Integrative Analysis of Gene Expression Patterns in Gastric Cancer:

Abstract:

This study investigates differential gene expression in gastric and smooth muscle tissues to uncover molecular mechanisms relevant to gastric cancer. RNA-Seq data was processed using trimming, alignment, counting, and quality control, followed by differential expression analysis using DESeq2 and edgeR. PCA and heatmap visualizations confirmed clear tissue separation. Genes like CNN2 and MYLK-AS1 were upregulated, while tumor suppressors like GKN1 and TFF2 were downregulated in stomach tissues. Enrichment analyses via DAVID and STRING identified potential pathways and molecular interactions related to gastric cancer.

Introduction:

Gastric cancer remains a leading cause of cancer-related mortality worldwide. It arises from the epithelial lining of the stomach and progresses through a series of genetic and epigenetic alterations. Understanding transcriptomic changes in gastric cancer can reveal biomarkers for diagnosis and potential targets for therapy.

Calponin 2 (CNN2) has been linked to enhanced cell proliferation in gastric tumors (Hu et al., 2017). MYLK-AS1, a long non-coding RNA, contributes to tumorigenesis through epigenetic silencing of LATS2 via EZH2 (Luo & Xiang, 2021). GKN1, a gastric-specific tumor suppressor, is often silenced in tumors through epigenetic regulation and miRNA targeting (di Stadio et al., 2019). Additionally, ghrelin (GHRL), a peptide hormone involved in appetite regulation, exhibits lower expression in gastric cancer tissues but paradoxically associates with worse outcomes when elevated in precancerous states (Wang et al., 2023).

To elucidate gene expression dynamics, we performed RNA-Seq analysis on stomach and smooth muscle tissues, employing methods such as rlog transformation, PCA, and distance heatmap clustering. The findings aim to enhance our molecular understanding of gastric cancer.

Hypothesis:

We hypothesize that in-depth transcriptomic analysis of stomach and smooth muscle tissues using RNA-Seq and advanced bioinformatics will uncover key molecular signatures involved in gastric cancer. Specifically, we anticipate the upregulation of genes such as CNN2 and MYLK-AS1, supporting their role as oncogenic drivers, and the downregulation of tumor-suppressive genes like GKN1, potentially influenced by regulators such as miR-544a. These findings are expected to provide deeper insight into the gene regulatory networks driving gastric cancer progression and offer avenues for identifying prognostic markers or therapeutic targets.

Methods:

RNA-Seq data from stomach and smooth muscle samples underwent a series of preprocessing steps to ensure data quality and accurate quantification.

• Quality Control: FastQC (v0.11.9) was used to evaluate the quality of raw reads and detect potential issues such as low-quality base calls or adapter contamination.

• Trimming: Trim Galore (v0.6.6), built on Cutadapt, was employed to remove low-quality bases and sequencing adapters from raw reads.

• Mapping: Cleaned reads were aligned to the reference genome using HISAT2 (v2.2.2), a splice-aware aligner suitable for transcriptomic data. For organisms without a reference genome, de novo transcriptome assembly would be an alternative approach.

• File Conversion and Visualization: Aligned BAM files were converted from SAM using SAMtools (v1.10), and visualized with IGV (v2.8.0) to inspect mapping quality and gene coverage.

• Counting: Gene-level expression counts were generated using HTSeq, which quantifies reads per gene. Attention was paid to non-specific mapping and how multimapping reads were treated by the software.

• Normalization and Differential Expression: Using R (v4.3.1), we performed rlog transformation and statistical modeling with DESeq2 (v1.42.0) and edgeR (v4.0.1). Gene filtering and visualization were supported by Genefilter (v1.84.0), Pheatmap (v1.0.12), RColorBrewer (v1.1-3), and Statmod (v1.5.0).

• Data Visualization: PCA plots were generated to assess variance across samples, while heatmaps (both sample-wise and gene-wise) provided insights into clustering patterns and expression dynamics.

• Enrichment Analysis: Differentially expressed genes were further analyzed using DAVID for functional enrichment. Protein-protein interaction networks were visualized via STRING.

Human tissue samples were ethically sourced from the Uppsala Biobank under Swedish regulations, governed by the Uppsala Ethical Review Board (Refs: 2002-577, 2005-338, 2007-159 for protein, and 2011-473 for RNA). Samples were obtained through the Uppsala Biobank under Swedish ethical guidelines (Refs: 2002-577, 2005-338, 2007-159, 2011-473).

