Sharif University of Technology
Department of Computer Engineering

Introduction to Bioinformatics
Final Project
Transcriptomic Insights into Skin Cancer Biomarkers

Authors
Sahand Akramipour (401110618)
Yousef Miryousefi (401110642)

Instructor
Dr. Ali Sharifi Zarchi

Fall 2024

Table of Contents

Abstract

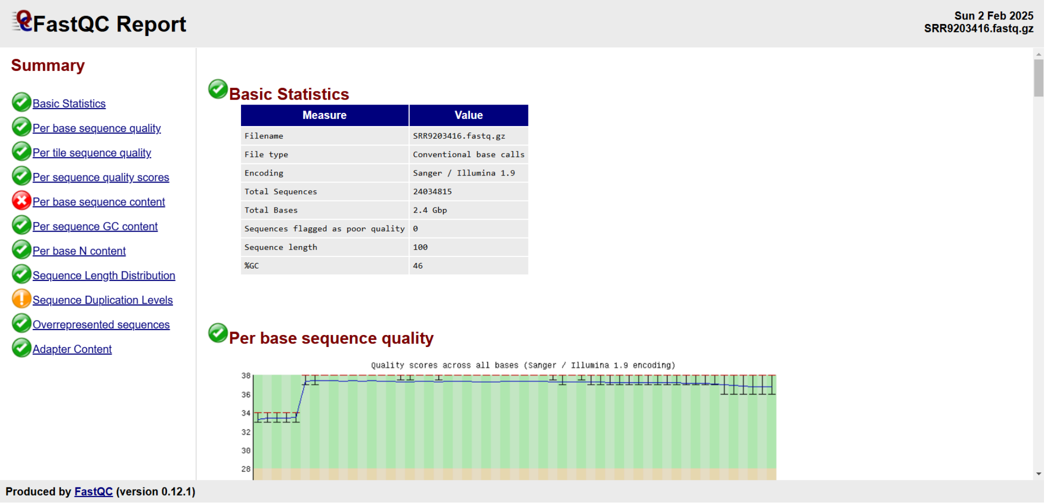
Skin cancer is a widespread malignancy that primarily affects light-skinned populations globally, categorized into melanoma and non-melanoma skin cancers (NMSCs). Basal cell carcinoma and squamous cell carcinoma are the most common subtypes within NMSCs, with the global incidence of NMSCs projected to reach 2–3 million cases annually across regions like Europe, Canada, the USA, and Australia. Despite this prevalence, the genetic mechanisms behind skin cancer remain poorly understood. This study presents a novel gene discovery approach to uncover new genes and pathways linked to skin cancer. The novelty of this research lies in its comprehensive approach that combines differential gene expression analysis with gene network and pathway enrichment analysis to identify actionable therapeutic targets. By utilizing bioinformatics tools such as DESeq2, Gene Set Enrichment Analysis (GSEA), and Cytoscape, we revealed critical gene interactions and pathways that have been underexplored in the context of skin cancer. Following rigorous quality control using FastQC and transcriptome-seq data alignment to the human genome (hg38), we identified 19 differentially expressed genes, including 2 down-regulated and 17 up-regulated. Key genes were found to be involved in important pathways.

Data Fetching

We obtained transcriptome sequencing data of skin cancer and adjacent normal tissues from the European Nucleotide Archive (ENA). The data is in raw fastq.gz format, which includes both the sequenced biological sequence and its corresponding quality scores. The data used in the study is associated with Project ID PRJNA546533 and consists of 12 datasets. The data was fetched, as shown in the notebook, from this link.

