

MULTIPLICITY IN HIGH-DIMENSIONAL DATA

From thousands of findings to a handful of evidence

Technical report

Planimeter Kft.

Version 1.0.0

23 September 2026

Methodological demonstration only. The analysis uses public TCGA-BRCA expression data to illustrate multiplicity and evidence interpretation. It is not a biomarker-validation study and should not be used for clinical decision-making.

Executive summary

Objective. To demonstrate how multiplicity changes the interpretation of high-dimensional analyses and how empirical Bayes variance moderation and effect-size ranking add information beyond a binary significant/not-significant screen.

Data. The stage-based rule retained 441 TCGA-BRCA expression samples representing 391 unique participant IDs and 60,660 transcript features. Early stage was defined as Stage I/IA/IB and late stage as Stage III/IIIA/IIIB/IIIC.

Key result. Conventional feature-wise testing produced 8,538 raw p-values below 0.05. Benjamini-Hochberg adjustment reduced the retained set to 35 features and Bonferroni adjustment to 5. A separate limma empirical Bayes analysis identified 610 FDR-controlled features. Only 6 features appeared in both the top 100 ranked by adjusted significance and the top 100 ranked by absolute effect size.

Interpretation. Error-control choices and uncertainty modelling materially changed the set of retained findings. Statistical evidence and estimated magnitude were related but not interchangeable (correlation between $|\log FC|$ and $-\log_{10}(\text{FDR})$: $r = 0.72$).

1. Data and analytical framework

1.1 Source data

The development workflow used TCGA-BRCA STAR Counts and associated clinical metadata stored in a Bioconductor SummarizedExperiment object. The large raw object is not redistributed. The public package contains compact derived CSV files that reproduce the published summaries and figures.

1.2 Clinical grouping

Samples were classified as Early when AJCC pathologic stage was Stage I, IA or IB, and Late when stage was Stage III, IIIA, IIIB or IIIC. Samples outside these definitions or with missing stage were excluded from this demonstration.

1.3 Reproducibility

The released data files preserve the numerical results of the exploratory development analysis. All code paths in the public scripts are relative. A separate data-preparation script documents the transformation from the local source RDS object, while the main public analysis runs directly from the released CSV files.

Analysis group	Expression samples	Unique participants
Early	197	174
Late	244	217
Total	441	391

1.4 Statistical analysis

- Independent screening: a two-sample t-test was applied feature by feature to the retained raw STAR count matrix. Raw p-values were adjusted using Benjamini-Hochberg false discovery rate control and Bonferroni family-wise error control.
- Empirical Bayes analysis: expression values were transformed as $\log_2(\text{count} + 1)$, a design matrix containing the stage group was fitted with limma, and empirical Bayes moderation was applied to the feature-wise linear-model statistics.
- Ranking comparison: limma features were ranked independently by FDR-adjusted p-value and by absolute log fold change. The top-100 sets were compared by overlap.
- Association of magnitude and statistical evidence: Pearson correlation was calculated between absolute limma log fold change and $-\log_{10}(\text{FDR-adjusted p-value})$.

1.5 Sample structure

The retained metadata contain 441 expression samples but 391 unique TCGA participant identifiers. 48 participants contributed more than one retained expression sample. Barcode-derived sample-type codes comprise 395 primary-tumor samples (01), 2 metastatic samples (06) and 44 solid-tissue-normal samples (11). The development analysis did not model this within-participant clustering or restrict sample type.

2. Results

2.1 Multiplicity changes the evidence set

Stage of inference	Retained features	% of 60,660
Raw $p < 0.05$	8,538	14.1%
BH/FDR adjusted $p < 0.05$	35	0.058%
Bonferroni adjusted $p < 0.05$	5	0.008%

The same 60,660-feature search space produced substantially different retained sets depending on the error criterion. The comparison illustrates why the number of nominally significant features cannot by itself be interpreted as the size of a credible evidence set.

2.2 Empirical Bayes moderation

The limma empirical Bayes analysis retained 610 features at FDR-adjusted $p < 0.05$, compared with 35 retained after BH adjustment of the independent feature-wise t-tests. Empirical Bayes moderation borrows information across features when estimating variability, which can materially alter precision in high-dimensional settings.

This comparison is descriptive rather than a controlled method benchmark: the two analyses differ not only in variance modelling but also in analysis scale, because the independent tests used raw STAR counts whereas limma used $\log_2(\text{count} + 1)$ values.

2.3 Statistical evidence versus effect magnitude

Only 6 features were shared by the top 100 ranked by limma FDR-adjusted p-value and the top 100 ranked by absolute log fold change. Across the full finite result set, absolute log fold change and $-\log_{10}(\text{FDR})$ had correlation $r = 0.717$.

The result illustrates that statistical evidence and effect magnitude answer different questions. A strong ranking by p-value reflects a combination of estimated magnitude and precision; a large estimated effect can rank differently when uncertainty is also considered.

3. Interpretation

The central lesson is not that one correction or model is universally preferable. Rather, high-dimensional inference is conditional on the error criterion, the variance model, the data transformation and the ranking target. Those choices define what is retained as evidence. Reporting them explicitly is therefore part of the scientific result, not merely an implementation detail.

The example also shows why a single list of “significant genes” can be an incomplete endpoint. Multiplicity control addresses the number of opportunities for false positives, empirical Bayes methods can improve variance estimation, and effect-size ranking adds a separate view of magnitude. These dimensions should be interpreted together.

4. Limitations

- The early-versus-late stage contrast was selected as a methodological demonstration and is not a prespecified biomarker-validation study.
- The 441 retained rows are expression samples, not 441 independent patients: they represent 391 unique participant IDs, and repeated samples were not modelled as correlated observations.
- The retained data include multiple TCGA sample types (395 code 01, 2 code 06 and 44 code 11). No sample-type restriction was applied in the original exploratory analysis.
- The independent t-tests operate on raw STAR counts. The limma analysis uses $\log_2(\text{count} + 1)$ without a count-specific normalization/voom workflow. These choices are retained for reproducibility, not recommended as a general RNA-seq differential-expression pipeline.
- The difference between 35 independent-test FDR discoveries and 610 limma discoveries should not be interpreted as a pure measure of the benefit of empirical Bayes moderation because the two workflows differ in more than one modelling choice.

- An FDR-adjusted discovery in this example is a statistical result within this analysis and should not be interpreted as a validated biological or clinical biomarker.

5. Reproducibility package

The public package contains four compact CSV files, a complete R analysis script, a separate data-preparation script, validated output tables, a data dictionary and the GPL license text. The large raw TCGA-BRCA RDS object is not redistributed. All paths in the released code are relative.

File	Purpose
sample_metadata.csv	Stage groups and retained TCGA sample identifiers.
independent_testing_results.csv	Raw, BH and Bonferroni results plus log-scale mean differences.
limma_results.csv	Empirical Bayes feature-level results.
multiplicity_summary.csv	Published key numerical results.
multiplicity_analysis_planimeter_v1.0.0.R	Primary reproducible analysis and five figures.
multiplicity_data_preparation_planimeter_v1.0.0.R	Transformation from the local source RDS object to public CSVs.
README.md / DATA_DICTIONARY.md	Execution notes, provenance, limitations and field definitions.

6. License and AI use

The Planimeter R source code is released under the GNU General Public License v3.0 or later (GPL-3.0-or-later). Generative AI was used as a support tool during drafting, code development, review and validation. The analytical design, interpretation, final materials and responsibility for the release remain with Planimeter Kft.

Sources and methodological references

NCI Genomic Data Commons (GDC). TCGA-BRCA data are distributed through the GDC Data Portal: <https://portal.gdc.cancer.gov/>

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B. 1995;57(1):289-300.

Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47. doi:10.1093/nar/gkv007