Check for updates

METHOD ARTICLE

REVISED Hobotnica: exploring molecular signature quality [version 2; peer review: 2 approved]

Alexey Stupnikov[1,2], Alexey Sizykh[1], Anna Budkina[1], Alexander Favorov[3,4], Bahman Afsari[3], Sarah Wheelan[3], Luigi Marchionni[5], Yulia Medvedeva[1,2,6]

[1]Moscow Institute of Physics and Technology, Moscow, Russian Federation
[2]National Medical Research Center for Endocrinology, Moscow, Russian Federation
[3]Johns Hopkins University, Baltimore, USA
[4]Vavilov Institute for General Genetics RAS, Moscow, Russian Federation
[5]Weill Cornell Medicine, New York, USA
[6]Center of Biotechnology RAS, Moscow, Russian Federation

v2 **First published:** 08 Dec 2021, **10**:1260
https://doi.org/10.12688/f1000research.74846.1
Latest published: 16 Aug 2022, **10**:1260
https://doi.org/10.12688/f1000research.74846.2

Abstract

A Molecular Features Set (MFS), is a result of a vast diversity of bioinformatics pipelines. The lack of a "gold standard" for most experimental data modalities makes it difficult to provide valid estimation for a particular MFS's quality. Yet, this goal can partially be achieved by analyzing inner-sample Distance Matrices (DM) and their power to distinguish between phenotypes.
The quality of a DM can be assessed by summarizing its power to quantify the differences of inner-phenotype and outer-phenotype distances. This estimation of the DM quality can be construed as a measure of the MFS's quality.
Here we propose Hobotnica, an approach to estimate MFSs quality by their ability to stratify data, and assign them significance scores, that allow for collating various signatures and comparing their quality for contrasting groups.

Keywords

Molecular signature, Distance Matrix, Differential Gene Expression, Gene Signature, Rank statistics

This article is included in the Bioinformatics gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 16 Aug 2022	✓ view	✓ view
version 1 08 Dec 2021	? view	? view

1. **Roberto Malinverni**, Josep Carreras Leukemia Research Institute (IJC), Badalona, Spain

2. **Shailesh Tripathi**, University of Applied Sciences Upper Austria, Linz, Austria
FH Austria, Steyr, Austria

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Alexey Stupnikov (aleksej.stupnikov@phystech.edu), Yulia Medvedeva (ju.medvedeva@gmail.com)

Author roles: Stupnikov A: Conceptualization, Formal Analysis, Investigation, Methodology, Project Administration, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Sizykh A**: Formal Analysis, Investigation, Software, Visualization; **Budkina A**: Formal Analysis, Investigation, Visualization; **Favorov A**: Conceptualization; **Afsari B**: Conceptualization, Writing – Original Draft Preparation; **Wheelan S**: Conceptualization, Funding Acquisition, Writing – Original Draft Preparation; **Marchionni L**: Conceptualization, Funding Acquisition, Methodology; **Medvedeva Y**: Funding Acquisition, Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Ministry of Science and Higher Education of the Russian Federation (agreement no. 075-15-2020-899) and by the NIH grants R01DE027809 and P30CA006973.

Copyright: © 2022 Stupnikov A *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Stupnikov A, Sizykh A, Budkina A *et al.* **Hobotnica: exploring molecular signature quality [version 2; peer review: 2 approved]** F1000Research 2022, **10**:1260 https://doi.org/10.12688/f1000research.74846.2

First published: 08 Dec 2021, **10**:1260 https://doi.org/10.12688/f1000research.74846.1

REVISED **Amendments from Version 1**

In the new version we have expanded the definition and introduction for our approach.
We have added additional analysis for Methylation data type to illustrate and validate our approach.
New figures were added to demonstrate this.
Several minor editions to improve the manuscript were added.

Any further responses from the reviewers can be found at the end of the article

Introduction

A signature based on a predefined Molecular Features Set (MFS), which is designed to distinguish biological conditions or phenotypes from each other, is a crucial concept in bioinformatics and precision medicine. In this context, signatures typically originate from MFS from contrasting experimental data from two or more sample groups, which differ phenotypically. These MFS incorporate information on the differences between the groups. The nature of the MFS depends on the modality of the original data. For instance, the MFS provided by the Differential Gene Expression approach is a list of Differentially Expressed genes (DEG); Differential Methylation analysis provides Differentially Methylated Cytosines or regions (DMC and DMR) as MFS.