Results:

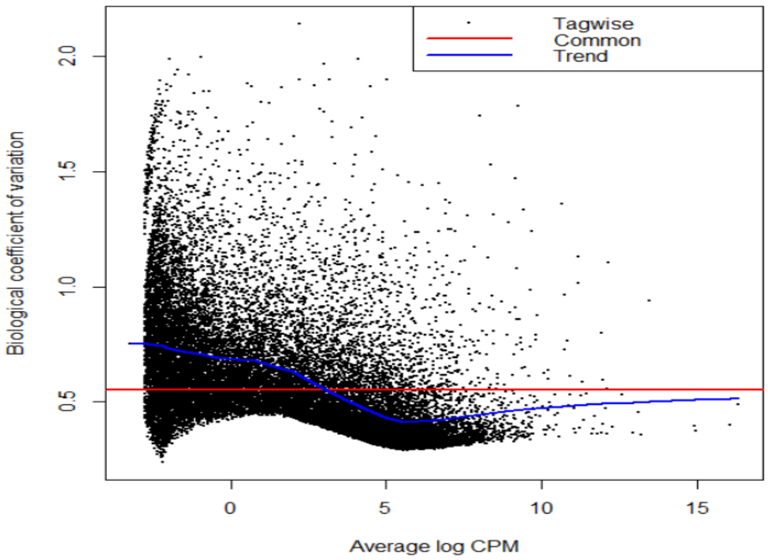


Figure 1: BCV Plot

The biological coefficient of variation (BCV) plot shows the relationship between expression level and variability across genes. The fitted trend line reflects how variance is adjusted based on expression magnitude, helping to highlight genes with true biological differences while reducing technical noise.

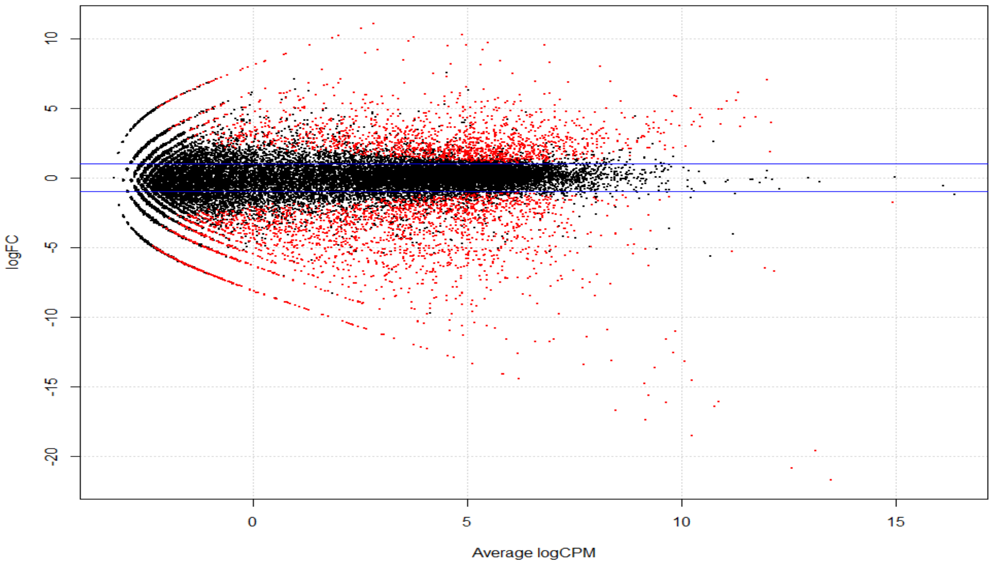


Figure 2: Smear plot

This plot visualized log-fold changes versus average expression. Genes that deviate significantly from the center horizontal line were differentially expressed, with red points indicating significance.

