*Bioinformatics Advances*, 2026, **6**(1), vbag128
https://doi.org/10.1093/bioadv/vbag128
Published: 18 May 2026
**Application Note | Software**

OXFORD

# Differentially Expressed Gene Annotator (DEGAn): automated annotation and analysis of DEGs datasets with OS and PFS data

**Giuseppe Agapito**[1,2,3,4,*], **and Mario Cannataro**[2,5]

[1]Department of Law, Economics and Sociology, University Magna Græcia of Catanzaro, Catanzaro, 88100, Italy
[2]Data Analytics Research Center, University Magna Græcia of Catanzaro, Catanzaro, 88100, Italy
[3]Laboratorio di Storia Giuridica ed Economica Research Center, University Magna Græcia of Catanzaro, Catanzaro, 88100, Italy
[4]Cultura Romana del Diritto e Sistemi Giuridici Contemporanei Research Center, University Magna Græcia of Catanzaro, Catanzaro, 88100, Italy
[5]Department of Medical and Surgical Sciences, University Magna Græcia of Catanzaro, Catanzaro, 88100, Italy
*Corresponding author. Department of Law, Economics and Sociology, University Magna Græcia of Catanzaro, Viale Europa, Catanzaro, 88100, Italy.
E-mail: agapito@unicz.it.
Associate Editor: Lina Ma

## Abstract

***Motivation:*** **High-throughput transcriptomic platforms routinely generate large differentially expressed gene (DEG) datasets, but linking these matrices to clinical outcomes still often requires manual data integration and repeated gene-by-gene survival analyses. This limits scalability, reproducibility, and practical use in translational research.**

***Results:*** **We present DEGAn (Differentially Expressed Gene Annotator), a Java-based application for automated survival analysis of DEG matrices annotated with clinical data. DEGAn integrates gene expression with overall survival (OS) and progression-free survival (PFS), performs Kaplan–Meier and log-rank analyses across all genes in a single run, and returns ranked results with false discovery rate correction. The updated version extends DEGAn with configurable stratification strategies, alternative missing-value handling, multi-group survival analysis, and enhanced reporting to support sensitivity assessment and reproducible exploratory screening. DEGAn provides a graphical user interface and is distributed as a standalone executable for local, privacy-preserving analysis.**

***Availability and implementation:*** **DEGAn is implemented in Java and released under the LGPL v3.0+ license. Source code and binaries are available at https://gitlab.com/giuseppeagapito/degan.git.**

## Introduction

RNA-seq, next-generation sequencing, and microarray platforms generate large-scale gene expression data that are routinely summarized as differentially expressed gene (DEG) matrices (Heller 2002, Marguerat and Bähler 2010, Wang and Karamyshev 2020). Although these data are widely used for biomarker discovery, their clinical interpretation remains limited when survival endpoints and other annotations must be integrated manually and analyzed one gene at a time (Brazma and Vilo 2000, Swanton et al. 2010). Such workflows are time-consuming, difficult to reproduce, and poorly suited to transcriptome-wide screening.

Kaplan–Meier survival analysis and log-rank testing are widely available in commercial, open-source, and web-based tools. However, most existing solutions are oriented toward single-gene or small-scale analyses, require substantial preprocessing, or depend on predefined public datasets (Chandrashekar et al. 2017, 2022, Posta and Győrffy 2025). As a consequence, systematic survival screening across user-defined DEG matrices remains cumbersome.

To address this limitation, we developed DEGAn (Differentially Expressed Gene Annotator), a standalone Java application for automated annotation and survival analysis of DEG datasets with OS and PFS information. DEGAn performs matrix-level survival analysis in a single execution, ranks genes according to statistical significance with false discovery rate adjustment, and supports local execution to reduce privacy risks associated with sensitive genomic and clinical data (Agapito and Cannataro 2023, Agapito et al. 2023).

## Background

Kaplan–Meier survival analysis is widely used to relate gene expression to outcomes such as overall survival (OS) and progression-free survival (PFS). Although commercial software,

**Received:** 17 November 2025. **Revised:** 7 April 2026. **Accepted:** 30 April 2026

open-source statistical environments, and web-based portals support survival analysis, most existing solutions are geared toward single-gene or small-scale studies, often requiring manual preprocessing, repeated analyses, or programming expertise. These limitations reduce scalability, automation, and reproducibility when high-dimensional DEG matrices must be screened systematically.

To address this gap, we developed *DEGAn*, a standalone framework for automated survival analysis of annotated DEG matrices. DEGAn performs matrix-level analysis in a single run, integrating data harmonization, missing-value handling, expression stratification, and false discovery rate correction to produce ranked survival-associated genes. A broader comparison with existing tools and their limitations is provided in the Supplementary Material.

# Implementation

DEGAn is a platform-independent Java application that performs automated survival analysis of annotated DEG matrices. It integrates gene-expression and clinical data, performs one-click analysis of overall survival (OS) and progression-free survival (PFS), and provides interactive visualization of Kaplan–Meier curves for prioritized genes.

For each gene, DEGAn estimates survival curves and evaluates differences among expression-defined groups using Kaplan–Meier analysis and log-rank testing. DEGAn supports multiple stratification strategies, including median split, tertile-based grouping, quartile-based grouping, and manually defined thresholds. Different stratification strategies may capture distinct biological patterns of gene-expression effects on prognosis. Median dichotomization is widely used for its simplicity and interpretability, but it may miss non-linear effects or patterns visible only at expression extremes. Tertile- and quartile-based stratification can reveal more graded survival trends and better reflect patient heterogeneity, whereas manual thresholds may be preferable when prior biological or clinical knowledge suggests specific cutoffs. For missing data, mean imputation is a simple and computationally efficient default, but it has limitations: it reduces variance, may weaken biologically meaningful differences among samples, and can affect assignment to expression-based groups. In survival analysis, this may influence Kaplan–Meier separation and log-rank significance. For this reason, DEGAn also supports KNN-based imputation and complete-case analysis. Median-based grouping and simple imputation remain the defaults because they provide an accessible, fast, and easily interpretable baseline workflow, while more advanced methods are available when justified by data structure and missingness patterns.

Algorithm 1 summarizes the same configurable workflow described in the Supplementary Material, here reported in compact form.To improve scalability, DEGAn processes genes independently in parallel on multi-core architectures. Results are ranked after false discovery rate correction, and significant genes can be explored interactively through survival-curve visualization. Additional implementation details are reported in the Supplementary Material.

**Algorithm 1 DEGAn workflow for configurable transcriptome-wide survival analysis**

**Require:** DEG matrix $X \in \mathbb{R}^{G \times S}$ (genes × samples); clinical table *100* with survival time $T$ and event indicator $\delta$; stratification mode $M$; imputation mode $I$
**Ensure:** Ranked list of genes with log-rank statistics, adjusted $q$-values, missingness rates, and optional Kaplan–Meier curves
Load DEG matrix $X$ and clinical table $C$
Harmonize sample identifiers and reorder columns of $X$ to match $C$
Validate survival fields: ensure $T \geq 0$ and $\delta \in \{0, 1\}$
Initialize result container $\mathcal{R}$
**for** $g \leftarrow 1$ to $G$ **do**
    Compute missing-data rate $r_g$
    **if** $I =$ mean or median imputation **then**
        Estimate the selected summary statistic over observed values of gene $g$
        Impute missing values of gene $g$
    **else if** $I =$ KNN imputation **then**
        Impute missing values of gene $g$ using KNN
    **else if** $I =$ complete-case analysis **then**
        Remove samples with missing values for gene $g$
    **end if**
**end for**
$P \leftarrow$ number of available CPU cores
Partition gene indices $\{1, \ldots, G\}$ into disjoint blocks
**for all** gene blocks $\mathcal{B}_i$ **in parallel do**
    **for all** $g \in \mathcal{B}_i$ **do**
        Define expression groups for gene $g$ according to stratification mode $M$
        Construct survival bundles linking expression values with $(T_s, \delta_s)$
        Estimate Kaplan–Meier curves for all detected groups
        Compute global log-rank test and raw $P$value $p_g$
        **if** number of groups $> 2$ **then**
            Compute pairwise post-hoc log-rank comparisons
        **end if**
        Store gene identifier, $r_g$, $p_g$, group statistics, and survival summaries in $\mathcal{R}$
    **end for**
**end for**
Apply Benjamini–Hochberg correction to obtain adjusted $q$-values
Rank genes by increasing $q$-values
Provide ranked results, missingness summaries, and Kaplan–Meier curves for visualization

