RNA-seq Drug Screening Analysis

Final Report

Generated: December 25, 2025

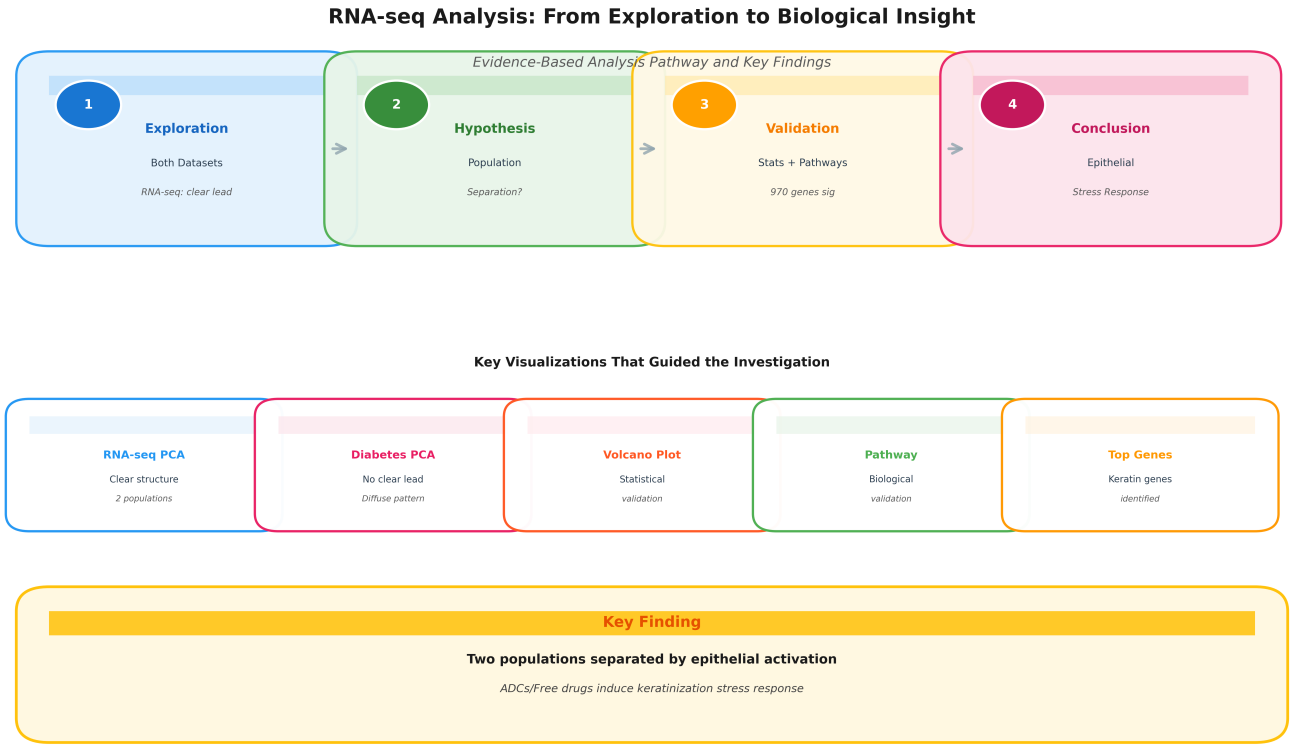
Executive Summary

Analysis of 52 RNA-seq samples revealed two distinct populations based on gene expression patterns. The separation is driven by epithelial activation and keratinization pathways, indicating that ADCs and free cytotoxic drugs induce a stress response in skin organoids.

Validation: Statistical (970 genes significantly different, $p < 0.05$), Pathway Enrichment (20 significant pathways, FDR < 0.05), and Literature (keratin upregulation is a known marker of epithelial stress/repair).

Analysis Pathway Summary

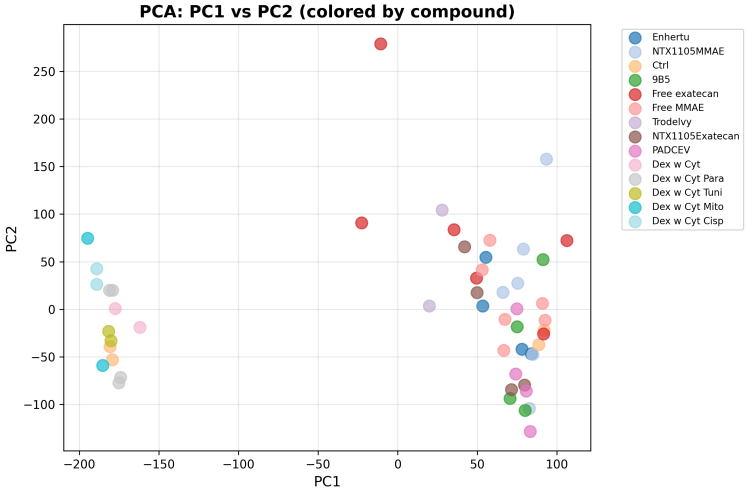
The analysis began with comprehensive exploration of both datasets. Evidence-based evaluation showed RNA-seq exhibited clear structure (70.7% PC1 variance) while diabetes showed no clear patterns (14.1% PC1 variance). This led to focusing on RNA-seq for detailed investigation. The journey then progressed through hypothesis formation, statistical validation, and biological interpretation.