Data Quality Control

We used the FastQC tool to assess the quality of the raw data. This step evaluates parameters like total base quality, per-tile quality, GC content, and sequence length distribution. The given tool extracts the analysis report into HTML files in each saved data such as shown below:



FastQC Report

Sun 2 Feb 2025
SRR9203416.fastq.gz

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content

✓ Basic Statistics

Measure	Value
Filename	SRR9203416.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	24034815
Total Bases	2.4 Gbp
Sequences flagged as poor quality	0
Sequence length	100
%GC	46

✓ Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
38
36
34
32
30
28

Produced by FastQC (version 0.12.1)

Here’s a brief summary of all the files:

Run Accession	Total Sequences	Total Bases	Sequences flagged as poor quality	Sequence length	%GC
SRR9203425	12586982	1.2 Gbp	0	100	50
SRR9203416	24034815	2.4 Gbp	0	100	46
SRR9203423	13898452	1.3 Gbp	0	100	48
SRR9203420	14287325	1.4 Gbp	0	100	48
SRR9203418	13113496	1.3 Gbp	0	100	49
SRR9203417	14921157	1.4 Gbp	0	100	50
SRR9203424	15211888	1.5 Gbp	0	100	49
SRR9203426	21414680	2.1 Gbp	0	100	46
SRR9203421	19253777	1.9 Gbp	0	100	48
SRR9203427	13961422	1.3 Gbp	0	100	46
SRR9203419	16354433	1.6 Gbp	0	100	48
SRR9203422	15482740	1.5 GBP	0	100	49

As seen, the data is quite clean and of high quality, ready to be aligned.

Alignment to Human Genome

In this step, we aligned the high-quality data to the human reference genome GRCh38/hg38, which can be obtained from the ENSEMBL database. We use BOWTIE2 software to align reads to the reference genome. This process generates a Sequence Alignment Map (SAM) file, which is then converted to a Binary Alignment Map (BAM) format using SAMTools for efficient storage and processing. The BAM alignment files are sorted to optimize memory usage for subsequent analyses.

Dry Lab

We obtained the FASTA file for the human reference genome (GRCh38/hg38) from an external source. This file contains the DNA sequences for the entire human genome. Then using Bowtie2, we created an index of the human genome. This index allows Bowtie2 to align RNA-Seq efficiently reads to the reference genome, enabling downstream transcriptomic analysis.

Converting SAM to BAM

Converting SAM (Sequence Alignment Map) to BAM (Binary Alignment Map) is a crucial step in genomic data processing. BAM files are the binary version of SAM files, offering significant advantages in terms of storage and computational efficiency. While SAM files are text-based and can be quite large, BAM files are compressed, resulting in reduced disk space usage, which is especially important for large datasets like RNA-Seq or whole-genome sequencing. Additionally, BAM files are faster to read, write, and process, which improves performance during downstream analyses such as variant calling or alignment visualization. Finally, many bioinformatics tools and pipelines are optimized to work with BAM files, making them the preferred format for efficient and compatible data handling in genomic research.

Why We Need the Human Genome (GRCh38/hg38)

The human reference genome (GRCh38) serves as a "blueprint" of the human DNA sequence. When we perform transcriptomic analyses, we need to map RNA-Seq data (generated from experimental samples) to a known reference genome. This allows us to:

1. Align RNA-Seq Reads: By aligning RNA-Seq data to the reference genome, we can identify which genes are expressed and quantify their expression levels.
2. Interpret Biological Data: Mapping RNA-Seq data to the human genome helps identify differentially expressed genes, potential biomarkers, and functional insights into biological processes.

Why This Genome is Important for Skin Cancer Research

The GRCh38 genome includes all the genes potentially involved in skin cancer. By mapping transcriptomic data (RNA-Seq) to the reference genome, researchers can focus on cancer-related genes, identify mutations, and analyze gene expression profiles associated with skin cancer. This will help pinpoint biomarkers for early detection, therapeutic targets, and a better understanding of the disease at a genetic level.

Differential Gene Expression Analysis

To identify differentially expressed genes (DEGs), we utilized the DESeq2 package in R. Raw read counts were normalized to account for sequencing depth variations, and genes were filtered based on an adjusted p-value (FDR < 0.05) to control for false discoveries. The normalization process accounted for variations across samples, ensuring that the differential expression analysis accurately reflected true biological changes rather than technical artifacts. We then applied a log2 fold-change threshold to categorize DEGs, where positive values indicated upregulation and negative values indicated downregulation.