A significant number of mutational, expression and methylation-based signatures have recently been published and they are actively used in research and translational medicine. Examples of expression-based signatures involve gene sets for clinical prognosis (e.g., PAM50,[1] MammaPrint[2]), for pathways and gene enrichment analysis (e.g., MsigDB collections[3]), and for drug re-purposing (e.g., LINCS project[4]).

Direct quality assessment for MFS is currently hardly possible, since there are no 'gold standard' datasets where active Molecular Features are explicitly known. In this manuscript, we propose a novel approach - Hobotnica, that allows for measurement of MFS quality by addressing the key property of the signature, namely, its quality for data stratification.

Hobotnica leverages the quality of distance matrices obtained from any source, in order to assess the quality of the MFS from any data modality compared to a random MFS. In this study, we demonstrate its application to transcriptomic and methylation signatures.

Methods

Approach

The Hobotnica approach is as follows: For a given data set W and a given MFS (S) we derive the inter-sample distance matrix ($DM(S,W)$). Then we assess the quality of DM (and, thus, of S) with a summarizing function ($\alpha(DM(S)) = \alpha(DM(S),Y)$ or by abuse of notation $\alpha(DM(S))$) where (Y) represents the labels of samples. In shorter notation,

$$\begin{aligned} H &: S \\ f(S|D) &\rightarrow DM \\ g(DM|Y) &\rightarrow \alpha \end{aligned} \tag{1}$$

We desire the function α to gauge if the inner-class samples are closer to each other than to outer-class samples. If no difference exists from one class to another, α must be close to zero and as the difference grows, α grows. In the ideal case of a perfect separation, α reaches its maximum at 1:

- $\alpha \in [0,1]$
- $\alpha \rightarrow 1 \Leftrightarrow$ High groups stratification quality
- $\alpha \rightarrow 0 \Leftrightarrow$ Low groups stratification quality

Under the null hypothesis of Hobotnica ((H_0)), no significant difference exists between $\alpha(S)$ and the α of an equal-sized general random set. On the contrary, the alternative (H_A) hypothesizes that S generates higher α than most random S' of the same size. To estimate a null distribution for Hobotnica's α, we applied a permutation test. As our default options, we use Kendall distance as the distance measure and Mann-Whitney-Wilcoxon test as the summarizing function.

When instead of a single *MFS* a set of hypotheses $\{H_1 : MFS_1, H_2 : MFS_2, \ldots, H_n : MFS_n\}$ is in place, for each Molecular Feature Set MFS_i corresponding Distance Matrix DM_i can be generated, and than, in turn, particular value of the measure α_i:

$$\begin{cases} H_1 : MFS_1 \\ H_2 : MFS_2 \\ \ldots \\ H_n : MFS_n \end{cases} \rightarrow \begin{cases} f(MFS_1|D) \rightarrow DM_1 \\ f(MFS_2|D) \rightarrow DM_2 \\ \ldots \\ (MFS_n|D) \rightarrow DM_n \end{cases} \rightarrow \begin{cases} g(DM_1|A) \rightarrow \alpha_1 \\ g(DM_2|A) \rightarrow \alpha_2 \\ \ldots \\ g(DM_n|A) \rightarrow \alpha_n \end{cases}. \quad (2)$$

Thus, for every MFS MFS_i from set of hypotheses $\{H_1 : MFS_1, H_2 : MFS_2, \ldots, H_n : MFS_n\}$ H-score α_i may be computed, resulting in a set $\langle \alpha_1, \alpha_2, \ldots \alpha_n, \rangle$. Comparing α values allows for corresponding Feature Sets qualities ranking and selecting the most informative Signatures for the Data *D*.

Model

For two groups of samples, *R* and *G*, taking into account the nature of Distance Matrix (0-diagonal and symmetry), the corresponding lower triangle matrix may be considered (Figure 1A). The distance values of this matrix d_{pq} can be ranked (Figure 1B), producing new matrix with natural values r_{pq} (Figure 1C).

Subseting the matrix to values corresponding to in-class distances (Figure 1D), we can compute sum of these ranks Σ.

