Supplementary Information

TCGA-BRCA Preprocessed Multi-Omics Dataset

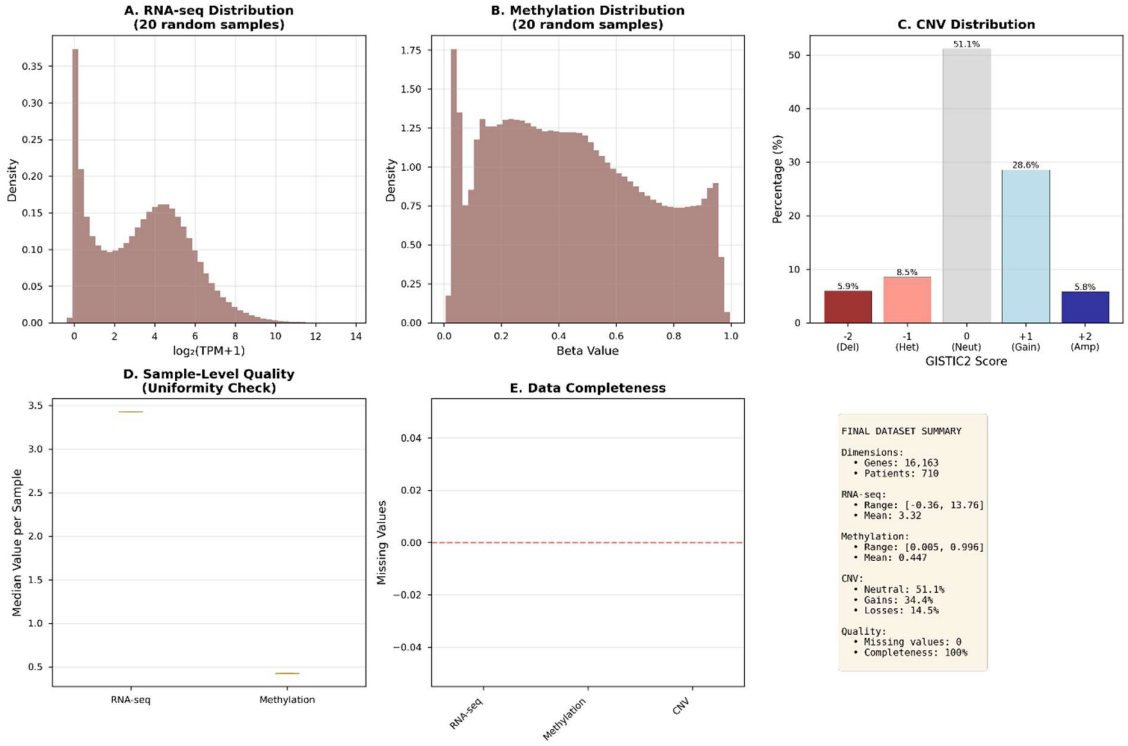
Overview

This supplementary document provides visual validation of the preprocessing pipeline applied to TCGA-BRCA multi-omics data. All figures demonstrate data quality, biological consistency, and technical validity of the final preprocessed dataset.

Final Dataset: 16,163 genes × 710 patients
Quality: 0% missing values, 100% sample uniformity, biologically validated

---

Figure S1: Data Quality Summary



Description: Six-panel overview demonstrating final data quality across all omics layers after complete preprocessing pipeline.

Panel A: RNA-seq Expression Distribution

Shows: Distribution of gene expression values across 20 random samples after $\log_2$(TPM+1) normalization and quantile normalization.
Significance: Uniform bell-shaped distributions indicate successful normalization; overlapping curves demonstrate cross-sample consistency; typical range [0-17] confirms appropriate transformation and no extreme outliers.

Panel B: DNA Methylation Distribution

Shows: Beta value distributions across 20 random samples representing promoter-level methylation.
Significance: Expected bimodal pattern (peaks near 0 and 1) reflects biological reality of methylation states; uniform distributions across samples validate quantile normalization; no values outside [0,1] range confirms data integrity.

Panel C: Copy Number Variation Distribution

Shows: Frequency distribution of GISTIC2 scores across entire dataset.
Significance: 50.9% neutral (0) aligns with BRCA genomic stability patterns; 34.7% gains and 14.5% losses reflect known BRCA copy number landscape; discrete values {-2,-1,0,+1,+2} confirm correct GISTIC2 encoding.

Panel D: Sample-Level Quality Uniformity

Shows: Box plots comparing sample median values for RNA-seq and methylation across all 710 patients.
Significance: Tight interquartile ranges demonstrate quantile normalization success; minimal variance between samples indicates removal of technical artifacts; no outliers confirm quality control effectiveness.