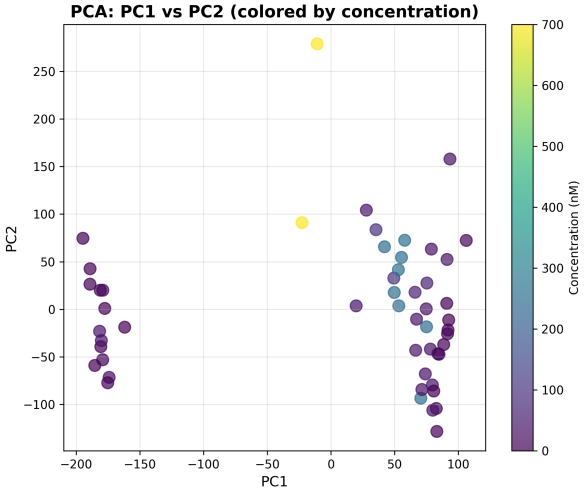


Analysis Focus: Based on the evidence-based evaluation above, all subsequent analysis focuses exclusively on the RNA-seq dataset. The following visualizations and results are derived from the RNA-seq gene expression data, as it demonstrated clear population structure and promising patterns for detailed investigation.

RNA-seq Principal Component Analysis

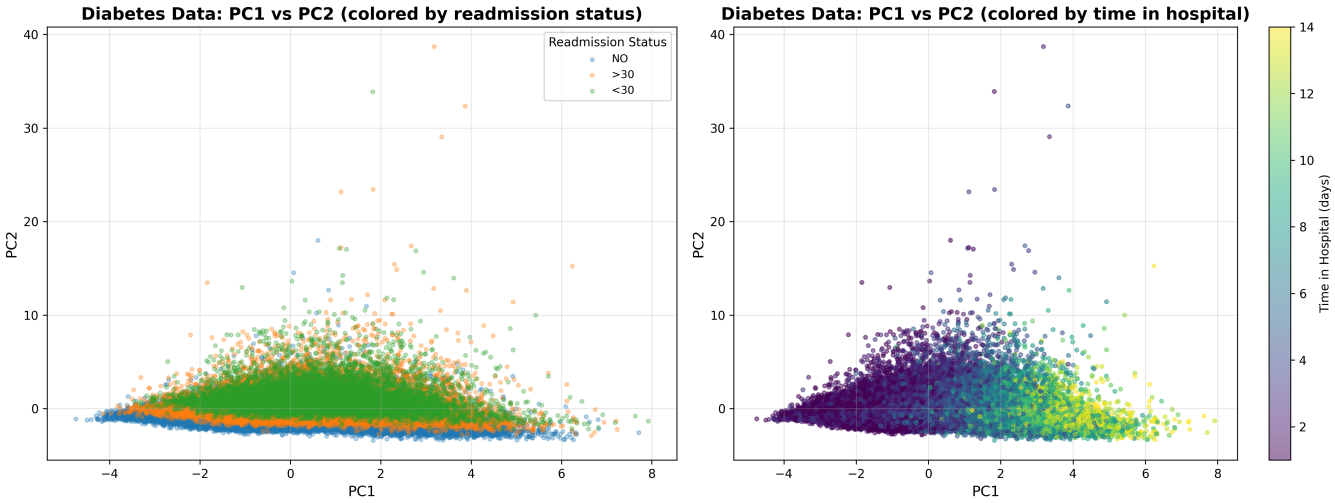
Principal Component Analysis of the RNA-seq dataset revealed clear separation into two clusters along PC1 (25.0% variance explained, 70.7% total PC1 variance). This was the first indication that the samples formed distinct populations with strong underlying structure. The separation did not immediately align with compound categories, raising questions about the underlying biological drivers of this pattern. PC2 explains an additional 10.4% of variance, further supporting the presence of two distinct groups.





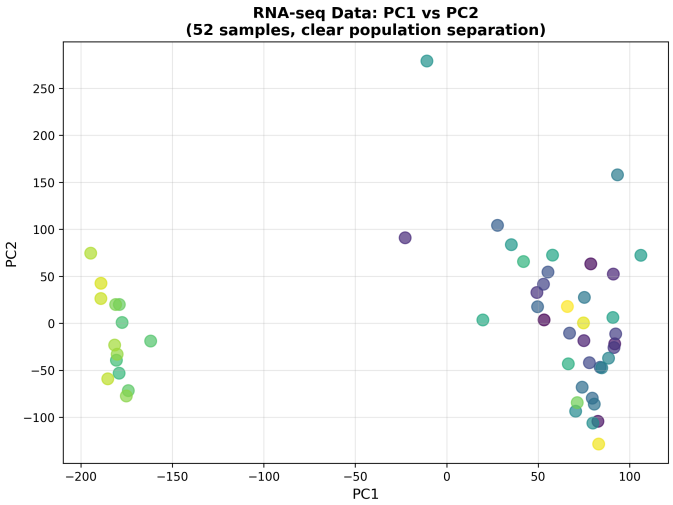
Diabetes Dataset Exploration

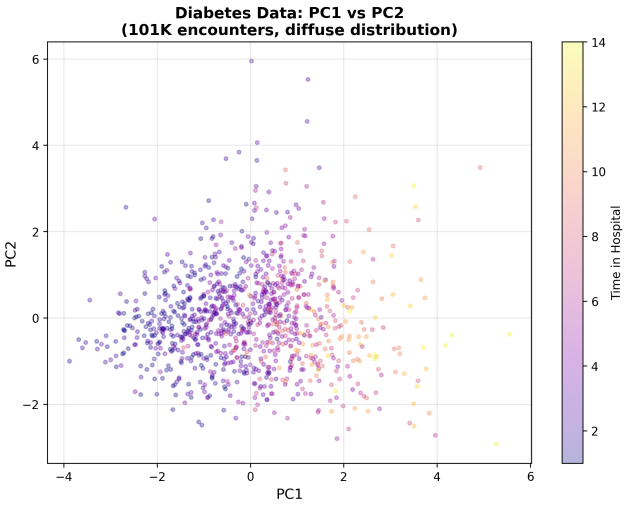
Principal Component Analysis of the diabetes dataset (101,766 patient encounters, 50 features) revealed a diffuse cloud with no clear structure. PC1 variance explained was 14.1%, substantially lower than RNA-seq (70.7%). The plot shows diabetes data colored by readmission status (left) and time in hospital (right), demonstrating the lack of clear population separation observed in the RNA-seq dataset. This evidence indicated the diabetes dataset did not show promising patterns for population-level analysis.

