Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

BMEntropy v1.0: Entropy-Based Analysis of Gene Expression in Multiple Myeloma Plasma Cells

Elham Khalesi

khalesi.elham53@gmail.com

Iran National Standards Organization

Research Article

Keywords: Bone marrow, multiple myeloma, gene expression, Shannon entropy, Rényi entropy, Tsallis entropy, computational biology

Posted Date: October 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11159712/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

This study presents BMEntropy v1.0, an exploratory workflow for summarizing expression-derived entropy in multiple-myeloma plasma-cell samples. The supplied analysis uses 500 GSE7039 samples and a processed matrix of 10,638 genes by 500 samples. Shannon, Rényi, and Tsallis summaries were calculated; principal-component analysis (PCA) and K-means clustering were then applied to standardized entropy measures. Mean Shannon entropy was 12.214255 (SD 0.038184), and three clusters contained 224, 250, and 26 samples. A separate exploratory procedure produced a 97-gene candidate set from an intersection of ranked statistical and Random Forest lists, followed by name cleaning. The results describe computational patterns within this dataset; they do not demonstrate independent replication, causal biological meaning, diagnostic accuracy, or clinical utility. The candidate genes should therefore be treated as hypotheses for follow-up, not as a validated biomarker panel.

Introduction

Recent advances in bone marrow computational pathology and hematologic profiling have increasingly relied on information-theoretic metrics. Radiomic and texture entropy extracted from [18F]FDG PET/CT scans and whole-slide histopathology have been studied in multiple myeloma and myeloproliferative neoplasms [1–3]. Related work has explored automated bone marrow cell detection and leukemia transcriptomic stratification [4–6]. Cellular and Shannon diversity entropies have also been used in studies of mesenchymal stem cells and B-cell receptor repertoire maturation [7–10]. These studies span different data types and objectives. Accordingly, this study presents BMEntropy as an exploratory workflow for comparing several entropy summaries in a supplied bulk-expression dataset; it does not claim explainability, diagnostic performance, or clinical validation.

Entropy can summarize uncertainty or dispersion in a specified representation of molecular data, but its interpretation depends on how the quantity is defined. For example, Teschendorff and Enver's single-cell framework estimates signalling entropy by combining transcriptomes with an interaction network to model cellular signalling and differentiation potency [11]. BMEntropy instead computes summary entropies from a processed gene-expression matrix; it does not implement that network-based signalling-entropy framework. This distinction matters: similar terminology does not make the biological interpretation interchangeable. Here, we document the outputs of the supplied workflow, assess the structure of its entropy summaries, and report the resulting exploratory gene list while avoiding claims of disease classification or clinical validation.

Materials and Methods

The steps below reflect the supplied notebook and the scope and limitations documented in the accompanying analysis; they are not an independent reanalysis.

The BMEntropy v1.0 analysis was implemented using open-source scientific Python packages [12], applying Shannon, Rényi, and Tsallis entropy definitions [13–14] and drawing conceptual distinction from

network-based transcriptional entropy modeling [11].

Dataset and preprocessing

The supplied notebook analyzes 500 patient sample files associated with GSE7039 and platform GPL4819. Probe identifiers were mapped to genes, non-gene entries were removed, and the highest-variance probe was retained for each gene. The resulting expression matrix contained 10,638 genes and 500 samples. The notebook then applied $\log_2(x+1)$ transformation and custom quantile normalization. The supplied metadata file contains 500 unique sample identifiers.

Entropy summaries and clustering

The notebook computed Shannon entropy, Rényi entropy for α = 0.5, 2, and 3, and Tsallis entropy for q = 0.5, 1.5, and 2. Samples were divided into equal-sized high- and low-Shannon groups using the median. The seven entropy measures were standardized for PCA and K-means clustering. PCA retained two components for summary, and K-means used k = 3, random_state = 42, and n_init = 20 [12]. These analytical choices describe the supplied run; no independent replication or stability analysis is documented.

Gene-list construction and statistical comparisons

The notebook compared entropy measures across the derived clusters using ANOVA and Kruskal–Wallis tests. Because clusters were formed from entropy features and those same features were subsequently compared across clusters, these tests are descriptive of internal separation, not independent tests of predictive performance. For gene ranking, the supplied method defines Signature_Score as $-\log_{10}(\text{FDR})$, with zero FDR values replaced by 10^{-300}, multiplied by Effect_Size. It intersects the top 100 statistical and top 100 Random Forest features, then cleans gene names to yield 97 entries. This procedure is exploratory and is not a prospectively specified or externally validated biomarker-selection protocol.

Workflow illustration

Figure 1 provides a schematic overview of the analysis sequence. It is an explanatory illustration, not an output generated by reanalysis of the expression data.