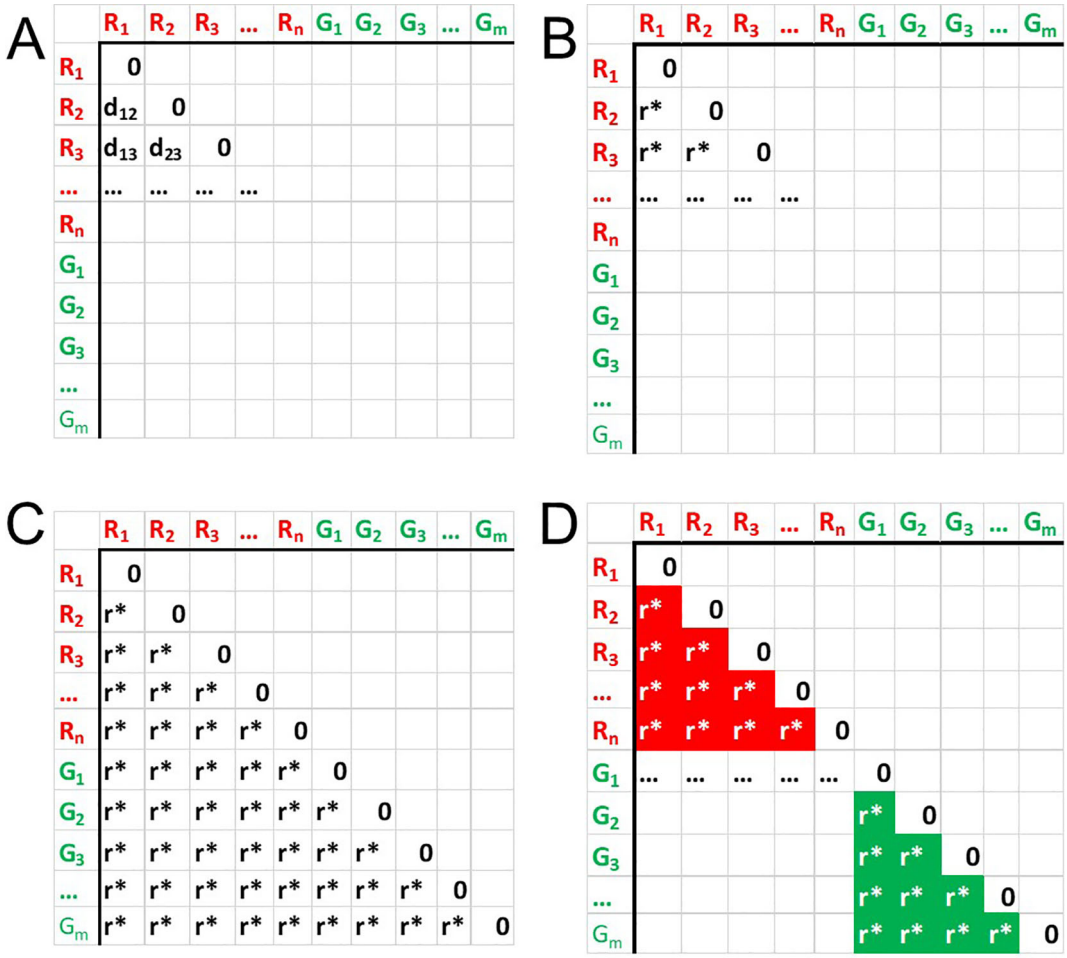

Figure 1. Distances ranking procedure A: Distance matrix structure. B: each distance element d_{pq} is substituted with corresponding rank value. C: Repeating the process for all elements of the distance matrix. D: Sub-setting in-class ranks corresponding to in-class distance elements.

Given the numbers of samples is n in group 1 and in m group 2, total number of ranks (and therefore max value of a rank) in un-subsetted triangle matrix will be

$$M = \frac{(m+n)(m+n-1)}{2} \tag{3}$$

Likewise, total number of ranks is subsetted to in-class values will be

$$N = \frac{(m^2+n^2-(m+n))}{2} \tag{4}$$

Σ reaches its min value A if the subset procedure selects minimal values of ranks, or, in other words, ranks in the selected part of rank matrix are values from 1 to M:

$$A = \sum_{i=1}^{M} i \tag{5}$$

The minimal value is reached when the lowest ranks reside in the diagonal squares of the distance matrix that correspond to in-group distances, and, therefore, best groups separation. Similarly, max value B is delivered when maximal values of Ranks Matrix are selected - from $(N - M)$ to N:

$$B = \sum_{i=(N-M)}^{N} i \tag{6}$$

Thus, $\Sigma \in [A, B]$. Now performing a scaling of compact $[A, B]$ to $[0, 1]$ finally allows us to retrieve value α with requested properties:

$$F : [A,B] \rightarrow [0,1], F(A) = 1, F(B) = 0. \tag{7}$$

$$F : S \rightarrow \alpha \tag{8}$$

This procedure finally allows to introduce measure α with necessary listed properties. We refer to this measure as *Hobotnica*, or H-score.