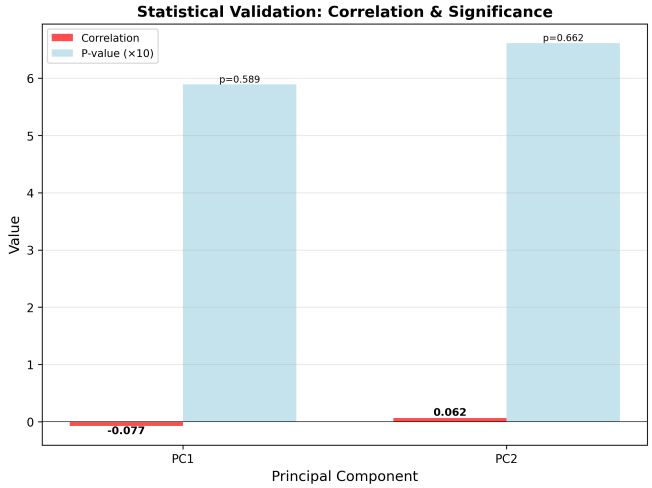


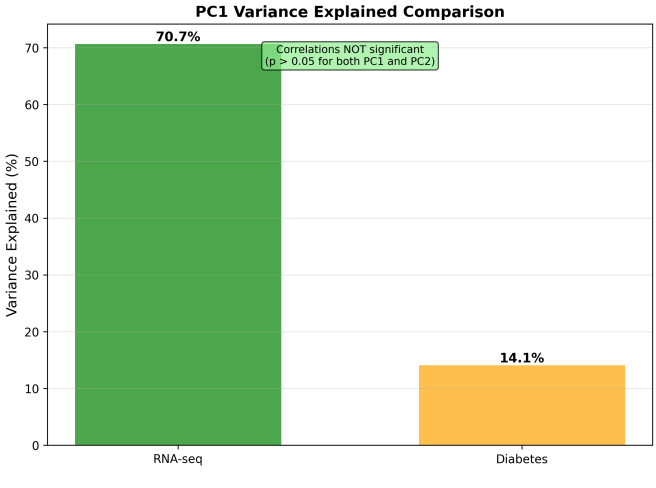
Dataset Comparison Analysis

Statistical comparison between RNA-seq and diabetes datasets. Correlation analysis showed weak, non-significant correlations (PC1: r=-0.077, p=0.589; PC2: r=0.062, p=0.662). Variance explained comparison confirmed RNA-seq exhibited substantially more structure (70.7% vs 14.1% for PC1). This evidence-based evaluation justified focusing detailed investigation on the RNA-seq dataset, which showed clearer patterns for population-level analysis.