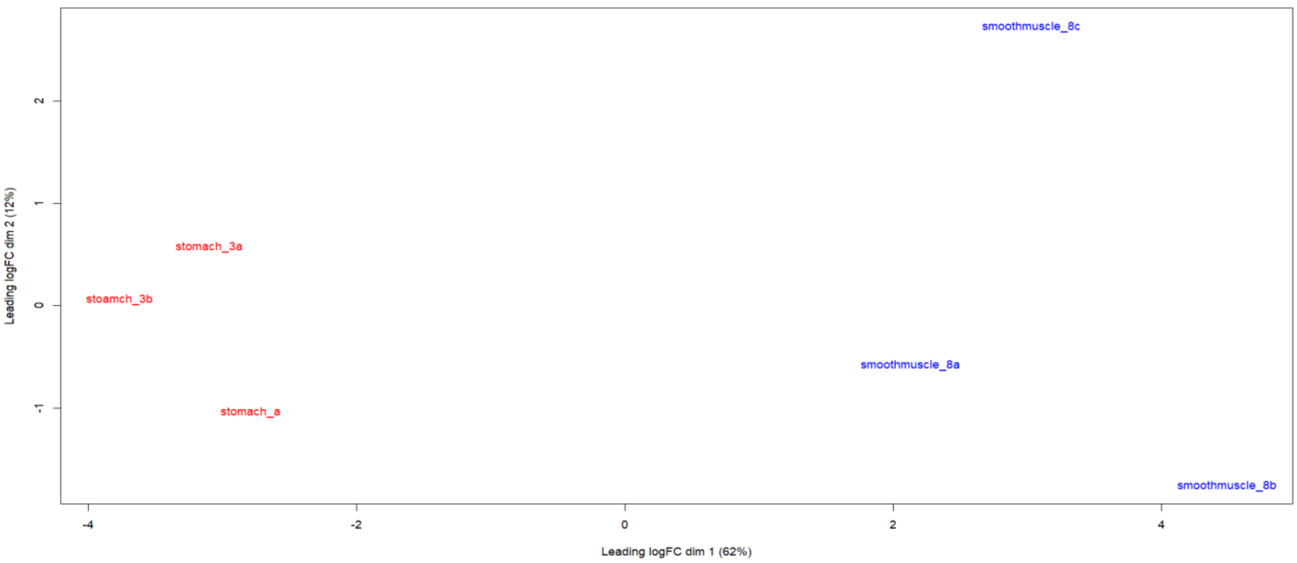


Fig 3: MDS plot

Multidimensional Scaling (MDS) was used to visualize the similarity between samples based on their gene expression profiles. In Figure 3, the two primary dimensions explain 62% and 5% of the total variance, respectively. This suggests that most of the variation can be attributed to differences along the first axis. The plot clearly separates 'stomach' and 'smooth muscle' samples, indicating that gene expression profiles within each tissue type are more like one another than between groups.