Thus, for every Gene Signature GS_i from set of hypotheses $\{H_1 : GS_1, H_2 : GS_2, ..., H_n : GS_n\}$ H-score α_i may be computed, resulting in a set $\langle \alpha_1, \alpha_2, ... \alpha_n \rangle$. Comparing and ranking α values allows for corresponding Gene Signatures qualities ranking and comparison.

The proposed approach is different from existing metrics, such as often used Rand index[5] and other clusteringbased measures, as it allows one to avoid clustering procedure, that itself may be carried out with various approaches and parameters.[6–8] In contract to clustering-based methods, H-scores directly reflects the sample stratification quality.

Validation

To validate our approach, we conducted four case studies.

In the first case study we extracted RNA-seq expression dataset for prostate cancer from the Cancer Genome Atlas (TCGA) on counts level.[9] As MFSs, we recruited the C2 collection of molecular signatures from MSigDB,[3,10] that contains a number of prostate-related gene sets. This way, every candidate MFS (gene set from the collection) produced its specific H-score.

For the second case study, we recruited the PAM50 molecular signature,[11] which was designed for classifying various breast cancer subtypes, as MFS. Then, we applied it to several datasets containing these breast cancer subtypes.[9,12–15]

In the third case study, we explored H-scores delivered by various DGE approaches. We performed DGE analysis for two groups of mice samples with different response to MYC factor treatment (Mycfl/fl vs MycΔIE, ERT2 genotypes)[16] with DESeq2[17] and edgeR.[18]

The top 100 genes for each method were then retrieved. In addition, we extracted a list of genes genes with the highest variance in expression, as well as a number of random gene sets.

In each case, the counts were normalised to counts per million (cpm). For every geneset an H-score and its p-value with BH[19] correction were computed.

In the last case the Hobotnica application for differential methylation signatures assessment was demonstrated. Hobotnica was applied to the signature from study[20] that distinguish B-cell subpopulations with mutated and unmutated IGHV from patients with chronic lymphocytic leukemia (M-CLL and U-CLL). The signature was derived from the data obtained on 450k Human Methylation Array, the length of the signature is 3265 sites.

The signature was validated on datasets from other experiments containing the same comparison groups (M-CLL and U-CLL): GSE136724 from,[21] GSE143411 from[22] and GSE144894 from.[23] Datasets GSE136724 and GSE143411 contain samples from patients after chemo(immune) therapy and untreated samples, only untreated groups were used for validation. H-score was calculated for each dataset using matrices of beta values with beta-mixture quantile normalization. The signature was reduced to 3089 sites for GSE136724, 3091 sites for GSE143411 and 3254 for GSE144894. The p-value was calculated based on 100000 random signatures of the same length using pseudocount.

Results

Prostate-related C2 gene sets clearly demonstrated highest H-score values and sufficient statistical significance (Table 1, Figure 2A), as well as data stratification (Figure 2B), which is expected for prostate cancer as opposed to control contrast. Gene sets not attributed to prostate cancer-related processes did not achieve statistically significant p-values (Table 1).

H-scores for the PAM50 signature were evidently higher for all datasets in the second case study than those for random gene sets for the same datasets (Figure 3, Figure 2C). This implies that the PAM50 signature exhibits a high stratification quality for various breast cancer subtypes samples. PAM50-delivered H-scores also demonstrated highly statistically significant p-values (Table 2).

In the third case study, various DGE approaches resulted in gene sets that delivered significantly different H-scores (Figure 4). For this dataset, edgeR provided a signature with the best quality score, while DESeq2 still demonstrated a higher separation quality than that of random signatures. Genes with the highest variance showed lower scores compared to random gene sets. This result stresses the importance of the Hobotnica procedure to evaluate the quality of a particular DGE analysis.

Table 3 contains the result H-scores and the corresponding p-values for differential methylation signature validation. H-score values are close to 1, and the p-values are less than 0.05 for all tested datasets. The distributions of Hscores

Table 1. Ten C2-chemical and genetic perturbations (GCP) Gene Signatures with the highest H-scores.