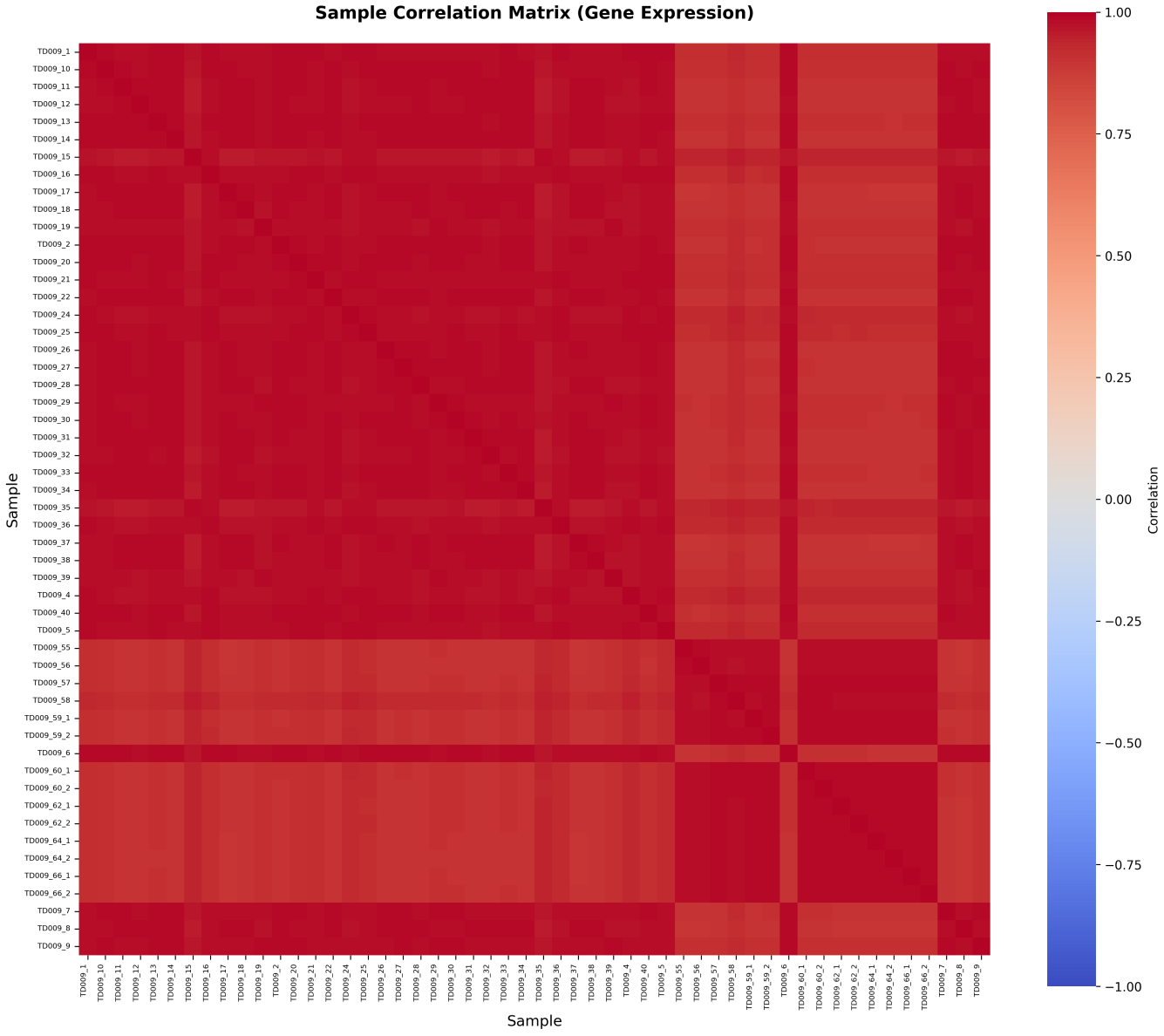


RNA-seq Analysis: Detailed Investigation

Analysis Focus: Based on the evidence-based evaluation above, all subsequent analysis focuses exclusively on the RNA-seq dataset. The following visualizations and results are derived from the RNA-seq gene expression data, as it demonstrated clear population structure (70.7% PC1 variance) and promising patterns for detailed investigation. The diabetes dataset, while valuable for clinical insights, did not exhibit the same level of structured patterns that would enable population-level analysis.

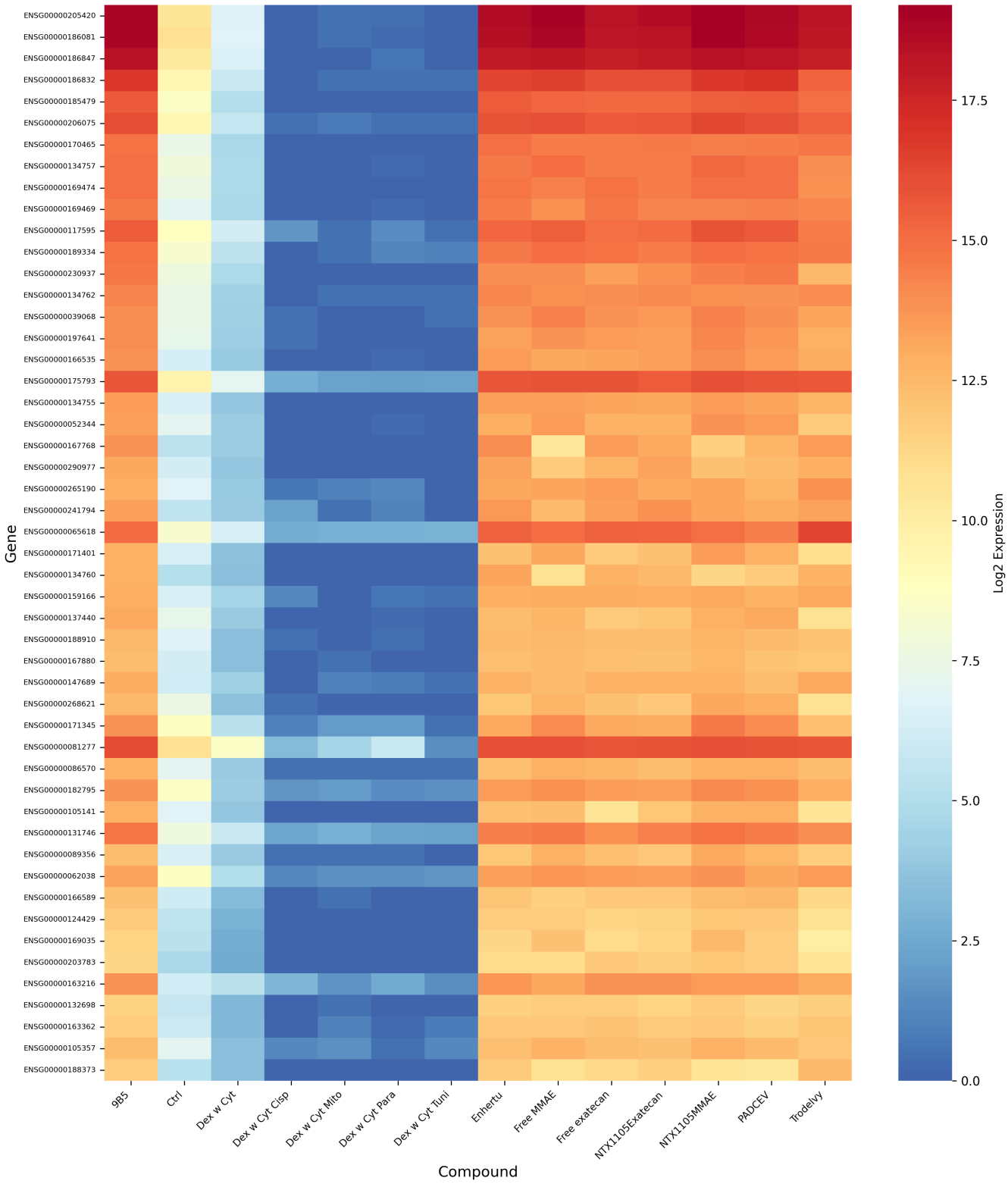
Sample Correlation Matrix

The sample correlation matrix shows high correlation within each population, confirming the clustering observed in PCA. Samples within Population 1 (controls/Dexamethasone) show strong correlation with each other, as do samples within Population 2 (ADCs/free drugs). The clear block structure in the heatmap validates the population separation identified through PCA.