# Performance evaluation

We evaluated DEGAn on synthetic DEG datasets with increasing dimensionality to characterize scalability and computational efficiency. Two matrix sizes were considered, corresponding to 20 000 and 100 000 genes, while the number of samples was progressively increased from 100 to 1000. Across all configurations, execution time increased smoothly with cohort size and remained only moderately affected by the five-fold increase in the number

of genes. Energy consumption followed a similar trend, increasing approximately linearly with the number of samples. Memory usage remained bounded across all tested settings, indicating that DEGAn avoids large intermediate in-memory structures and is suitable for execution on standard workstations.

Overall, these results show that DEGAn scales efficiently with the width of the dataset while remaining robust to increases in feature dimensionality, supporting its use for transcriptome-wide exploratory survival analysis.

Extended validation results, detailed methods, algorithmic workflow, complexity analysis, and additional figures and tables are reported in the Supplementary Material.

# Algorithm complexity analysis

Let $G$ denote the number of genes, $S$ the number of samples, and $C$ the number of available CPU cores. The preprocessing phase depends on the selected missing-data handling strategy. Simple mean/median imputation requires a linear scan of each gene and therefore costs $O(GS)$ overall, whereas complete-case analysis also requires $O(GS)$ time to identify and remove missing entries. KNN-based imputation is more computationally demanding; in the worst case, its cost depends on the number of incomplete samples and nearest-neighbor searches, leading to a higher preprocessing overhead than summary-statistic imputation.

For each gene, stratification requires at most a linear scan over $S$ samples once the required cutoff values have been determined. Median-based grouping has linear cost, while tertile- and quartile-based grouping require quantile computation, which can be performed in $O(S \log S)$ time if based on sorting. Manual-threshold grouping remains linear in $S$. Survival analysis then requires construction of the expression-defined groups, estimation of Kaplan–Meier curves, and computation of the log-rank statistic. This step has per-gene complexity $O(S \log S)$ because event times must be ordered.

Accordingly, under the default workflow based on summary-statistic imputation and median stratification, the overall complexity remains $O(GS \log S)$. When quantile-based stratification is selected, the asymptotic cost is unchanged, since sorting dominates. When KNN imputation is used, total runtime increases by an additional preprocessing term reflecting nearest-neighbor computation, but the downstream survival-analysis stage remains unchanged.

If more than two expression-defined groups are generated, DEGAn first computes a global log-rank test and then performs pairwise post-hoc comparisons. For $K$ groups, this introduces up to $O(K^2)$ pairwise tests per gene; however, since $K$ is small in practice (e.g., $K = 3$ for tertiles and $K = 4$ for quartiles), this does not change the dominant asymptotic term.

Since genes are independent, DEGAn parallelizes computation across $C' \approx C - 1$ workers, leading to an expected wall-clock time of $O\left(\frac{G}{C'} S \log S\right)$ under the default workflow, with modest synchronization overhead. Memory usage is dominated by storage of the input expression matrix and therefore scales as $O(GS)$. Additional per-gene statistics, missingness information, and survival summaries are maintained using compact shared data structures, yielding stable memory consumption across increasing feature dimensionalities.