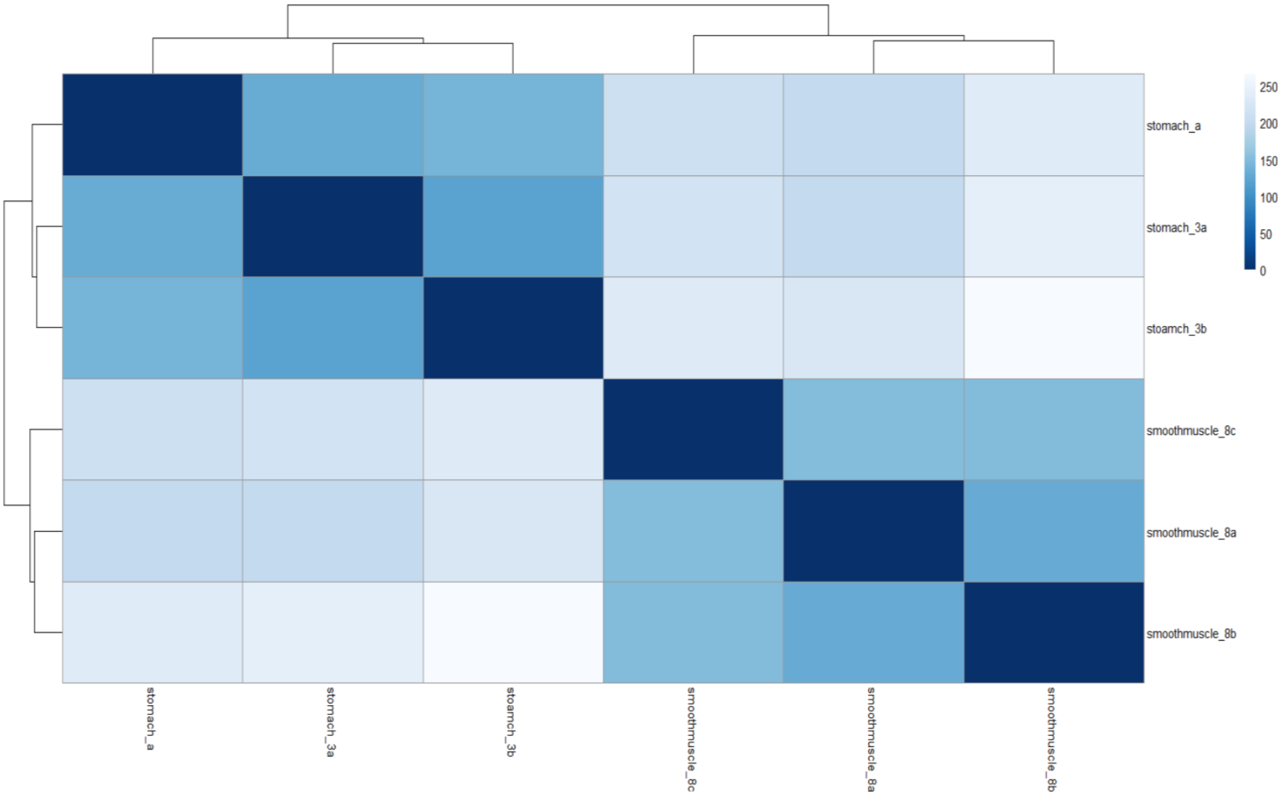


Figure 4: Distance Heatmap

This matrix visualized pairwise sample distances based on rlog-transformed counts. Darker blue indicates smaller distances (greater similarity), while lighter blue indicates larger distances. The darkest blue diagonal represents self-comparison (zero distance). The upper left quadrant reveals that stomach samples are more similar to each other, while the lower right quadrant shows a similar pattern for smooth muscle samples. This supports consistent clustering within each tissue type.

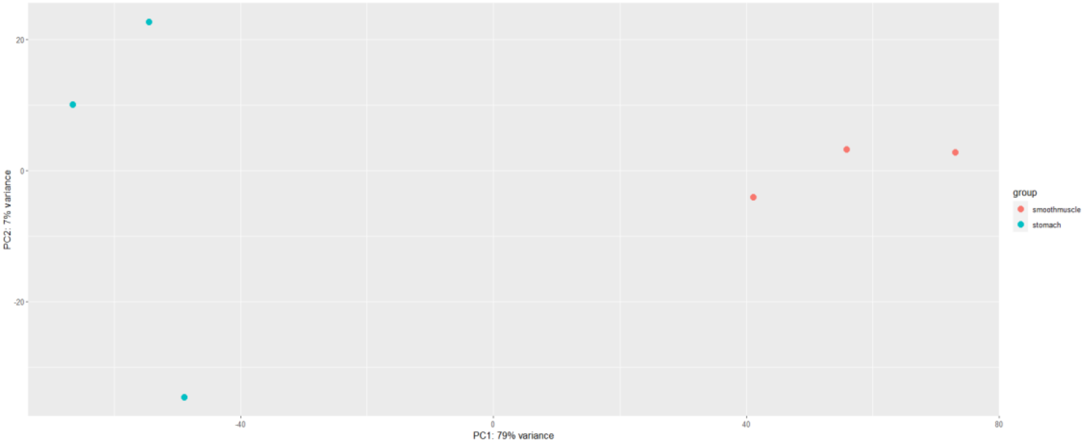


Figure 5: PCA Plot

Principal Component Analysis (PCA) was used to reduce dimensionality and visualize the overall variance in gene expression between samples. PC1 accounted for 79% of the variance and PC2 for 7%, together explaining 86% of the total variance. This indicates a strong separation between the two tissue types. Samples within each group (stomach and smooth muscle) formed tight clusters, reflecting internal consistency, while separation along PC1 highlights distinct expression profiles between groups.

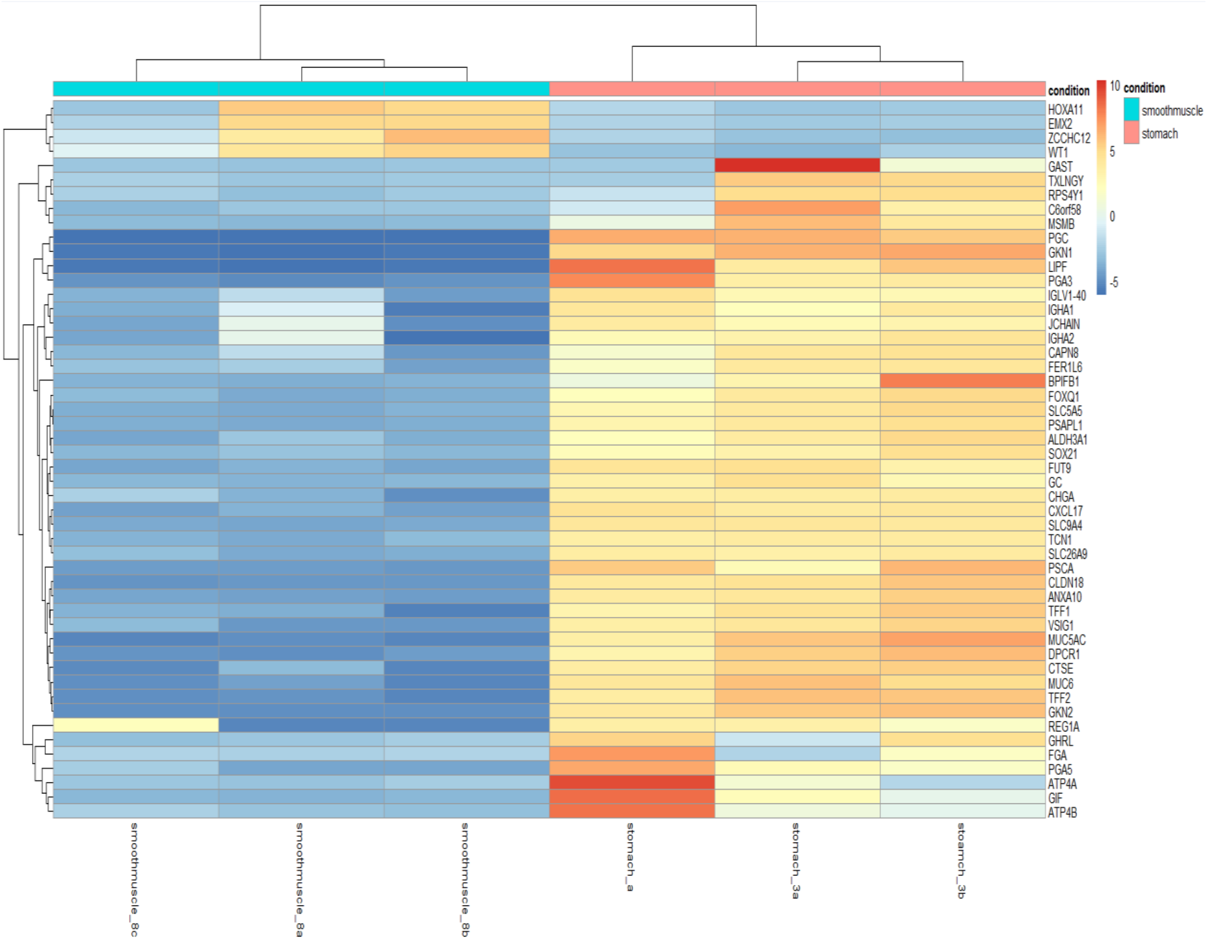


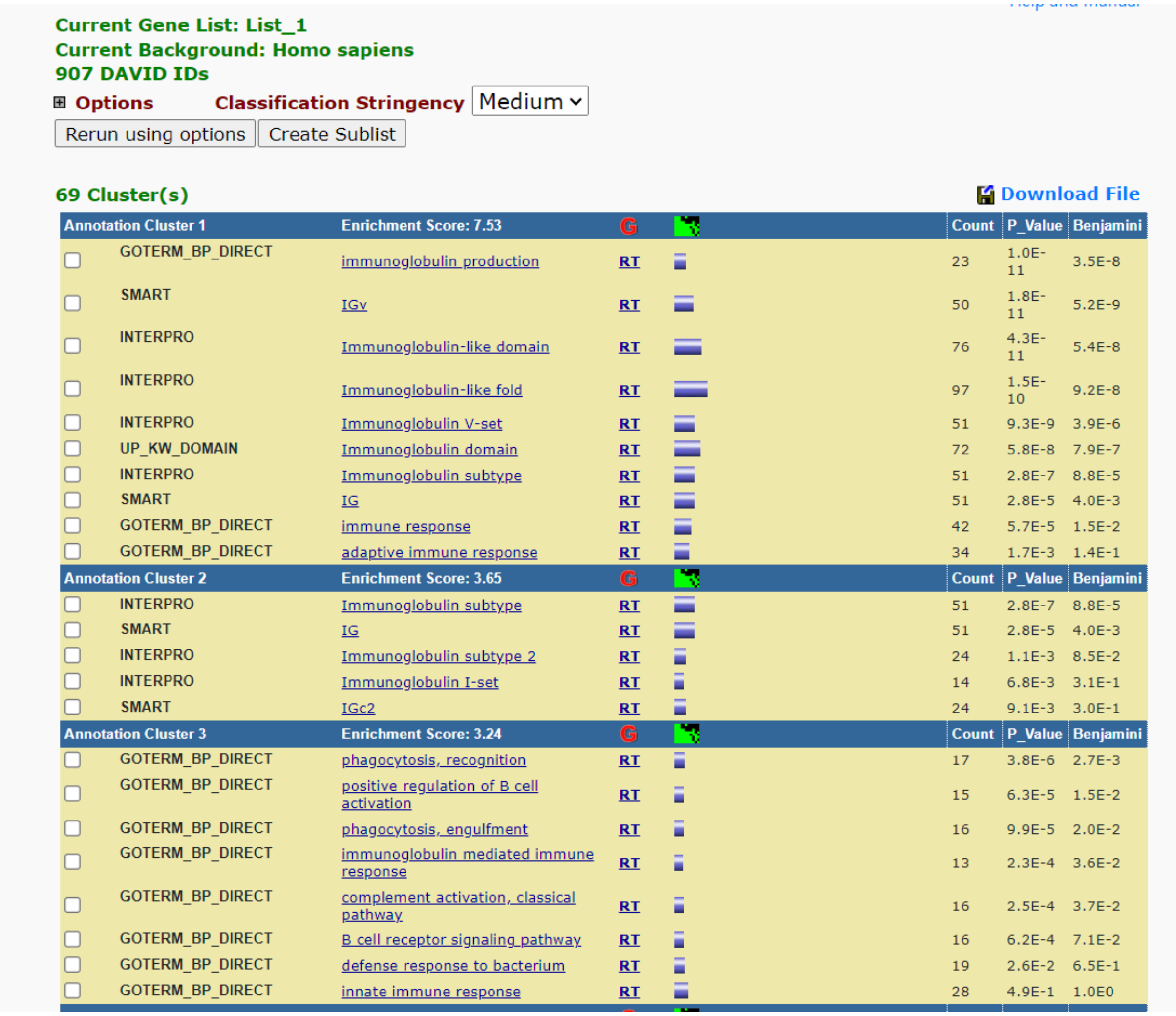
Figure 6: Heatmap

Expression levels of the top 50 most variable genes were visualized. The heatmap uses a color gradient where blue indicates low expression, red indicates high expression, and yellow/orange represent intermediate levels. Both samples and genes are hierarchically clustered.

Stomach samples exhibited high expression in genes toward the right side of the heatmap, while smooth muscle samples showed high expression in genes toward the left. Samples such as smoothmuscle_8a and smoothmuscle_8b clustered closely due to moderately high expression of initial genes and low expression of others. Stomach samples like stomach_3a and stomach_3b shared expression trends distinct from the smooth muscle group. This clustering supports strong separation between tissue types, confirming transcriptomic differences.

Enrichment:

The differentially expressed genes identified using edgeR were analyzed through DAVID for functional enrichment. Clustered annotation terms included immune response, immunoglobulin domain, and actin cytoskeleton organization. Genes such as MYLK, ACTA2, and CNN1 were enriched in pathways related to muscle contraction and cytoskeletal structure, while GKN1 and TFF2 were linked to gastric-specific mucosal maintenance. These enrichment results help contextualize the functional roles of DEGs within gastric cancer biology.



Current Gene List: List_1
Current Background: Homo sapiens
907 DAVID IDs

⊞ **Options** **Classification Stringency** Medium

69 Cluster(s)

Annotation Cluster 1	Enrichment Score: 7.53	G	Count	P_Value	Benjamini
☐ GOTERM_BP_DIRECT	immunoglobulin production	RT	23	1.0E-11	3.5E-8
☐ SMART	IGv	RT	50	1.8E-11	5.2E-9
☐ INTERPRO	Immunoglobulin-like domain	RT	76	4.3E-11	5.4E-8
☐ INTERPRO	Immunoglobulin-like fold	RT	97	1.5E-10	9.2E-8
☐ INTERPRO	Immunoglobulin V-set	RT	51	9.3E-9	3.9E-6
☐ UP_KW_DOMAIN	Immunoglobulin domain	RT	72	5.8E-8	7.9E-7
☐ INTERPRO	Immunoglobulin subtype	RT	51	2.8E-7	8.8E-5
☐ SMART	IG	RT	51	2.8E-5	4.0E-3
☐ GOTERM_BP_DIRECT	immune response	RT	42	5.7E-5	1.5E-2
☐ GOTERM_BP_DIRECT	adaptive immune response	RT	34	1.7E-3	1.4E-1
Annotation Cluster 2	**Enrichment Score: 3.65**	**G**	**Count**	**P_Value**	**Benjamini**
☐ INTERPRO	Immunoglobulin subtype	RT	51	2.8E-7	8.8E-5
☐ SMART	IG	RT	51	2.8E-5	4.0E-3
☐ INTERPRO	Immunoglobulin subtype 2	RT	24	1.1E-3	8.5E-2
☐ INTERPRO	Immunoglobulin I-set	RT	14	6.8E-3	3.1E-1
☐ SMART	IGc2	RT	24	9.1E-3	3.0E-1
Annotation Cluster 3	**Enrichment Score: 3.24**	**G**	**Count**	**P_Value**	**Benjamini**
☐ GOTERM_BP_DIRECT	phagocytosis, recognition	RT	17	3.8E-6	2.7E-3
☐ GOTERM_BP_DIRECT	positive regulation of B cell activation	RT	15	6.3E-5	1.5E-2
☐ GOTERM_BP_DIRECT	phagocytosis, engulfment	RT	16	9.9E-5	2.0E-2
☐ GOTERM_BP_DIRECT	immunoglobulin mediated immune response	RT	13	2.3E-4	3.6E-2
☐ GOTERM_BP_DIRECT	complement activation, classical pathway	RT	16	2.5E-4	3.7E-2
☐ GOTERM_BP_DIRECT	B cell receptor signaling pathway	RT	16	6.2E-4	7.1E-2
☐ GOTERM_BP_DIRECT	defense response to bacterium	RT	19	2.6E-2	6.5E-1
☐ GOTERM_BP_DIRECT	innate immune response	RT	28	4.9E-1	1.0E0