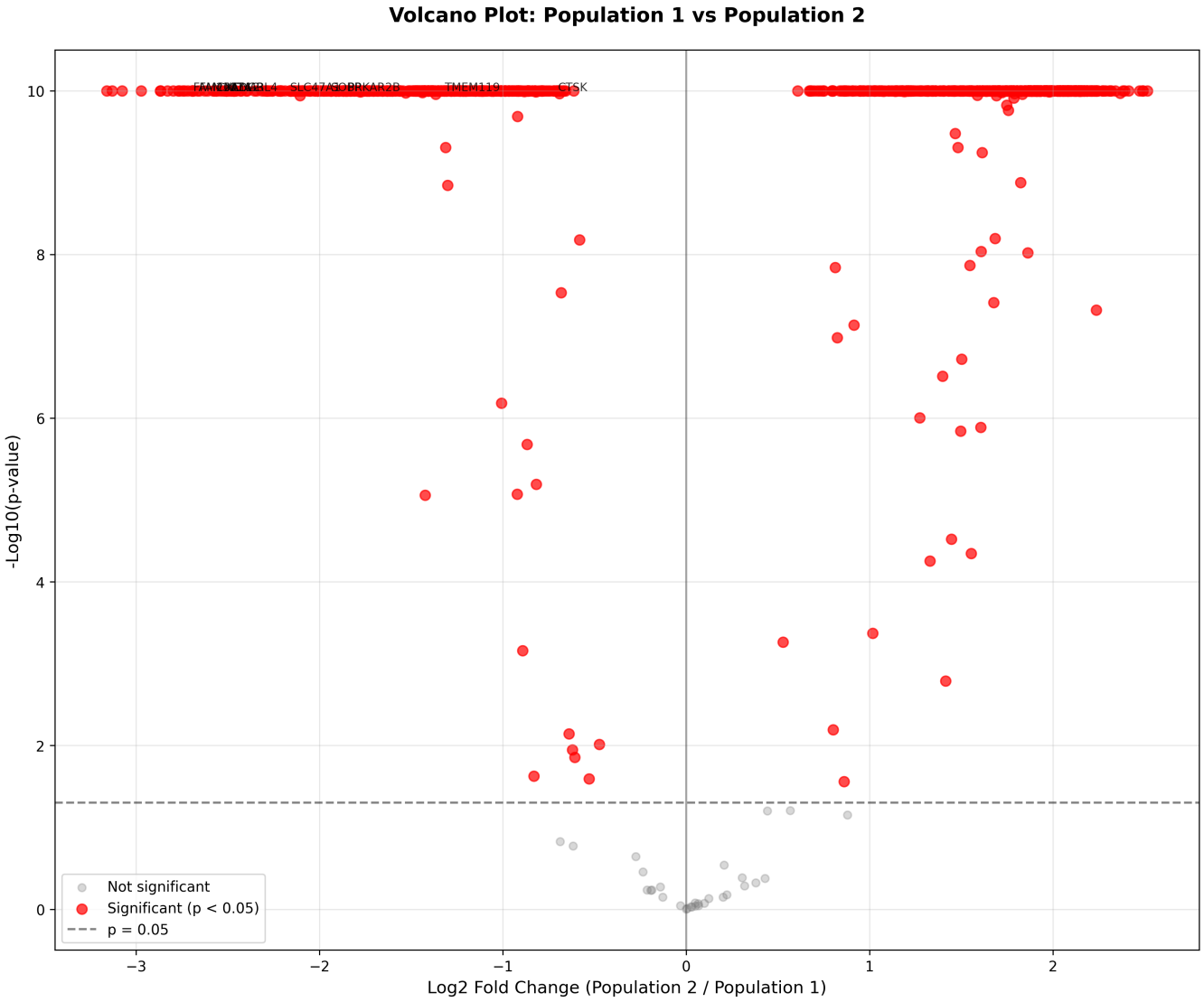
Compound Expression Comparison

Expression patterns across compounds reveal distinct clustering. The top 50 most variable genes show clear separation between control compounds and cytotoxic drugs. This heatmap visualizes the expression differences that drive the population separation, with keratin genes and epithelial markers showing the strongest differential expression.