# Validation and discussion

DEGAn was validated on the public GEO dataset GSE30219, which contains gene expression profiles and overall survival information for lung cancer patients (Rousseaux et al. 2013). DEGAn directly processed the full dataset and produced a ranked list of genes associated with survival. Table 1 lists the first three top-ranked probes identified by DEGAn in the GSE30219 validation dataset. For validation, representative top-ranked probes were exported into an SPSS-compatible format and reanalyzed using Kaplan–Meier estimation and log-rank testing. The resulting survival curves and test statistics were consistent with those obtained using DEGAn, confirming the correctness of the implementation.

The current version improves methodological flexibility while preserving automation. Configurable stratification and imputation options enable users to assess the impact of analytical choices on survival results, while multi-group analysis broadens the range of supported study designs. At the same time, local execution and one-click matrix-level processing remain key

**Table 1** Top probes identified by DEGAn in the GSE30219 validation dataset.

| Probe ID | Gene | *p*-value |
|---|---|---|
| 15,53,015_a_at | RECQL4 | $1.0037 \times 10^{-3}$ |
| 2,28,775_at | EMC3 | $1.1620 \times 10^{-3}$ |
| 15,52,680_a_at | KNL1 | $1.4874 \times 10^{-3}$ |

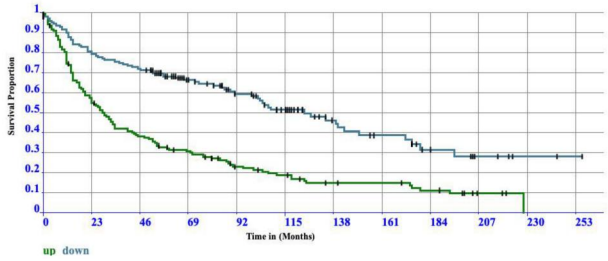


**Figure 1** Overall survival curve generated by DEGAn for the probe set *15,52,680_a_at*.

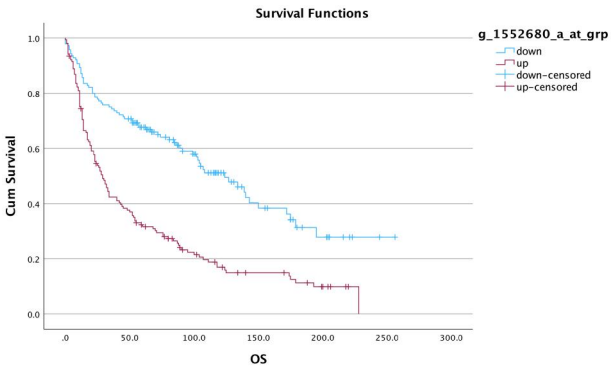


**Figure 2** Corresponding overall survival curve generated by SPSS for the probe set *15,52,680_a_at*.

practical advantages over manual gene-by-gene workflows and web-based systems restricted to predefined cohorts.

In the analyzed lung cancer dataset, DEGAn prioritized probes associated with genes such as *RECQL4*, *EMC3*, and *KNL1*, supporting its ability to highlight biologically and clinically meaningful candidates for downstream investigation (Wang et al. 2024, Tang et al. 2024, Li et al. 2026).

The close agreement between Figures 1 and 2 further confirms the consistency of DEGAn with SPSS statistical platform, while highlighting the practical advantage of transcriptome-wide automated analysis in a single executable workflow. Extended analyses, validation results, and additional figures and tables are provided in the Supplementary Material.

## Conclusions

DEGAn provides an automated framework for integrating DEG matrices with clinical survival data and performing transcriptome-wide OS and PFS analysis in a single workflow. By replacing repetitive gene-by-gene procedures with matrix-level processing, DEGAn improves scalability, reproducibility, and usability in translational research settings. The updated implementation further strengthens the software by introducing configurable stratification, alternative imputation strategies, multi-group survival analysis, and richer reporting for sensitivity assessment. DEGAn therefore represents a practical and extensible tool for survival-oriented analysis of transcriptomic data.