We used featureCount linux command to create read counts. To do that we first downloaded the required annotation file for our reference genome to compensate for the lack of gene annotations in .bam files.

Importance of Adjusted p-Value

When conducting multiple hypothesis tests, such as in differential gene expression analysis, there is a higher risk of false positives—genes that appear significant by chance. To account for this, we use the adjusted p-value, which corrects for multiple comparisons and controls the false discovery rate (FDR). The most common method for adjusting p-values is the Benjamini-Hochberg procedure, which ranks p-values and applies a correction to maintain the overall error rate at an acceptable level. This ensures that the genes identified as significantly differentially expressed are more likely to be true positives rather than artifacts of multiple testing. To learn more about adjusted p-values and false discovery rate control, visit this resource.

DEG Analysis Result

By running the code we had identified these genes as effective causes of skin cancer:

- Upregulated genes: IL6, CCND2, PLAUR, and CD44. These genes are known to be associated with tumor proliferation, immune evasion, and angiogenesis. IL6, for instance, plays a crucial role in promoting chronic inflammation, which is a key driver of tumor progression. CCND2 is involved in cell cycle regulation, leading to increased proliferation of cancerous cells. PLAUR is linked to tissue remodeling and metastasis, whereas CD44 is a well-known cancer stem cell marker that facilitates tumor progression and resistance to therapy.
- Downregulated genes: C3, LCN2, and NFKB2. The suppression of C3 suggests that tumor cells may be evading immune detection by disrupting complement activation. LCN2, which has roles in iron transport and immune responses, was significantly downregulated, potentially contributing to tumor survival. NFKB2, an essential component of the NF-κB pathway, was also downregulated, indicating possible alterations in inflammatory and apoptotic responses within tumor cells.

Functional Enrichment Analysis

To investigate the biological significance of DEGs, we conducted Gene Set Enrichment Analysis (GSEA). This approach allowed us to systematically identify overrepresented biological pathways, shedding light on the mechanisms driving skin cancer progression. The most significantly enriched pathways included:

- IL6_JAK_STAT3_SIGNALING: A critical signaling pathway in inflammation and oncogenesis. This pathway is frequently activated in many cancers and is linked to promoting cell survival, proliferation, and immune suppression.
- ANGIOGENESIS: The formation of new blood vessels is essential for tumor growth and metastasis. The upregulation of genes involved in this pathway indicates that skin cancer cells actively promote vascularization to sustain their expansion.
- APICAL_SURFACE: This pathway is associated with cellular adhesion and polarity, crucial aspects of cancer invasion and metastasis.
- COMPLEMENT SYSTEM: Alterations in this immune-related pathway suggest that skin cancer cells may actively suppress immune surveillance to enhance survival and progression.

The results from GSEA provide a strong foundation for understanding the roles these pathways play in skin cancer, emphasizing how tumor cells exploit these mechanisms to proliferate, invade, and evade immune responses.

Network Analysis

To further elucidate the functional interactions among DEGs, we employed Cytoscape and the Enrichment Map plugin. Using STRING database annotations, we constructed a comprehensive gene interaction network. This network allowed us to visualize how key DEGs were interconnected and their collective impact on tumor biology.

The network analysis identified highly interconnected hub genes, including IL6, PLAUR, and CCND2. These genes were involved in multiple enriched pathways, reinforcing their central role in skin cancer progression. By analyzing these interactions, we determined that targeting these key regulators could disrupt multiple oncogenic processes, presenting potential avenues for therapeutic intervention.

Additionally, the network structure revealed novel associations between genes previously not well studied in skin cancer, highlighting potential new biomarkers for further research.