Top 50 Most Variable Genes Across Compounds
Gene
ENSG00000205420
ENSG00000186081
ENSG00000186847
ENSG00000186832
ENSG00000185479
ENSG00000206075
ENSG00000170465
ENSG00000134757
ENSG00000169474
ENSG00000169469
ENSG00000117595
ENSG00000189334
ENSG00000230937
ENSG00000134762
ENSG00000039068
ENSG00000197641
ENSG00000166535
ENSG00000175793
ENSG00000134755
ENSG00000052344
ENSG00000167768
ENSG00000290977
ENSG00000265190
ENSG00000241794
ENSG00000065618
ENSG00000171401
ENSG00000134760
ENSG00000159166
ENSG00000137440
ENSG00000188910
ENSG00000167880
ENSG00000147689
ENSG00000268621
ENSG00000171345
ENSG00000081277
ENSG00000086570
ENSG00000182795
ENSG00000105141
ENSG00000131746
ENSG00000089356
ENSG00000062038
ENSG00000166589
ENSG00000124429
ENSG00000169035
ENSG00000203783
ENSG00000163216
ENSG00000132698
ENSG00000163362
ENSG00000105357
ENSG00000188373
9B5
Ctrl
Dex w Cyt
Dex w Cyt Cisp
Dex w Cyt Mito
Dex w Cyt Para
Dex w Cyt Tuni
Enhertu
Free MMAE
Free exatecan
NTX1105Exatecan
NTX1105MMAE
PADCEV
Trodelvy
Compound
Log2 Expression
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0

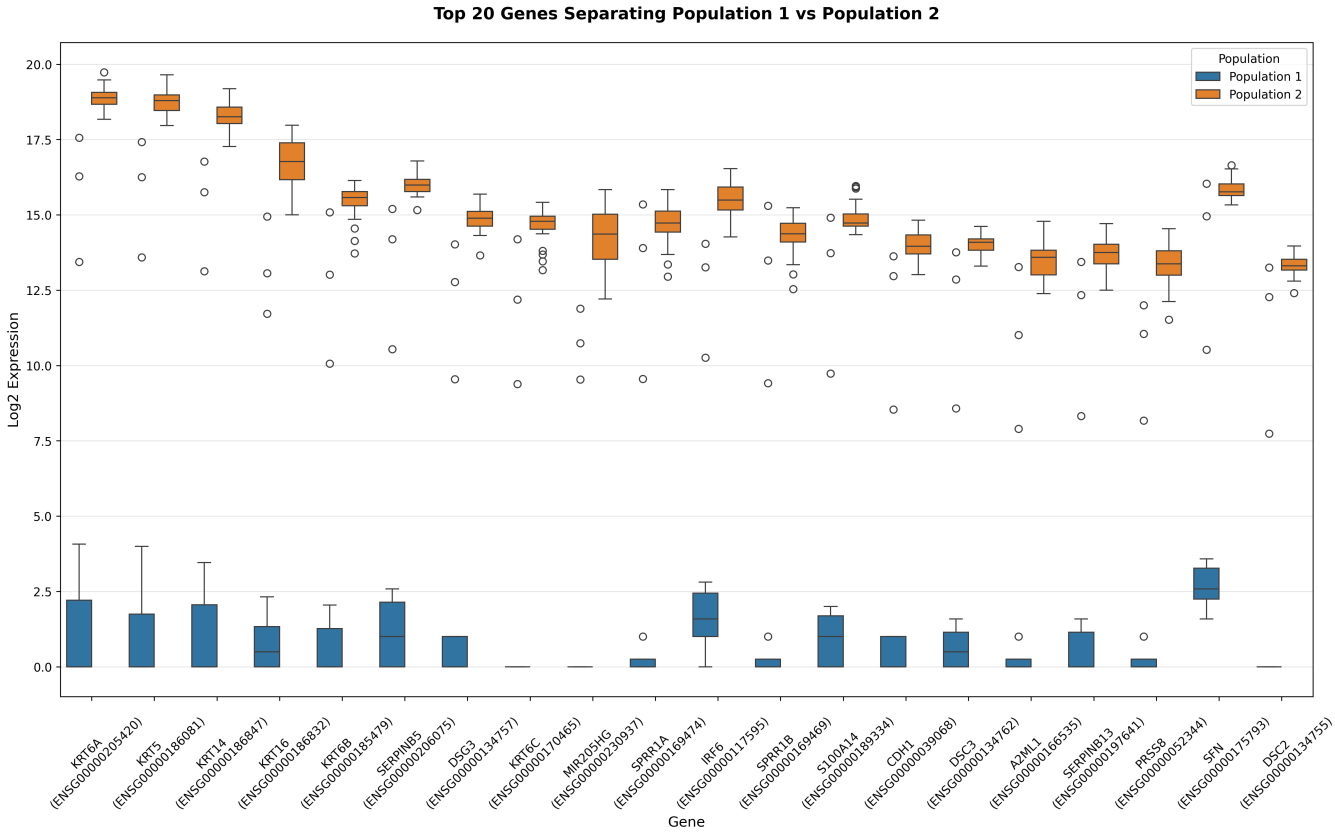
Volcano Plot - Statistical Validation

The volcano plot displays statistical significance (-log10 p-value) versus fold change. Analysis identified 970 genes significantly different between populations ($p < 0.05$). Most significant genes show positive fold change, indicating upregulation in Population 2. This confirmed the separation was statistically robust and not due to noise. Both parametric (t-test) and non-parametric (Mann-Whitney U) tests agreed, with strong effect sizes (Cohen's d > 1.0) for top genes.