## Author contributions

Giuseppe Agapito [Conceptualization (Lead), Software (Lead), Validation (Equal), Writing Original Draft (Equal), Writing Review Editing (Equal)] and Mario Cannataro [Validation (Equal), Writing Original Draft (Equal), Writing Review Editing (Equal)]

## Supplementary material

Supplementary material is available at *Bioinformatics Advances* online.

## Conflicts of interest

None declared.

## Funding

This work was supported by Next Generation EU—Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of “Innovation Ecosystems,” building “Territorial R&D Leaders” (Directorial Decree n. 2021/3277), project Tech4You—Technologies for climate change adaptation and quality of life improvement [ECS0000009]; and by PR CALABRIA FESR-FSE 2021–2027, Priorità 1—Una Calabria più competitiva e intelligente, Azione 1.1.1, project HeartTrack [CUP J59I24002660005].

## Data availability

The data underlying this article are available in the Gene Expression Omnibus database at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi? acc=GSE30219 under accession number GSE30219.

## References

Agapito G, Cannataro M. An overview on the challenges and limitations using cloud computing in healthcare corporations. *BDCC* 2023;**7**:68. https://doi.org/10.3390/bdcc7020068. https://www.mdpi.com/2504-2289/7/2/68.

Agapito G, Cannataro M, Cinaglia P *et al.* Use predictive learning model to tackle data breaches in healthcare domain. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp.3273–3279. IEEE, 2023.

Brazma A, Vilo J. Gene expression data analysis. *FEBS Lett* 2000; **480**:17–24.

Chandrashekar DS, Bashel B, Balasubramanya SAH *et al.* UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. *Neoplasia* 2017;**19**:649–58.

Chandrashekar DS, Karthikeyan SK, Korla PK *et al.* UALCAN: an update to the integrated cancer data analysis platform. *Neoplasia* 2022;**25**:18–27.

Heller MJ. Dna microarray technology: devices, systems, and applications. *Annu Rev Biomed Eng* 2002;**4**:129–53.

Li R, Yu W, Wang D *et al.* RECQL4 promotes the malignant progression of lung adenocarcinoma through the YBX1/G3BP1-mediated NF-kB signaling pathway. *Cell Death Discov* 2026;**12**: 8. https://doi.org/10.1038/s41420-025-02849-3.

Marguerat S, Bähler J. RNA-seq: from technology to biology. *Cell Mol Life Sci* 2010;**67**:569–79.

Posta M, Győrffy B. Pathway-level mutational signatures predict breast cancer outcomes and reveal therapeutic targets. *Br J Pharmacol* 2025;**182**:5734–47. doi: https://doi.org/10.1111/bph.70215. https://bpspubs.onlinelibrary.wiley.com/doi/abs/10.1111/bph.70215.

Rousseaux S, Debernardi A, Jacquiau B *et al.* Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med* 2013;**5**: 186ra66. https://doi.org/10.1126/scitranslmed.3005723. https://www.science.org/doi/abs/10.1126/scitranslmed.3005723.

Swanton C, Larkin JM, Gerlinger M *et al.* Predictive biomarker discovery through the parallel integration of clinical trial and functional genomics datasets. *Genome Med* 2010;**2**:53.

Tang X, Wei W, Sun Y *et al.* EMC3 regulates trafficking and pulmonary toxicity of the SFTPC$^{I73T}$ mutation associated with interstitial lung disease. *J Clin Invest* 2024;**134**.

Wang D, Karamyshev AL. Next generation sequencing (NGS) application in multiparameter gene expression analysis. In *Molecular Toxicology Protocols*. New York, NY: Humana, 2020, 17–34.

Wang Y, Li C, Jiang T. A comprehensive exploration of twist1 to identify a biomarker for tumor immunity and prognosis in pan-cancer. *Medicine* 2024;**103**:e37790. https://doi.org/10.1097/MD.0000000000037790

Bioinformatics Advances, 2026, 6, 1–4
https://doi.org/10.1093/bioadv/vbag128
Application Note