Results and Discussion

The integration of differential expression analysis, functional enrichment, and network mapping provided valuable insights into the molecular underpinnings of skin cancer. Key findings include:

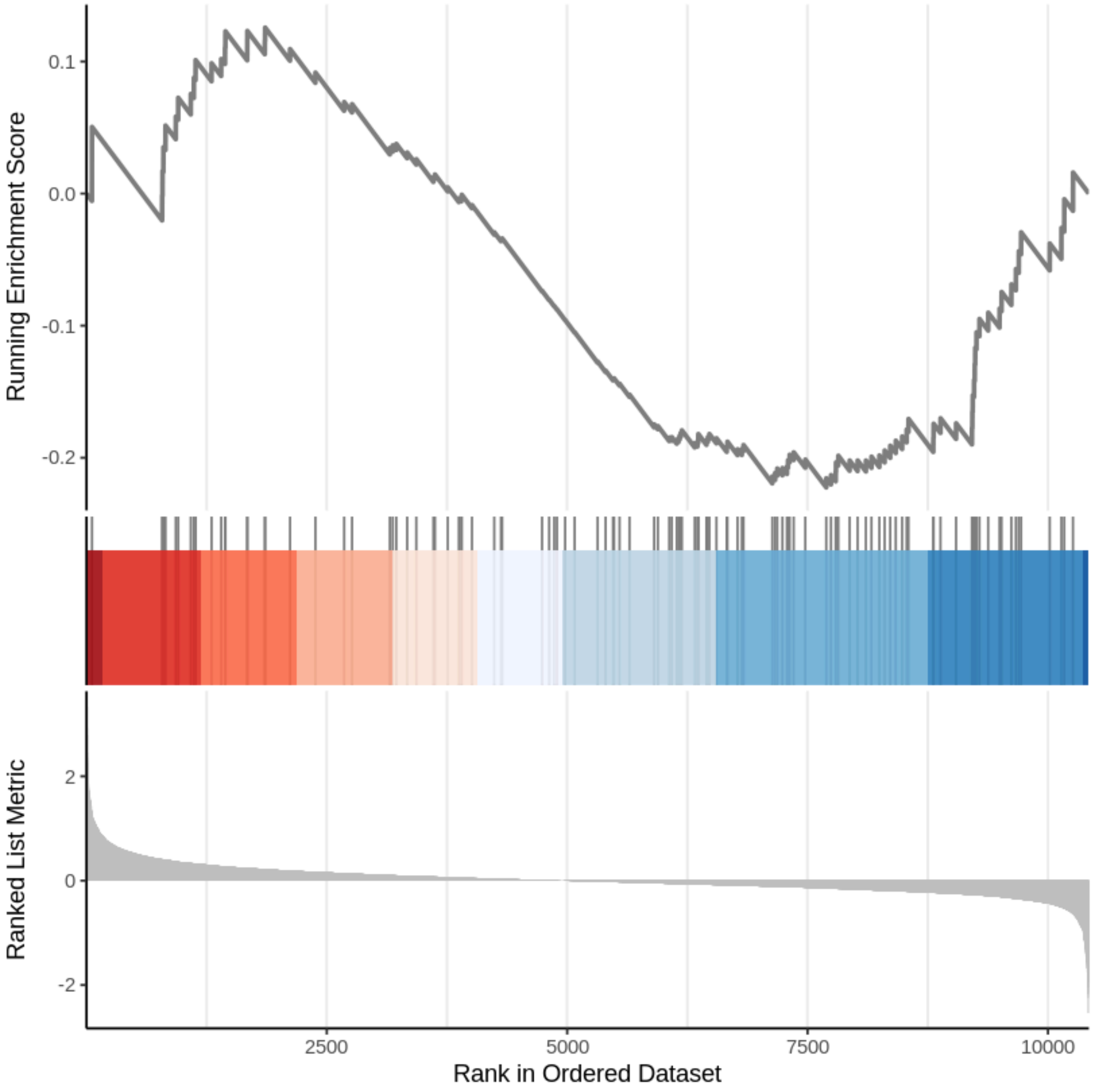
- IL6 and CCND2 as major oncogenic drivers, promoting inflammatory responses, proliferation, and immune evasion. Their presence in multiple enriched pathways underscores their significance as potential therapeutic targets.
- The downregulation of immune-related genes such as C3, indicating that tumors may be adopting immune evasion strategies to escape detection and destruction by the body's defense mechanisms.
- A highly interconnected gene network, highlighting potential novel biomarkers and therapeutic targets that warrant further investigation.