Signature	H-score	p-value
TOMLINS_PROSTATE_CANCER	0.795	**0.025**
WALLACE_PROSTATE_CANCER	0.747	**0.025**
OUYANG_PROSTATE_CANCER_PROGRESSION	0.745	**0.025**
LIU_PROSTATE_CANCER	0.735	**0.025**
PIEPOLI_LGI1_TARGETS	0.724	0.059
SMID_BREAST_CANCER_RELAPSE_IN_LIVER	0.712	0.164
TIMOFEEVA_GROWTH_STRESS_VIA_STAT1	0.708	0.240
GENTILE_UV_LOW_DOSE	0.705	0.308
JOHANSSON_BRAIN_CANCER_EARLY_VS_LATE	0.701	0.377
HOWLIN_CITED1_TARGETS_1	0.700	0.377

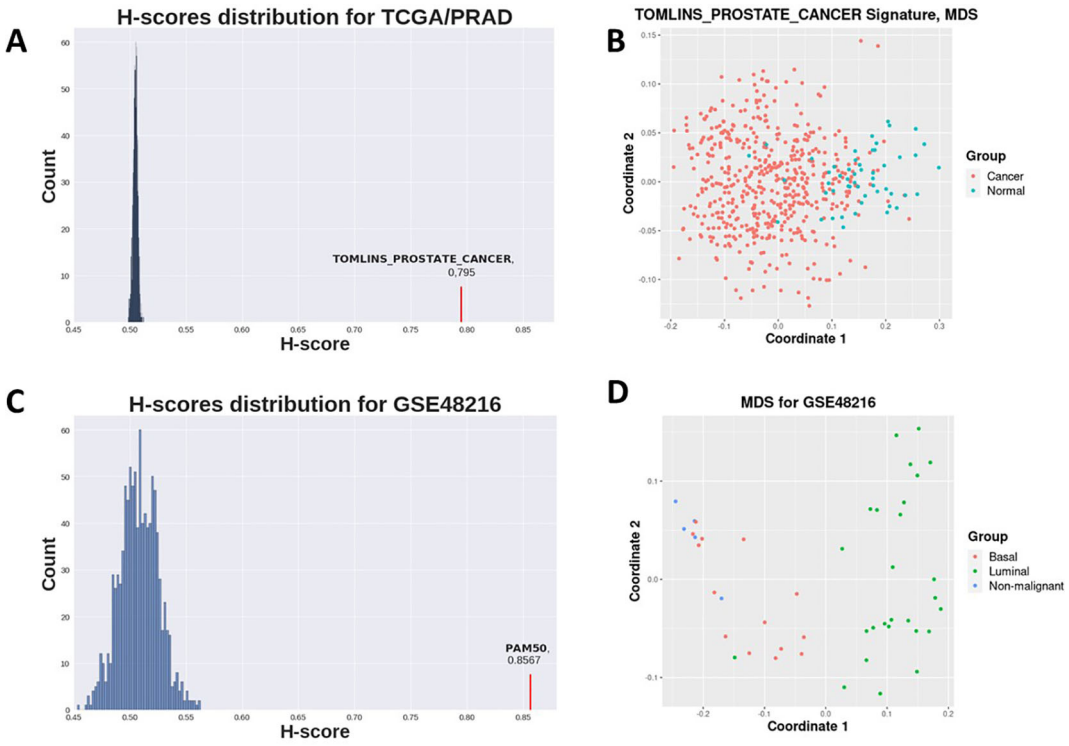

Figure 2. A : Distribution of H-scores for random genesets (blue) on TCGA prostate cancer vs normal dataset (see Table 1) and Tomlins prostate geneset H-score (red). B: MDS for TCGA prostate demonstrates samples separation with Tomlins geneset. C: Distribution of H-scores for random genesets (blue) on GSE48216 breast cancer dataset (see Table 2) and PAM50 geneset H-score (red). D: MDS for GSE48216 breast cancer dataset samples separation with PAM50 geneset.