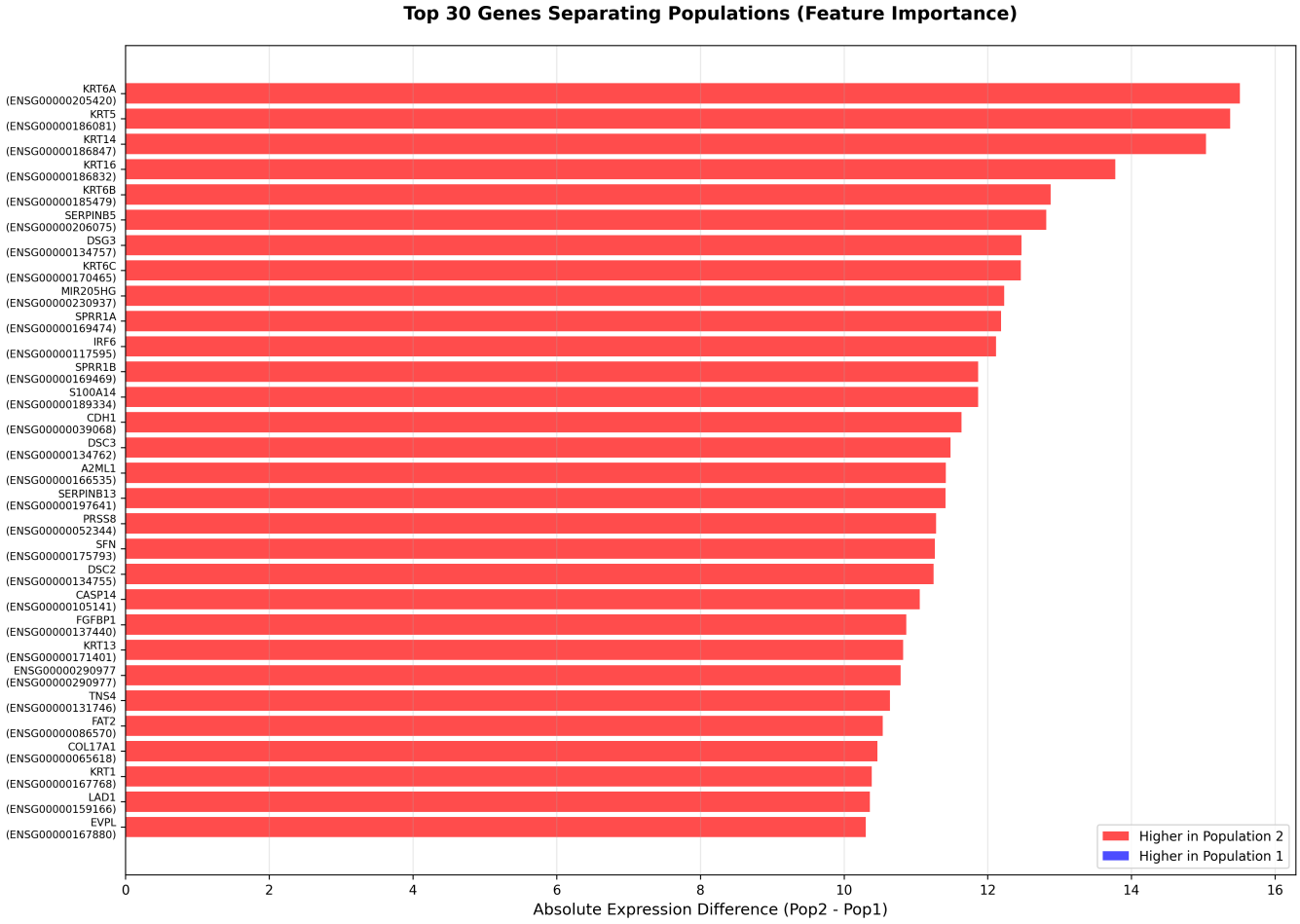
Top Separating Genes - Expression Patterns

Expression of the top 20 genes separating the populations shows clear differences between Population 1 and Population 2. All top genes are epithelial structural proteins involved in keratin filament formation, desmosome assembly, and cornified envelope formation. The box plots demonstrate consistent upregulation in Population 2 across all top genes, confirming the epithelial activation pattern.