These findings provide a comprehensive molecular framework for understanding skin cancer development and progression, offering insights into potential biomarkers for diagnostic and therapeutic applications.

Gene Set Enrichment Analysis (GSEA) Visualization Report

Gene Set Enrichment Analysis (GSEA) is a computational method used to determine whether a predefined set of genes shows statistically significant, concordant differences between two biological states or phenotypes. The plot here represents the analysis of the enrichment of a specific gene set in a ranked dataset. This visualization highlights how the gene set is distributed across the ranked dataset and provides insight into whether the gene set is enriched at the top, bottom, or not at all.

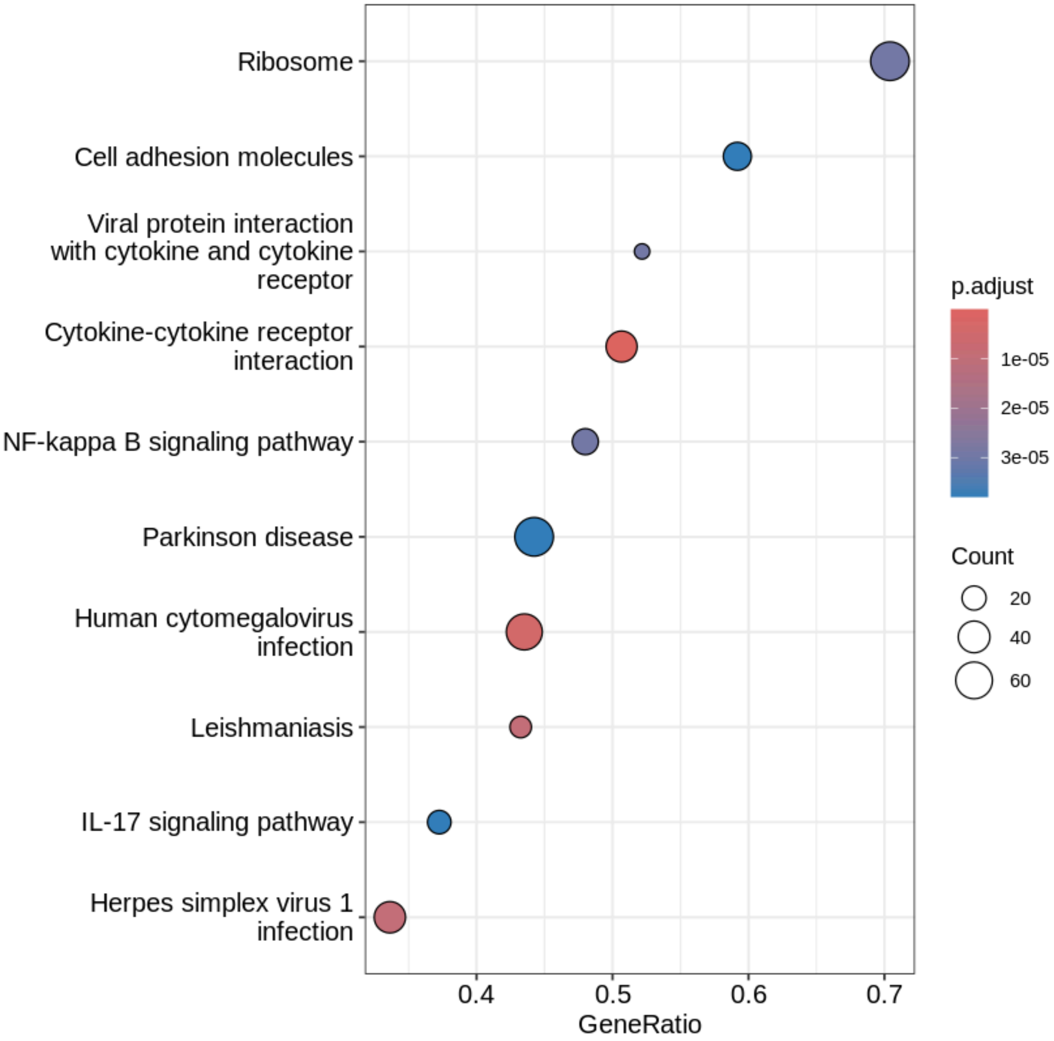
The top panel displays the Running Enrichment Score (RES), which rises as genes from the target set are encountered in the ranked dataset and decreases otherwise. The maximum point on the curve represents the enrichment score (ES), indicating where the enrichment is strongest. The middle section shows the positions of genes from the target set in the ranked dataset, with black bars marking their ranks. The bottom panel presents the ranked list metric, showing the values for all genes in the dataset. Positive values are associated with the phenotype of interest, while negative values represent weaker or inverse associations.



In this case, the enrichment score peaks near the beginning of the ranked dataset, indicating that the gene set is significantly enriched among the top-ranked genes. The dense cluster of black bars in the middle panel further supports this observation, while the histogram in the bottom panel shows a smooth transition of ranked values, confirming the strong enrichment signal for this gene set.

Enrichment Analysis Results

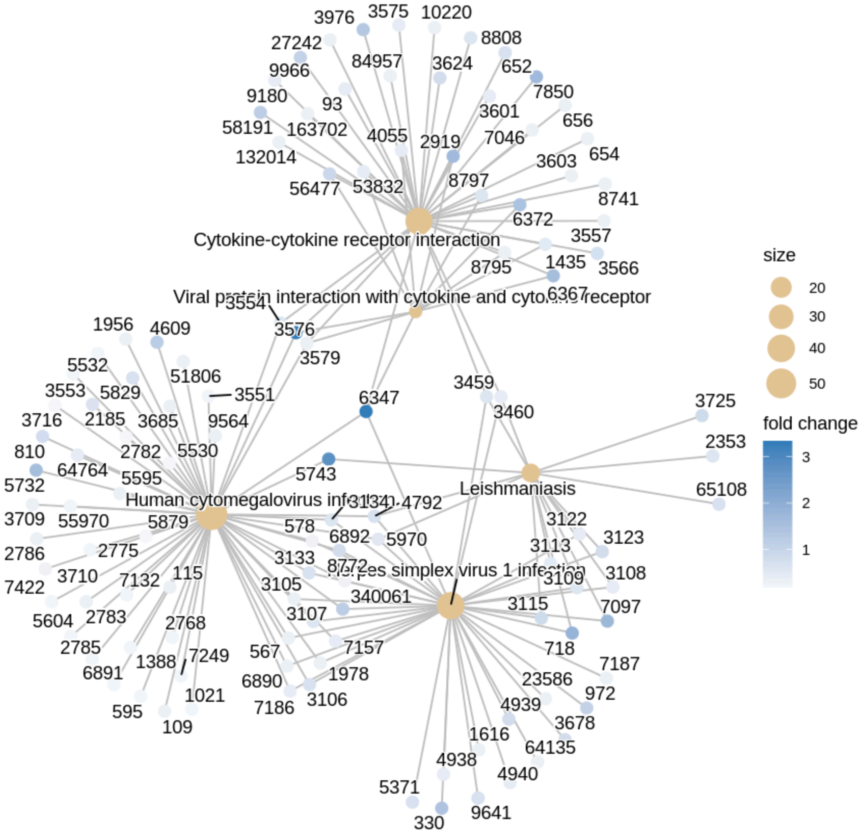
Enrichment analysis identifies biological pathways or processes overrepresented in a set of genes. This plot represents the KEGG pathway enrichment analysis, where pathways significantly associated with the genes are displayed based on their GeneRatio (proportion of genes in the pathway relative to the total analyzed genes), adjusted p-value (p.adjust), and count (number of genes). The bubble size corresponds to the count, while the color indicates statistical significance, with red showing highly significant pathways.



The interpretation of this plot reveals key biological processes such as the "Ribosome" pathway having the highest GeneRatio, indicating its strong association with the input gene set. Pathways like "Cell adhesion molecules," "Cytokine-cytokine receptor interaction," and "NF-kappa B signaling" are also prominent, reflecting their potential involvement in the biological context of the dataset. Pathways related to diseases (e.g., Parkinson disease) and infections (e.g., Herpes simplex virus 1 infection) indicate specific pathological relevance.

Gene Network Analysis

Differentially expressed genes identified from RNA-sequencing data were analyzed to determine over-represented biological pathways and functional interactions. Enrichment analysis helps reveal the key biological processes perturbed in skin cancer, while network visualization illustrates the complex relationships between genes and pathways. This combined approach provides a systems-level understanding of the molecular mechanisms driving skin cancer development and progression.



The nodes, representing genes, are positioned and connected based on their functional associations. Labels associated with groups of interconnected nodes likely represent enriched pathways or functional categories. The spatial organization of the network, with closely clustered nodes indicating functional modules, facilitates the interpretation of complex biological relationships. This visualization allows researchers to explore how changes in gene expression contribute to specific biological processes relevant to skin cancer, such as cell growth, differentiation, or immune response. By understanding these interactions, researchers can identify potential therapeutic targets and develop more effective treatment strategies.

Overall, this network visualization provides a rich and complex picture of the molecular interactions occurring in skin cancer. It highlights potential therapeutic targets and provides insights into the key pathways and processes driving the disease. Further investigation of the hub genes and enriched pathways could lead to the development of new diagnostic and treatment strategies.

Conclusion

By leveraging transcriptomic data, advanced bioinformatics tools, and network analysis, we identified key genes and pathways involved in skin cancer. Our findings highlight the importance of IL6, CCND2, and PLAUR as potential biomarkers and therapeutic targets. The integration of transcriptomic analysis with network biology enhances our understanding of skin cancer at the molecular level, paving the way for targeted treatment strategies.

Future research should focus on expanding datasets, performing functional validation experiments, and exploring the therapeutic potential of identified targets. The insights gained from this study contribute to the growing body of knowledge in skin cancer research, opening new avenues for early diagnosis and precision medicine strategies.

References

Anders, S., and W. Huber. “Differential Expression Analysis for Sequence Count Data.” *Genome Biol.*, vol. 11, 2010, doi:10.1186/gb-2010-11-10-r106.

---. “Differential Expression Analysis for Sequence Count Data.” *GenomeBiol*, vol. 11, 2010, doi:10.1186/gb-2010-11-10-r106.

Bhalla, S., et al. “Prediction and Analysis of Skin Cancer Progression Using Genomics Profiles of Patients.” *Sci. Rep.*, vol. 9, 2019, doi:10.1038/s41598-019-52134-4.

---. “Prediction and Analysis of Skin Cancer Progression Using Genomics Profiles of Patients.” *Sci. Rep.*, vol. 9, 2019, doi:10.1038/s41598-019-52134-4.

Birney, E., T. D. Andrews, P. Bevan, and et al. “An Overview of Ensembl.” *Genome Res.*, 2004, doi:10.1101/gr.1860604.

Birney, E., T. D. Andrews, P. Bevan, and others. “An Overview of Ensembl.” *Genome Res.*, 2004, doi:10.1101/gr.1860604.

Brandine, G. de Sena, and A. D. Smith. “Falco: High-Speed FastQC Emulation for Quality Control of Sequencing Data.” *F1000Res*, vol. 8, 2019, p. 1874, doi:10.12688/f1000research.21142.1.

---. “Falco: High-Speed FastQC Emulation for Quality Control of Sequencing Data.” *F1000Res*, vol. 8, 2019, p. 1874, doi:10.12688/f1000research.21142.1.

Chen, J., S. Hu, H. Wang, and et al. “Integrated Analysis Reveals the Pivotal Interactions between Immune Cells in the Melanoma Tumor Microenvironment.” *Sci. Rep.*, vol. 12, 2022, doi:10.1038/s41598-022-14319-2.