Results

Mean Shannon entropy across the 500 samples was 12.214255 (SD 0.038184; minimum 12.156888; median 12.207682; maximum 12.254845). The median split produced 250 high-entropy and 250 low-entropy samples. Pairwise correlations among the seven entropy measures were very high

(approximately 0.998880–0.999999), indicating that the summaries were strongly redundant in this dataset [13].

The first two PCA components accounted for 99.9718% and 0.02765% of variance, respectively. K-means produced clusters of 224, 250, and 26 samples. The supplied notebook also reports significant between-cluster tests for entropy measures; these results are expected to reflect, at least in part, the use of those measures to construct the clusters. They should not be interpreted as independent evidence of biological subtypes.

The top-list intersection and name-cleaning procedure yielded 97 candidate genes. These are computationally selected candidates. The supplied files do not establish pathway enrichment, a validated biological panel, external-cohort replication, or association with clinical endpoints. See Supplementary Figure S3 (Fig. 3), placed immediately after this paragraph..

Entropy measures correlation heatmap.

Figure 2. Pairwise correlation heatmap of the seven entropy summaries computed for the 500 samples: Shannon entropy, Rényi entropy (α = 0.5, 2, and 3), and Tsallis entropy (q = 0.5, 1.5, and 2). The heatmap summarizes the strong pairwise associations among the measures; exact coefficients are reported in the analysis output.

Discussion

Mathematical Convergence and Metric Redundancy

A principal finding of this evaluation is the remarkably high mutual correlation (r > 0.99) observed across the seven evaluated entropy summaries: Shannon entropy, Rényi entropy at α = 0.5, 2, and 3, and Tsallis entropy at q = 0.5, 1.5, and 2. These closely related summaries captured nearly the same dominant axis of variation in the GSE7039 normalized bone-marrow microarray dataset. Although Shannon, Rényi, and Tsallis entropies differ mathematically in how they weight the expression distribution, their empirical convergence here indicates limited metric-specific separation in this bulk dataset. Consequently, PCA and K-means clustering based on the standardized summaries may provide largely redundant representations rather than distinct biological dimensions. Future analyses should test whether this pattern persists in independent cohorts and single-cell datasets, where cell-to-cell heterogeneity and zero inflation could produce more differentiated metric behavior [12–14].

Contextualization within Bone Marrow Neoplasm and Entropy Literature

The application of entropy-based analytics to bone marrow pathology has historically spanned distinct analytical modalities. In spatial and diagnostic imaging, entropy serves as a dominant feature: recent

radiomic profiling in multiple myeloma utilizes PET/CT texture and segmentation entropy to quantify spatial heterogeneity and evaluate measurable residual disease (MRD) [1–2]. Parallel advancements in digital hematopathology have implemented entropy-guided visual transformers and improved object detection (such as YOLOv7 variants) to systematically segment and classify bone marrow aspirate cytomorphology [4]. Moreover, whole-slide imaging of myeloproliferative neoplasms (MPN) has demonstrated the clinical utility of localized tissue entropy in flagging early architectural distortions [3]. In fluid-phase diagnostics, optimization models leveraging Shannon and Tsallis entropies have enhanced the discrimination of leukemic blasts in peripheral blood smears [5], while ensemble super-learners integrating entropy-filtered feature spaces have proven capable of stratifying leukemia molecular subtypes [6].

BMEntropy v1.0 extends these foundational principles directly to the transcriptome domain, establishing a standardized exploratory pipeline for bulk marrow arrays. However, while imaging and cytological entropy measure morphological and spatial chaos [1–5], transcriptomic entropy reflects functional state space dispersion. This distinction is critical when interpreting stem cell maturation dynamics, such as transitioning from fetal to adult mesenchymal stem cell compartments [7], resolving metabolic trade-offs in cell states [8], or tracing chemically directed trans-differentiation trajectories [9].

Distinction from Network-Signaling and Differentiation Frameworks

A vital theoretical boundary must be drawn between BMEntropy's static formulation and dynamical potency estimators such as SCENT (Single-Cell Entropy) [11]. SCENT integrates transcriptomic expression with external protein-protein interaction (PPI) topology to calculate signaling entropy, thereby estimating cellular differentiation potency and developmental plastic states [11]. In contrast, BMEntropy calculates distribution-level information entropy directly from raw or normalized gene vectors without a biological interactome scaffold. Therefore, while BMEntropy provides robust descriptive assessments of transcriptional dispersion, it does not measure signaling flux, biophysical cell-state potency, or regulatory network constraints. Furthermore, because marrow microenvironments involve delicate coordinated events—such as Ezrin-mediated molecular scaffolding during B-cell antigen receptor repertoire diversification [10]—isolated dispersion metrics cannot infer localized molecular mechanisms without multi-omic integration.

Study Limitations and Selection Optimism

Several methodological constraints warrant explicit emphasis:

1. Hypothesis-Generating Gene Signature: The 97-gene panel extracted via variance- and dispersion-driven filtering serves purely as a hypothesis-generating exploratory tool. Without prospective

independent validation across external cohorts (e.g., secondary AML, MDS, or normal bone marrow series), these markers remain susceptible to selection optimism and dataset-specific noise.