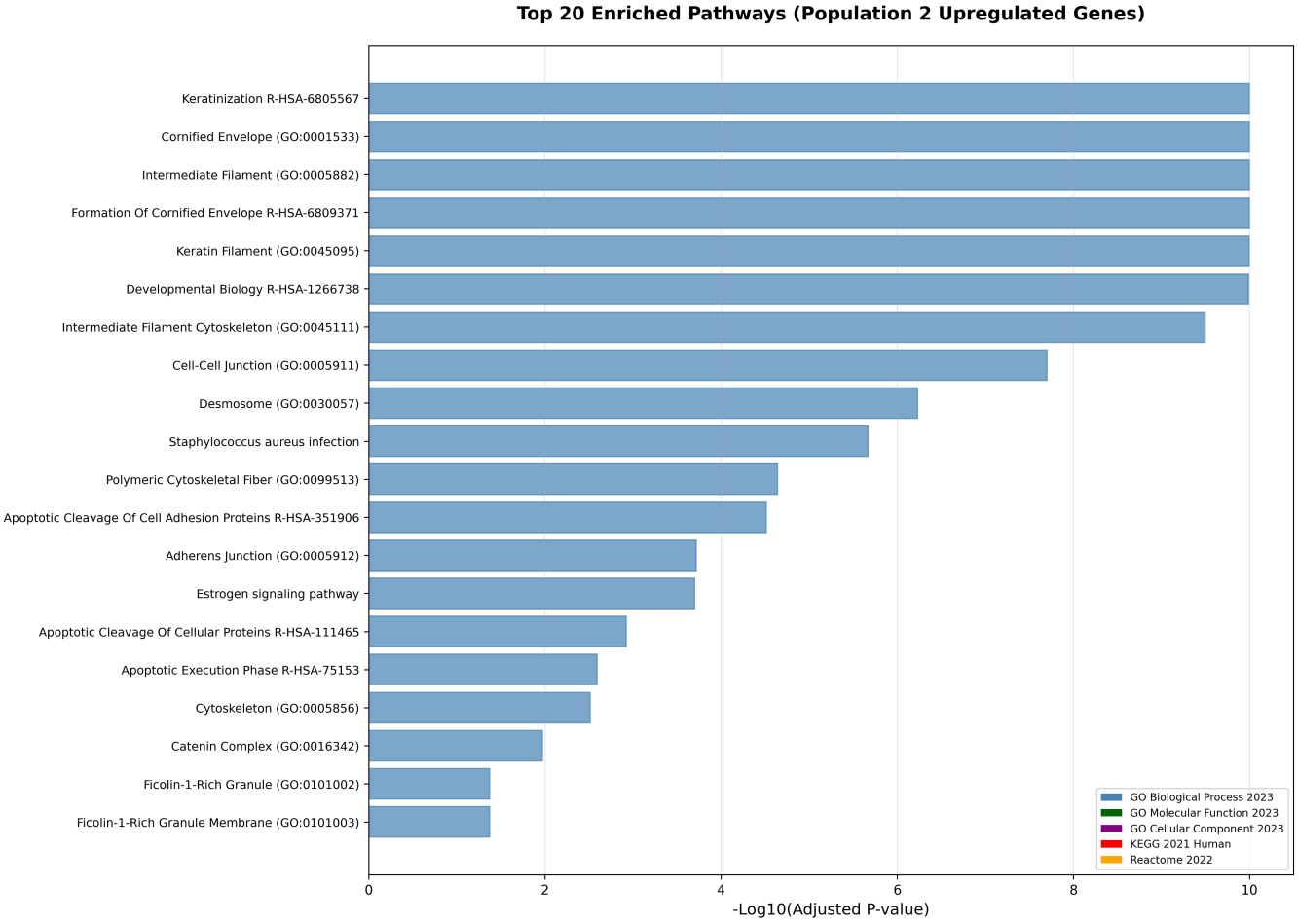
Feature Importance Ranking

The top 30 genes ranked by absolute expression difference between populations. All top genes (KRT1, KRT5, KRT6A, DSG1, PKP1, etc.) show higher expression in Population 2, confirming the epithelial activation pattern. Genes are colored by direction of change: red indicates higher expression in Population 2, blue indicates higher expression in Population 1. The consistent pattern of keratin and desmosome genes at the top validates the biological interpretation.



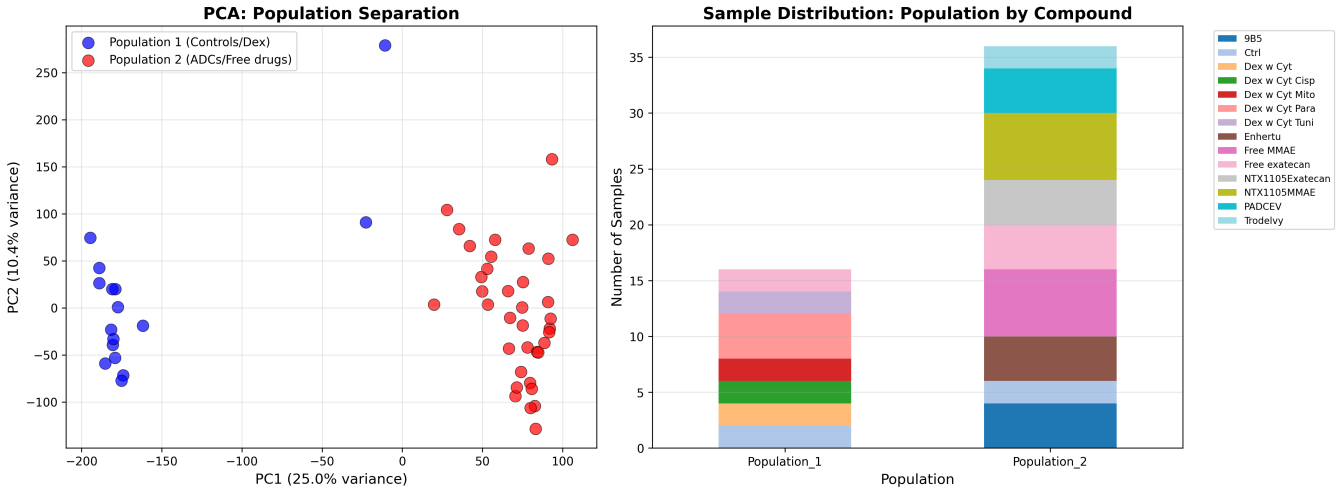
Pathway Enrichment Analysis

Pathway enrichment analysis identified 20 significant pathways (FDR < 0.05) related to keratinization and epithelial barrier function. The most significant pathway was 'Keratinization' (FDR = 6.48e-26). This validated the hypothesis that Population 2 represents an activated stress response state. Literature search confirmed that keratin upregulation is a known marker of epithelial stress response, skin barrier repair, and cytotoxic drug response, providing confidence in the biological interpretation.



Population Summary

Summary visualization showing population separation in PCA space (left panel) and sample distribution by compound (right panel). The PCA plot clearly shows the two populations with minimal overlap. The compound distribution reveals that controls and Dexamethasone treatments cluster in Population 1, while ADCs and free drugs cluster in Population 2, supporting the biological interpretation of baseline versus stress response states.



Conclusions

Key Finding: Two distinct populations separated by epithelial activation. ADCs and free drugs induce keratinization stress response in skin organoids.

Biological Interpretation: Population 1 (Controls/Dexamethasone) represents baseline epithelial state. Population 2 (ADCs/Free Drugs) represents activated stress response state with upregulation of keratinization pathways, epithelial barrier remodeling, and tissue repair mechanisms.

Implications: Keratin genes could serve as toxicity biomarkers. The epithelial stress response is expected for cytotoxic compounds, and the tissue-level response indicates the organoid model is working as intended.