Figure 7: Enrichment using EdgeR

Genes	logFC	logCPM	PValue	FDR
MYLK	5.488904	11.04349	7.75E-16	7.35E-13
ACTA2	5.540031	11.28194	3.12E-13	1.51E-10
MYH11	4.360866	11.73351	2.39E-13	1.26E-10
TAGLN	4.953564	10.87195	3.01E-13	1.48E-10
CNN1	5.828538	9.873856	2.97E-15	2.46E-12
CALD1	4.045615	10.29024	3.86E-10	7.85E-08
MYOCD	2.698763	5.64974	3.75E-08	4.27E-06
SMTN	2.734777	7.516364	2.36E-06	0.000143
ACTG2	7.032245	11.99322	2.06E-14	1.43E-11

Table 1: List of genes of smooth muscle

Genes	logFC	logCPM	PValue	FDR
TFF2	-14.5412	10.23261	3.45E-31	1.35E-27
MUC6	-13.1713	10.07954	9.18E-27	3.00E-23
PGC	-19.5956	13.12081	6.44E-67	3.79E-62
GKN1	-20.8346	12.57812	7.45E-50	1.10E-45
GKN2	-16.1191	9.640613	7.34E-40	4.81E-36
CLDN18	-12.5616	9.799127	1.36E-38	8.02E-35
GHRL	-9.04537	6.640259	5.11E-08	5.50E-06

Table 2: List of genes of stomach

From Tables 1 and 2, differential expression analysis between stomach and smooth muscle tissues identified a total of 1,587 upregulated and 2,243 downregulated genes. These values were based on an FDR cutoff of < 0.05. Remaining genes (55,004) showed no significant differential expression. The top DEGs demonstrate distinct molecular profiles between the two tissues, with smooth muscle showing enrichment for contractile and cytoskeletal components and stomach tissue enriched for mucosal and epithelial-specific genes.

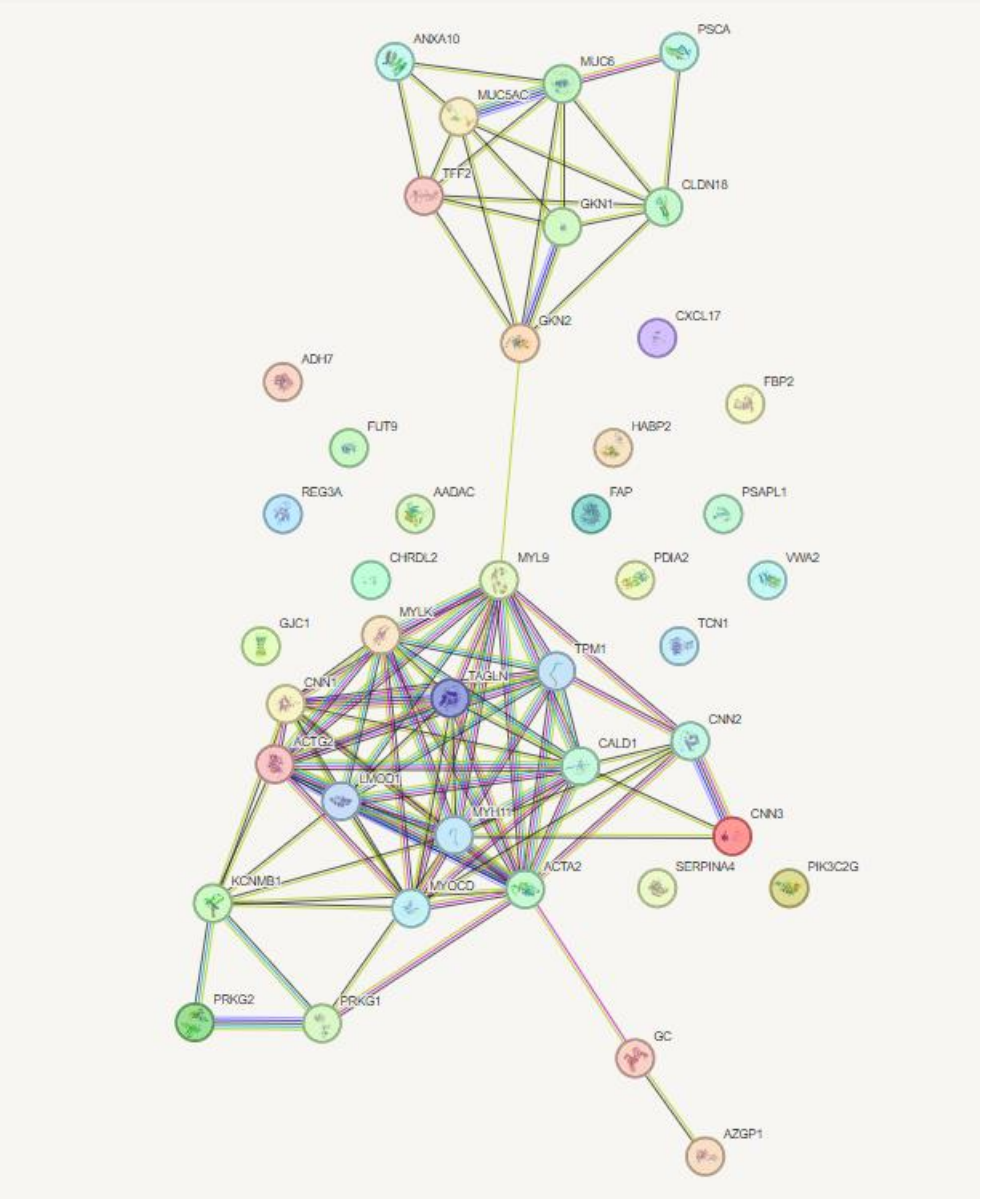


Figure 8: String network

The STRING network plot illustrates predicted protein-protein interactions among the differentially expressed genes. Densely connected nodes represent functionally related clusters. Genes with increased expression levels are often located in highly interconnected hubs (e.g., CNN2, ACTA2, MYLK), highlighting roles in muscle contraction and cytoskeletal regulation. Conversely, stomach-specific downregulated genes (e.g., GKN1, TFF2, GKN2) appear in more isolated subnetworks, reflecting their specialized and possibly non-overlapping functions in gastric mucosa. This network view reinforces the functional distinction between the two tissue types based on expression connectivity.

Discussion:

The findings confirm significant transcriptomic divergence between gastric and smooth muscle tissues. Upregulated genes such as CNN2, MYLK, and ACTA2 align with smooth muscle functionality. Downregulated gastric genes, including GKN1 and TFF2, are associated with tumor suppression and gastric epithelial maintenance. These results validate the potential of these genes as gastric cancer biomarkers or therapeutic targets.

Conclusion:

This integrative transcriptomic study reveals clear gene expression differences between gastric and smooth muscle tissues, highlighting critical molecular signatures linked to gastric cancer. The identification of tissue-specific expression profiles, validated through statistical analysis, clustering, and functional enrichment, underscores the utility of RNA-Seq in cancer research. The upregulation of muscle-related genes and downregulation of gastric tumor suppressors suggests potential biomarkers and therapeutic targets. These findings contribute to a deeper understanding of gastric cancer biology and lay the groundwork for future translational studies.

Citation:

- Hu, J. et al. (2017). Tumour Biology. https://doi.org/10.1177/1010428317706455
- Luo, J., & Xiang, H. (2021). Bioengineered. https://doi.org/10.1080/21655979.2021.1944019
- di Stadio, C. S. et al. (2019). Biochimie. https://doi.org/10.1016/j.biochi.2019.09.005
- Wang, J. et al. (2023). Frontiers in Oncology. https://doi.org/10.3389/fonc.2023.1142017