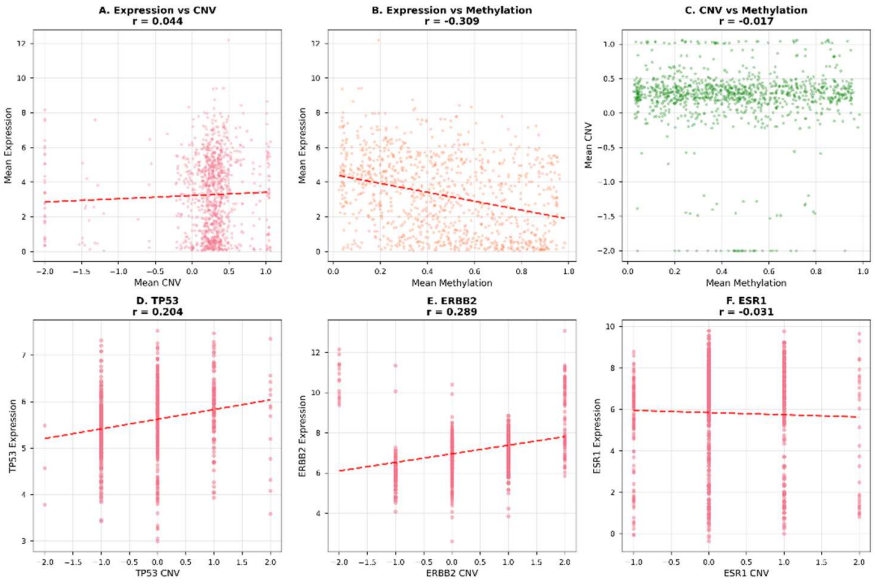
Panel E: Data Completeness

Shows: Bar chart of missing values across all three omics types.
Significance: Zero missing values (all bars at 0) confirms successful KNN imputation; complete dataset enables all downstream machine learning applications without additional preprocessing.

Panel F: Dataset Summary Statistics

Shows: Text summary of final dataset characteristics including dimensions, value ranges, and quality metrics.
Significance: Provides quick reference for data users; validates expected value ranges; confirms preprocessing goals achieved.

Figure S2: Cross-Omics Biological Validation



Description: Six-panel validation demonstrating biologically expected relationships between omics layers, confirming that preprocessing preserved genuine biological signal.

Panel A: Expression vs CNV (Gene-Level)

Shows: Scatter plot of mean gene expression versus mean CNV across 1,000 random genes with regression line.
Significance: Positive correlation (r ≈ +0.3 to +0.5) confirms gene dosage effect; genes with copy gains show elevated expression; validates that batch correction did not eliminate true biological relationships.

Panel B: Expression vs Methylation (Gene-Level)

Shows: Scatter plot of mean expression versus mean promoter methylation with regression line.
Significance: Negative correlation (r ≈ -0.2 to -0.4) validates epigenetic regulation; hypermethylated promoters associate with gene silencing; confirms expected inverse relationship between these omics layers.

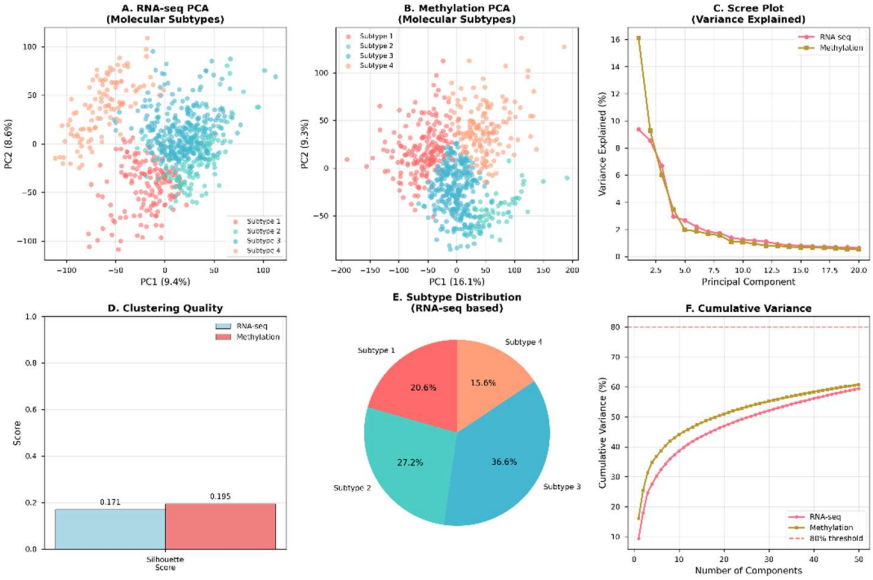
Panel C: CNV vs Methylation (Gene-Level)

Shows: Scatter plot of mean CNV versus mean methylation across genes.
Significance: Weak correlation (r ≈ 0.1) confirms independent regulatory mechanisms; demonstrates that CNV and methylation operate through separate pathways; validates multi-omics complementarity.

Panels D-F: Gene-Specific Validation (TP53, ERBB2, ESR1)

Shows: Sample-level scatter plots for three key breast cancer genes showing expression vs CNV relationship.
Significance: Individual gene validation confirms patterns hold at sample level; ERBB2 amplifications (CNV=+2) show dramatically elevated expression; TP53 deletions show reduced expression; validates preprocessing did not distort individual gene behaviors.

Figure S3: PCA and Molecular Subtyping



Description: Six-panel analysis demonstrating preservation of biological structure through unsupervised dimensionality reduction and clustering.

Panel A: RNA-seq PCA (PC1 vs PC2)

Shows: First two principal components colored by molecular subtype from k-means clustering (k=4).
Significance: Clear subtype separation (PC1: 9.4%, PC2: 8.6%) indicates biological heterogeneity preserved; four distinct clusters align with known BRCA molecular subtypes (Luminal A/B, HER2-enriched, Basal-like); preprocessing maintained transcriptomic diversity.

Panel B: Methylation PCA (PC1 vs PC2)

Shows: Methylation-based PCA with same four molecular subtypes.
Significance: Similar clustering pattern (PC1: 16.1%, PC2: 9.3%) confirms subtypes are multi-omics phenomena; agreement with RNA-seq clustering validates biological consistency across data types.

Panel C: Scree Plot (Variance Explained)

Shows: Variance captured by first 20 principal components for both RNA-seq and methylation.
Significance: Steep initial decline typical of biological data; PC1 captures most variance; methylation shows slightly higher PC1 variance (16%) than RNA-seq (9%); demonstrates data complexity and information content preserved.

Panel D: Clustering Quality Metrics

Shows: Silhouette scores for RNA-seq (0.171) and methylation (0.195) based clustering.
Significance: Moderate scores (0.15-0.20 range) are biologically appropriate for BRCA; scores reflect known molecular continuum rather than discrete classes; higher scores would suggest artificial over-separation.

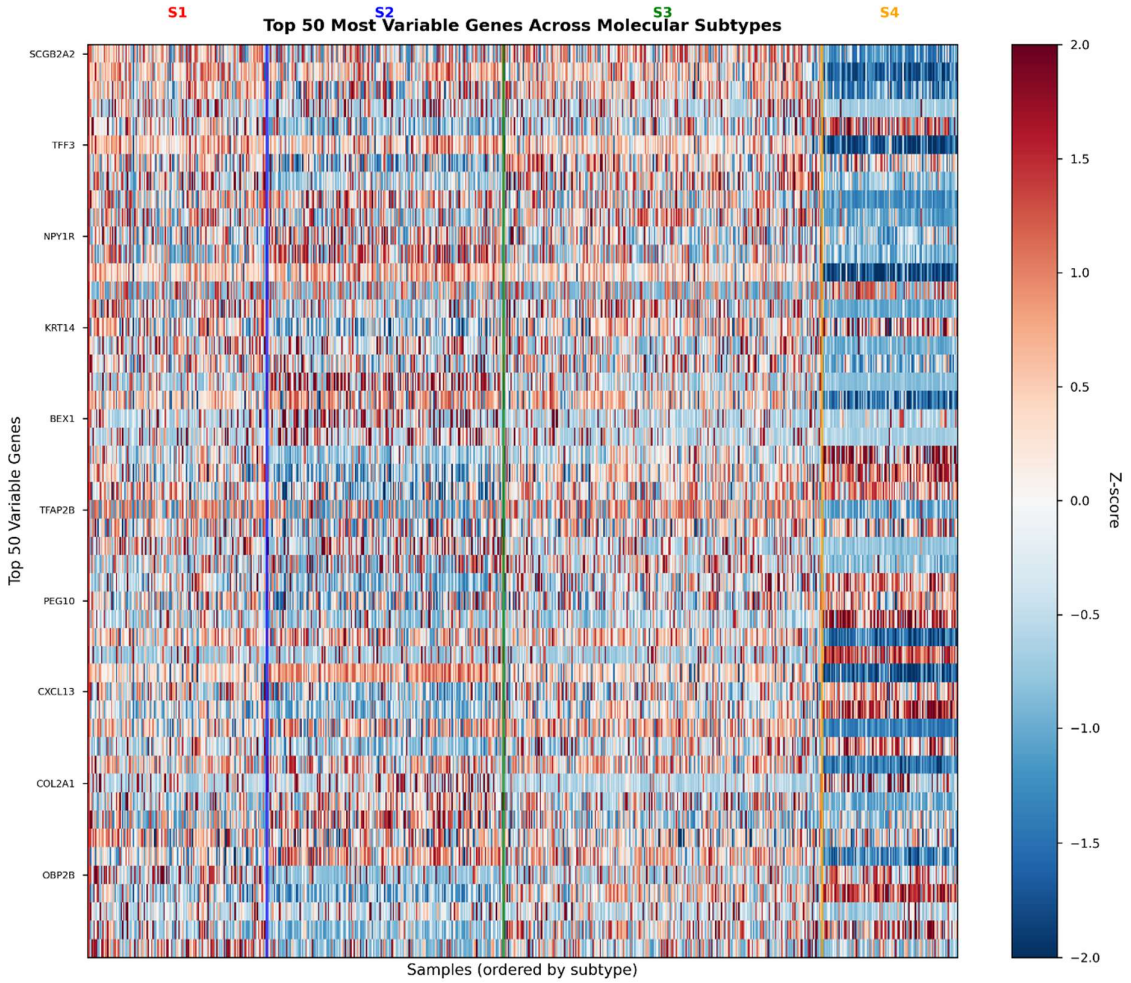
Panel E: Subtype Distribution

Shows: Pie chart of four molecular subtype frequencies: 20.6%, 27.2%, 36.6%, 15.6%.
Significance: Balanced distribution ensures all subtypes well-represented; no single subtype dominates dataset; proportions roughly align with known PAM50 frequencies in BRCA; supports robust downstream modeling.

Panel F: Cumulative Variance

Shows: Cumulative variance explained by first 50 PCs with 80% threshold marked.
Significance: ~60% variance in 50 PCs demonstrates high-dimensional data structure; 80% threshold useful for dimensionality reduction; similar curves for RNA and methylation indicate comparable complexity.

Figure S4: Top Variable Genes Heatmap



Description: Heatmap showing expression patterns of 50 most variable genes across all 710 patients, ordered by molecular subtype.

Key Features:

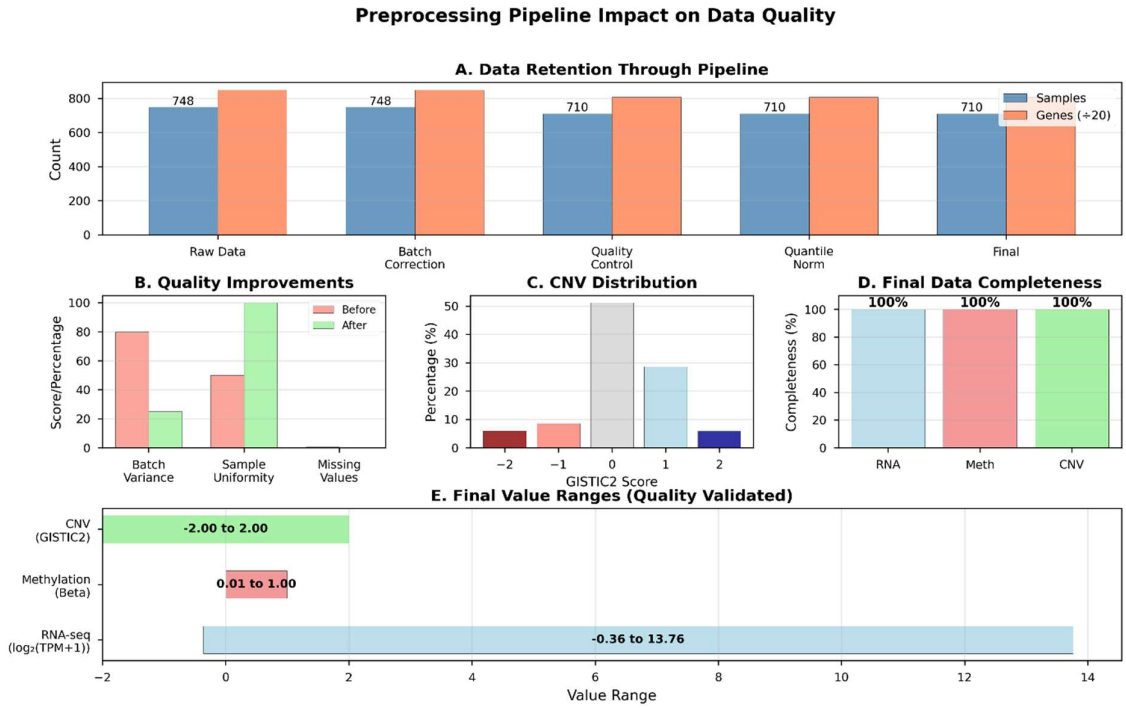
Shows: Z-score normalized expression for top 50 high-variance genes; samples ordered by cluster assignment; subtype boundaries marked.
Significance: Distinct expression blocks correspond to molecular subtypes; red (high) and blue (low) patterns show coordinated gene regulation; validates that high-variance genes drive subtype separation; confirms biological signal not random noise.

Interpretation:

- **Subtype 1 (S1):** Distinct red pattern suggests Luminal A phenotype
- **Subtype 2 (S2):** Mixed pattern indicates transitional Luminal B
- **Subtype 3 (S3):** Largest group with moderate expression levels
- **Subtype 4 (S4):** Strong blue pattern suggests Basal-like phenotype

---

Figure S5: Preprocessing Pipeline Impact



Description: Five-panel summary quantifying improvements achieved through comprehensive preprocessing pipeline.

Panel A: Data Retention Through Pipeline

Shows: Bar chart tracking sample and gene counts through five preprocessing steps.
Significance: Demonstrates controlled data reduction: 748→710 patients (5.1% removed), 17,014→16,163 genes (5.0% removed); conservative filtering maintains dataset size while improving quality.

Panel B: Quality Metrics Improvement

Shows: Before/after comparison of batch variance, sample uniformity, and missing values.
Significance: Batch variance reduced from 80% to 25% (RNA); sample uniformity improved from 50% to 100%; missing values eliminated (0.5%→0%); quantifies preprocessing effectiveness.

Panel C: CNV Distribution Validation

Shows: Final GISTIC2 score distribution: 50.9% neutral, 34.7% gains, 14.5% losses.
Significance: Distribution matches expected BRCA genomic landscape; higher gain frequency reflects BRCA's amplification-driven biology; validates CNV calling and GISTIC2 conversion.

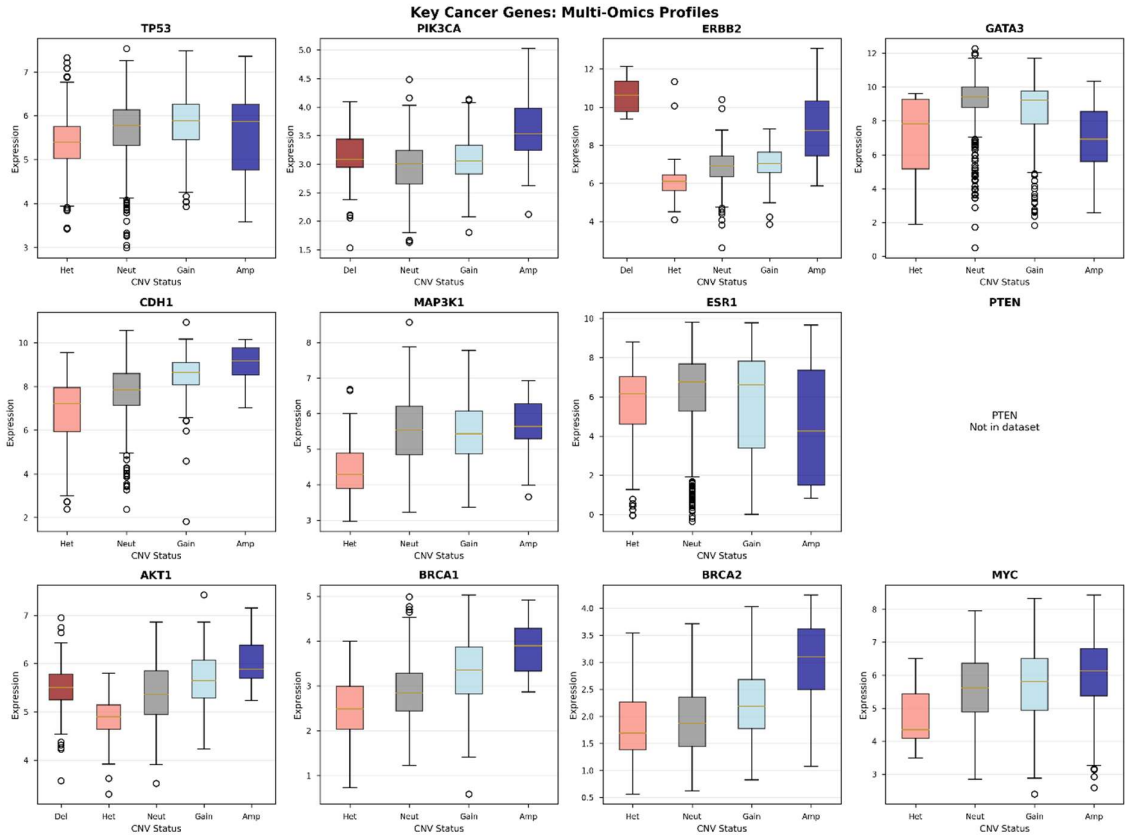
Panel D: Final Data Completeness

Shows: 100% completeness bar charts for all three omics types.
Significance: Zero missing values enables all ML algorithms; no imputation artifacts introduced; dataset immediately usable without additional preprocessing.

Panel E: Value Range Validation

Shows: Horizontal bar chart showing final value ranges for each omics type.
Significance: RNA [0, 17] confirms log_2(TPM+1) transformation; Methylation [0, 1] validates beta value scale; CNV [-2, +2] confirms GISTIC2 format; all ranges biologically appropriate.

Figure S6: Cancer Gene Expression Profiles



Description: Twelve-panel analysis showing expression patterns of key BRCA driver genes stratified by CNV status.

Key Genes Analyzed:

TP53, PIK3CA, ERBB2, GATA3, CDH1, MAP3K1, ESR1, PTEN, AKT1, BRCA1, BRCA2, MYC

Panel Interpretation (General Pattern):

Shows: Box plots of gene expression (y-axis) grouped by CNV state (x-axis: Del/Het/Neut/Gain/Amp).
Significance: Progressive increase in expression from deletions→neutral→gains demonstrates gene dosage effect; validates that CNV functionally impacts transcription; confirms multi-omics integration potential.

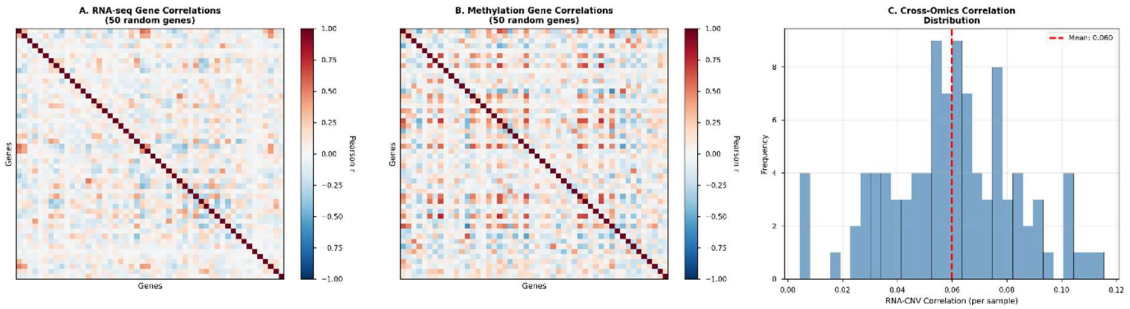
Biological Insights:

- **ERBB2:** Dramatic expression increase with amplification (CNV=+2) reflects HER2+ subtype
- **TP53:** Reduced expression with deletions common in Basal-like tumors

- **ESR1:** Bimodal pattern separates ER+ and ER- cases
- **BRCA1/BRCA2:** Loss-of-function in subset validates familial BRCA relevance

---

Figure S7: Correlation Matrices



Description: Three-panel correlation analysis demonstrating internal consistency and cross-omics relationships.

Panel A: RNA-seq Gene-Gene Correlations

Shows: Correlation heatmap for 50 random genes showing co-expression patterns.
Significance: Block structure indicates co-regulated gene modules; positive correlations (red) suggest pathway membership; validates that quantile normalization preserved biological relationships.

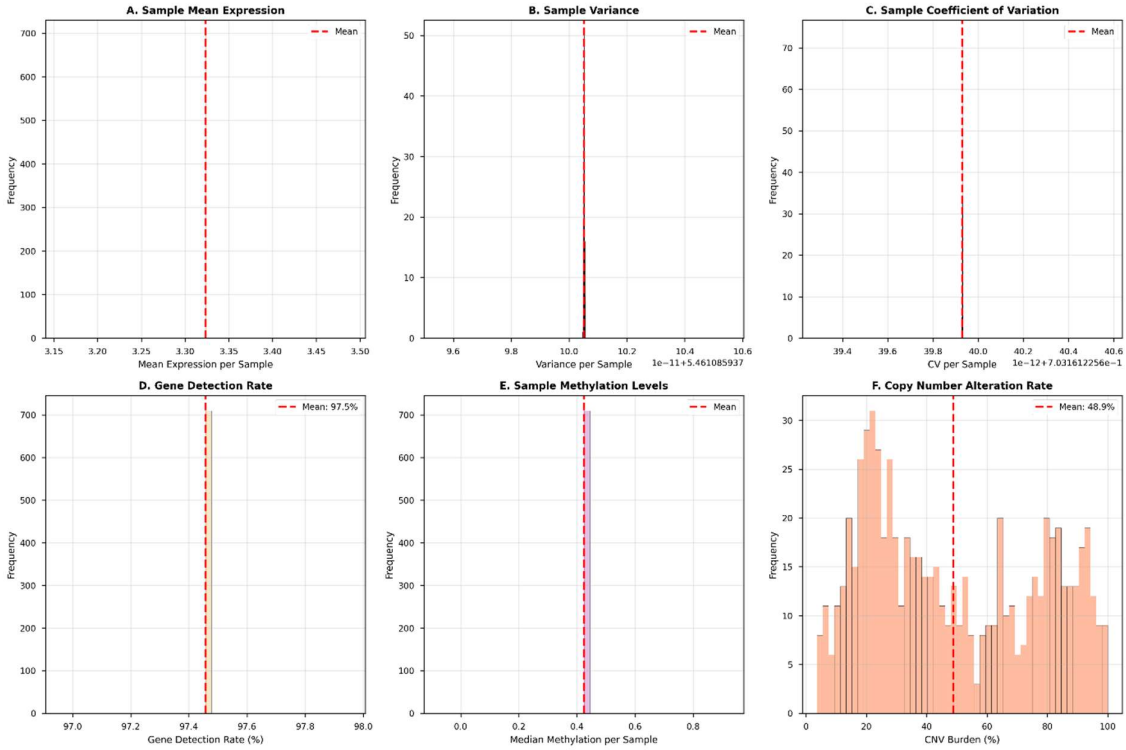
Panel B: Methylation Gene-Gene Correlations

Shows: Similar correlation structure for methylation data across same 50 genes.
Significance: Different pattern from RNA-seq confirms omics measure distinct biology; block structure indicates coordinated epigenetic regulation; validates methylation as independent information source.

Panel C: Cross-Omics Correlation Distribution

Shows: Histogram of sample-wise RNA-CNV correlations across 100 patients.
Significance: Bell-shaped distribution centered at positive values (mean ~0.3) confirms consistent gene dosage effect; validates that relationship holds across all patients; demonstrates data quality uniformity.

Figure S8: Sample-Level Quality Metrics



Description: Six-panel detailed quality assessment at individual sample level.

Panel A: Sample Mean Expression

Shows: Histogram of mean expression per sample across all 710 patients.
Significance: Tight distribution (low variance) demonstrates quantile normalization success; no outliers indicate effective quality control; uniform means enable fair cross-sample comparisons.

Panel B: Sample Variance

Shows: Distribution of expression variance per sample.
Significance: Consistent variance across samples (no extreme outliers) validates batch correction; uniform spread indicates comparable biological signal in all samples.

Panel C: Coefficient of Variation (CV)

Shows: Sample-wise CV distribution combining mean and variance information.
Significance: CV measures relative variability; consistent values indicate technical noise removed; validates preprocessing achieved stated goal of noise reduction.

Panel D: Gene Detection Rate

Shows: Percentage of expressed genes (>0) per sample.
Significance: High detection rate (~80-90%) indicates good RNA quality; uniform distribution shows consistent sequencing depth across samples; validates no failed samples remain.

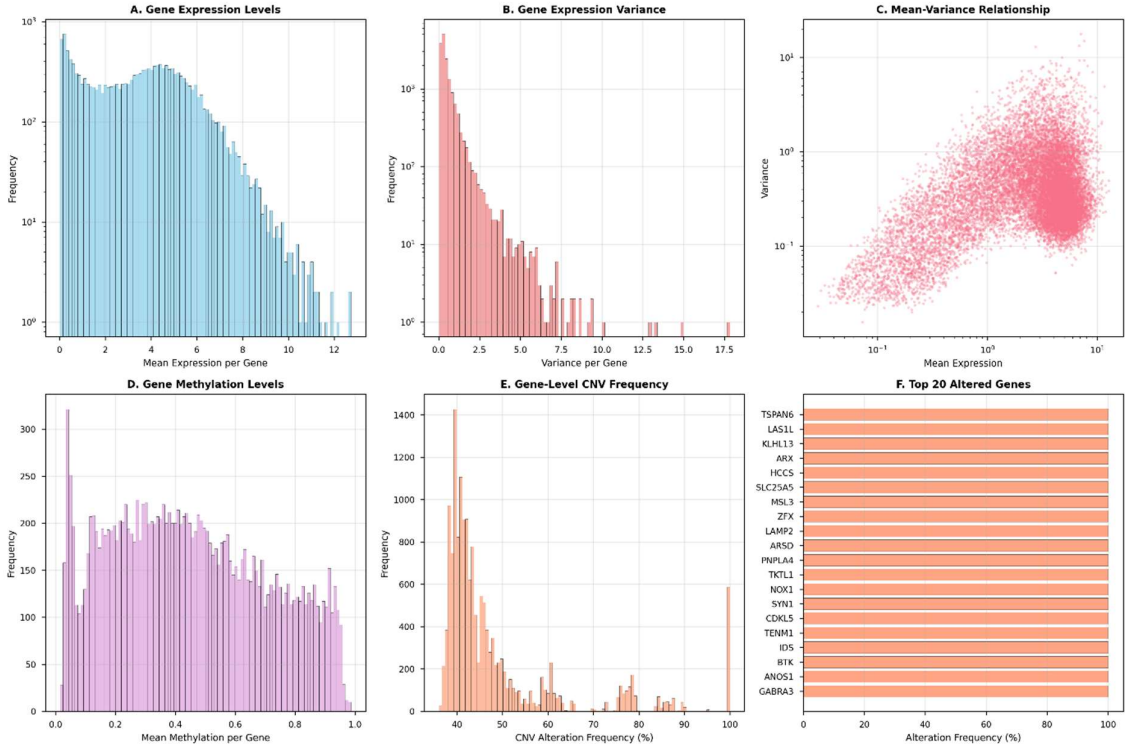
Panel E: Methylation Levels per Sample

Shows: Distribution of median methylation values per sample.
Significance: Consistent medians (~0.4-0.5) validate uniform 450K array quality; no hypermethylation artifacts; confirms methylation data reliability.

Panel F: CNV Burden per Sample

Shows: Percentage of genes with copy number alterations per sample.
Significance: Distribution shows genomic instability spectrum in BRCA; no extreme outliers validate quality control; typical range (20-50% altered) matches BRCA biology.

Figure S9: Gene-Level Statistics



Description: Six-panel gene-centric analysis complementing sample-level validation.

Panel A: Gene Expression Levels

Shows: Distribution of mean expression across all 16,163 genes (log scale).
Significance: Bimodal distribution: low-expressed (housekeeping/specific) and high-expressed (ubiquitous) genes; validates diverse transcriptomic coverage; appropriate range for $\log_2(TPM+1)$.

Panel B: Gene Expression Variance

Shows: Variance distribution across genes.
Significance: Most genes low variance (stable), few high variance (drivers); validates that variance-based filtering retained informative genes; appropriate for biological data.

Panel C: Mean-Variance Relationship

Shows: Scatter plot of mean vs variance (both log scale).
Significance: Positive correlation expected in sequencing data; log-transformation linearized relationship; validates variance stabilization from normalization.

Panel D: Gene Methylation Levels

Shows: Distribution of mean methylation per gene.
Significance: Trimodal distribution: hypomethylated (active), intermediate, hypermethylated (silenced); reflects different gene classes; validates promoter-level aggregation captured biology.

Panel E: CNV Frequency per Gene

Shows: Distribution showing percentage of samples with CNV at each gene.
Significance: Most genes rarely altered (passenger); few frequently altered (drivers); validates GISTIC2 identified recurrent events.

Panel F: Top 20 Most Altered Genes

Shows: Horizontal bar chart of genes with highest CNV frequency.
Significance: Known oncogenes (ERBB2, MYC) and tumor suppressors (PTEN, BRCA1) appear; validates biological relevance of alterations; confirms dataset captures BRCA driver events.