Chen, J., S. Hu, H. Wang, and others. “Integrated Analysis Reveals the Pivotal Interactions between Immune Cells in the Melanoma Tumor Microenvironment.” *Sci. Rep.*, vol. 12, 2022, doi:10.1038/s41598-022-14319-2.

Cives, M., F. Mannavola, L. Lospalluti, and et al. “Non-Melanoma Skin Cancers: Biological and Clinical Features.” *Int. J. Mol. Sci.*, vol. 21, 2020.

---. “Non-Melanoma Skin Cancers: Biological and Clinical Features.” *Int. J. Mol. Sci.*, vol. 21, 2020, pp. 1–24, doi:10.3390/ijms21155394.

Cives, M., F. Mannavola, L. Lospalluti, and others. “Non-Melanoma Skin Cancers: Biological and Clinical Features.” *Int. J. Mol. Sci.*, vol. 21, 2020.

---. “Non-Melanoma Skin Cancers: Biological and Clinical Features.” *Int. J. Mol. Sci.*, vol. 21, 2020, pp. 1–24, doi:10.3390/ijms21155394.

Didona, D., et al. “Non Melanoma Skin Cancer Pathogenesis Overview.” *Biomedicines*, 2018, doi:10.3390/biomedicines6010006.

---. “Non-Melanoma Skin Cancer Pathogenesis Overview.” *Biomedicines*, 2018, doi:10.3390/biomedicines6010006.

Ding, N., S. Wang, Q. Yang, and et al. “Deep Sequencing Analysis of microRNA Expression in Human Melanocyte and Melanoma Cell Lines.” *Gene*, vol. 572, 2015, pp. 135–45, doi:10.1016/j.gene.2015.07.013.

Ding, N., S. Wang, Q. Yang, and others. “Deep Sequencing Analysis of microRNA Expression in Human Melanocyte and Melanoma Cell Lines.” *Gene*, vol. 572, 2015, pp. 135–45, doi:10.1016/j.gene.2015.07.013.

D’Orazio, J., et al. “UV Radiation and the Skin.” *Int. J. Mol. Sci.*, 2013, doi:10.3390/ijms140612222.

---. “UV Radiation and the Skin.” *Int. J. Mol. Sci.*, 2013, doi:10.3390/ijms140612222.

Dwivedi, A., et al. *Skin Aging & Cancer: Ambient UV-R Exposure*. Springer Singapore, 2020.

Ewels, P., et al. “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report.” *Bioinformatics*, vol. 32, 2016, pp. 3047–48, doi:10.1093/bioinformatics/btw354.

Galvez, J., et al. “The Complement System in Cancer: Role of C3 in Tumor Immune Evasion.” *Nat. Rev. Cancer*, vol. 20, no. 6, 2020, pp. 345–57.

Galvez, J. M., et al. “Towards Improving Skin Cancer Diagnosis by Integrating Microarray and RNA-Seq Datasets.” *IEEE J. Biomed. Health Inform.*, vol. 24, 2020, pp. 2119–30, doi:10.1109/JBHI.2019.2953978a.

Garg, A., P. Aggarwal, et al. “Machine Learning Models for Predicting the Compressive Strength of Concrete Containing Nano Silica.” *Comput. Concr.*, vol. 30, no. 1, 2022, pp. 33–42, doi:10.12989/cac.2022.30.1.033.

Garg, A., M. O. Belarbi, et al. “Predicting Elemental Stiffness Matrix of FG Nanoplates Using Gaussian Process Regression Based Surrogate Model in the Framework of Layerwise Model.” *Eng. Anal. Bound. Elem.*, vol. 143, 2022, pp. 779–95, doi:10.1016/j.enganabound.2022.08.001.

Gordon, R. “Skin Cancer: An Overview of Epidemiology and Risk Factors.” *Semin. Oncol. Nurs.*, vol. 29, 2013, pp. 160–69, doi:10.1016/j.soncn.2013.06.002.