2. Bulk Cohort Confounding: The GSE7039 cohort comprises bulk-tissue extracts containing complex mixtures of hematopoietic and stromal cell types. Bulk entropy values inevitably average these distinct cellular proportions, potentially masking microenvironmental shifts that only single-cell deconvolution could resolve.

3. Absence of Direct Clinical Correlates: The baseline pipeline operates on raw dispersion metrics without immediate integration into clinical survival or cytogenetic outcome models. Future benchmarking must evaluate BMEntropy on single-cell RNA-seq datasets (scRNA-seq) to rigorously assess its diagnostic sensitivity relative to established network-entropy tools.

Conclusion

BMEntropy v1.0 provides an exploratory computational summary of expression-derived entropy in 500 GSE7039 samples, with highly correlated entropy measures, three K-means groups, and a 97-gene candidate list. These outputs support methodological description and hypothesis generation only. Independent cohort validation, feature-selection stability, biological interpretation, and evaluation against clinical outcomes are required before diagnostic or biomarker claims can be considered.

Abbreviations

BM: Bone Marrow; MM: Multiple Myeloma; NPC: Normal Plasma Cells; DEGs: Differentially Expressed Genes; GEO: Gene Expression Omnibus; PCA: Principal Component Analysis; t-SNE: t-Distributed Stochastic Neighbor Embedding; UMAP: Uniform Manifold Approximation and Projection; IQR: Interquartile Range; SD: Standard Deviation.

Declarations

Ethics approval and consent to participate

Not applicable. This study utilized publicly available, secondary, de-identified datasets from the NCBI Gene Expression Omnibus (GEO); therefore, formal ethical approval and consent to participate were not applicable. Clinical trial number: not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are publicly available in the NCBI Gene Expression Omnibus (GEO) repository under accession number GSE7039 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7039). To ensure full transparency during peer review, the source code and analytical scripts of BMEntropy v1.0 have been made available for the Editorial Board and peer reviewers as Additional file 1.

Competing interests

The author declares that they have no competing interests.

Funding

The author declares that no funding was received for this study.

Authors' contributions

EK conceived the study, developed the software BMEntropy v1.0, performed data analysis, and wrote the manuscript.

Acknowledgements

Not applicable.

Data availability. This revision is based on the supplied notebook, manuscript, and CSV files. The underlying GSE7039 accession is identified in the analysis, but raw-data reprocessing was not independently performed for this revision. The supplied files do not include an independent validation cohort or clinical outcome dataset.

References

1. Milara E, et al. Bone marrow segmentation and radiomics analysis of [18F]FDG PET/CT in multiple myeloma. Comput Biol Med. 2022; PMID: 36044803.
2. Milara E, et al. Radiomics analysis of bone marrow biopsy locations in [18F]FDG PET/CT images for measurable residual disease assessment in multiple myeloma. Phys Eng Sci Med. 2023. DOI: 10.1007/s13246-023-01265-0.
3. Frontiers in Oncology Study Group. Early myeloproliferative neoplasms (MPN) diagnosis using whole-slide imaging entropy analysis. Front Oncol. 2025; PMID: 41669264.
4. Zhang Y, et al. Improved YOLOv7 algorithm for detecting bone marrow cells (YOLOv7-CTA). Sensors (Basel). 2023;23(17):7640. PMID: 37688095.
5. Ahmad R, et al. White blood cell classification using entropy-controlled optimization in leukemia blood smears. Diagnostics. 2023; PMID: 36766457.

6. Selvaraj S, et al. Super learner model for classifying leukemia through gene expression and entropy-based feature selection. Discov Oncol. 2024; PMID: 39331180.
7. Gençer EB, et al. Transcriptomic and proteomic profiles of fetal versus adult mesenchymal stem cells. Stem Cell Res Ther. 2024;15:102. PMID: 38475970.
8. Lin DW, et al. Inferring metabolic objectives and trade-offs in single cells. Cell Syst. 2025; PMID: 39778581.
9. Weigao E, et al. Dynamics of cell fate decisions during chemically induced trans-differentiation. Adv Sci. 2025; PMID: 40051350.
10. Aysola V, et al. Ezrin promotes antigen receptor diversity during B cell development by supporting IgH chain repertoire diversification. ImmunoHorizons. 2022; PMID: 36241346.
11. Teschendorff AE, Enver T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. Nat Commun. 2017;8:15599. DOI: https://doi.org/10.1038/ncomms15599.
12. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.
13. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948;27(3):379–423. DOI: https://doi.org/10.1109/TIT.2003.809253.
14. Rényi A. On measures of entropy and information. Proc Fourth Berkeley Symp Math Stat Probab. 1961;1:547–561.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx