talk: A Protein-Protein Interaction (PPI) network is needed to determine if a single regulatory "hub" controls both the metabolic (folate) and signaling (calcium) changes. Finding this link could reveal a way to disable both survival mechanisms simultaneously.
    - Cellular Resolution: Since this project used Bulk RNA-Seq, the results represent an average of the entire tumor. Single-cell RNA-seq (scRNA-seq) would be the next logical step to confirm these genes are expressed by the cancer cells themselves and not by the surrounding healthy tissue or immune cells.
    - Laboratory Validation: These in silico results must be verified in a "wet lab." Future work should focus on using immunohistochemistry (IHC) to check protein levels on cell surfaces and conducting functional assays to see if blocking these targets actually stops cancer cell growth and migration.

8. References

- **Data Sources**
    - Kumar, R., & Horvath, A. (2013). mRNA-sequencing of breast cancer subtypes and normal tissue. NCBI Gene Expression Omnibus. GSE52194.
    - Basal Samples(1, 2, 3 respectively): SRR1027171, SRR1027172, SRR1027173.
    - Luminal Samples(1, 2, 3 respectively): SRR1027177, SRR1027178, SRR1027179
- **Literature & Tool References**
    1. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-752.
    2. Syed S, Fetrow LM, Alghamdi M, Kase AM, Geyer SM, Hartley C, et al. FOLH1 expression in triple-negative breast cancer: A TCGA analysis. Sci Rep. 2023;13(1):21743.
    3. So CL, Saunus JM, Roberts-Thomson SJ, Monteith GR. Calcium signaling in breast cancer progression and metastasis. Semin Cell Dev Biol. 2019;94:52-63.
    4. Munkácsy G, Krizsán S, Sztupinszki Z, Pipek O, Kozma K, Krzystanek M, et al. Metabolic Reprogramming in Triple-Negative Breast Cancer. Int J Mol Sci. 2023;24(9):8013.
    5. Joosse A, et al. Emerging Insights into Keratin 16 Expression during Metastatic Progression of Breast Cancer. British Journal of Cancer. 2021.
    6. Wang L, et al. Trefoil factor 1 (TFF1) is a potential prognostic biomarker with functional significance in breast cancers. Cell Death Dis. 2020;11(10):1-13.
    7. Bhateja P, et al. The Hedgehog Signaling Pathway: A Viable Target in Breast Cancer? Cancers. 2019;11(8):1126.
    8. The Galaxy Community. The Galaxy platform for accessible, reproducible and transparent biomedical research: 2022 update. Nucleic Acids Res. 2022;50(W1):W345-W351.



9. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907-915.
10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
11. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias (goseq). Genome Biol. 2010;11(2):R14.
12. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830-1831.

---

9. Appendix

- **Supplementary Materials**: Complete DEG Dataset: Full_DEG_List.
- The full list of 972 differentially expressed genes is provided as a supplementary electronic file.
- **Code/Workflow Links**: The complete bioinformatics pipeline, including all tool parameters and intermediate files, is publicly accessible via the Galaxy Project:

  The Analysis Workflow (Galaxy History)

---

10. Industry Impact Statement

- This workflow demonstrates a scalable pipeline for identifying subtype-specific biomarkers. By identifying **FOLH1** as a target in Basal-like samples, this study provides a framework for pharmaceutical companies to prioritize patients for **PSMA**-targeted clinical trials, potentially reducing drug development timelines and improving precision oncology outcomes.