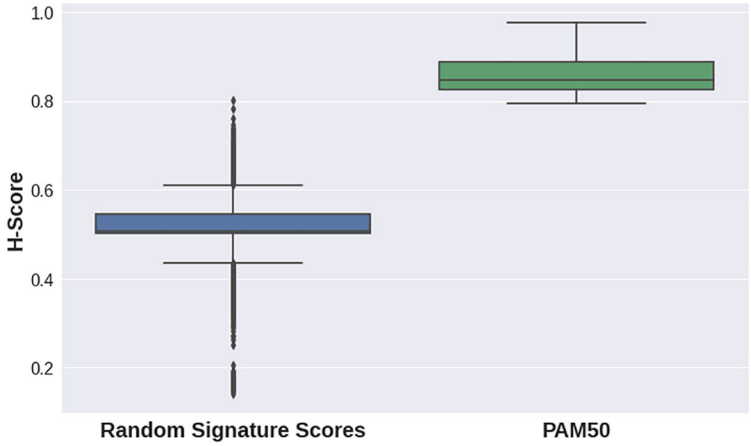

Figure 3. Distribution of random gene sets-delivered (blue) and PAM50 gene set-delivered (green) H-scores for breast cancer datasets (see Table 2).

Table 2. PAM50 results.

GEO Accession	Sample size	Groups in dataset	H-score	p-value
GSE58135	168	6	0.772	7e-4
GSE62944	1067	5	0.8892	0.0003
GSE48216	46	3	0.8567	0.0003
GSE80333	10	3	0.9765	0.0003

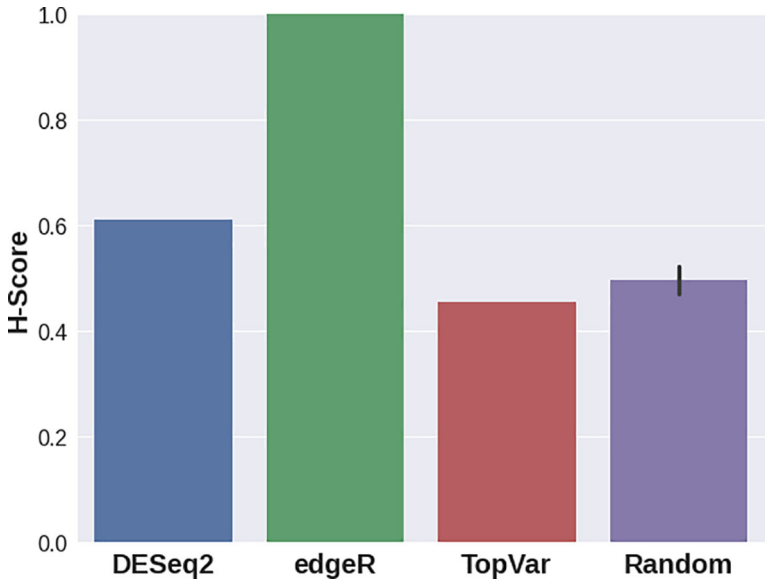

Figure 4. H-scores for the top 100 Gene Signatures delivered from DESeq2, edgeR, genes with highes variance and random gene sets applied to GSE155460 data.

Table 3. H-scores and p-values for the differential methylation signature that differentiate M-CLL and UCLL.

GEO Accession	Unmutated sample size	Mutated sample size	H-score	p-value
GSE136724	22	7	0.9586	9.9999e-06
GSE143411	8	2	1	9.9999e-06
GSE144894	44	76	0.9551	9.9999e-06

obtained from 100000 random signatures of the same length that were used for p-value calculation are shown in Figure 5. Figure 5 displays MDS plots based on submatrices that include only differential methylation signature sites.

Discussion

Hobotnica was designed to quantitatively evaluate MFS quality through their ability for data stratification, based on their inter-sample distance matrices, and to assess the statistical significance of the results. We demonstrated that Hobotnica can efficiently estimate the quality of a molecular signature in the context of expression data.

The suggested method can be used to evaluate various sorts of MFSs: those retrieved from DGE or DM analyses, Mutation/single nucleotide variation calling or pathways analysis, as well as data modalities of other types, that are suitable as differential problems.

The non-parametric statistic used in the approach not only allows for MFS of various types (differential, predefined, etc.) and data modalities (expression, methylation, etc.), but also for different structure of contrasted samples groups (sample size, preprocessing methods, etc.).

A possible application of Hobotnica is evaluating a particular model's performance (e.g., DGE model) for a particular dataset. This will allow researchers to choose a method that delivers a signature with the best data stratification from a number of proposed approaches.

Assessing H-score values for various lengths of the same set or signature can be explored with the proposed method, which will help to optimize MFS structure. Such procedures can be especially crucial in clinical applications.

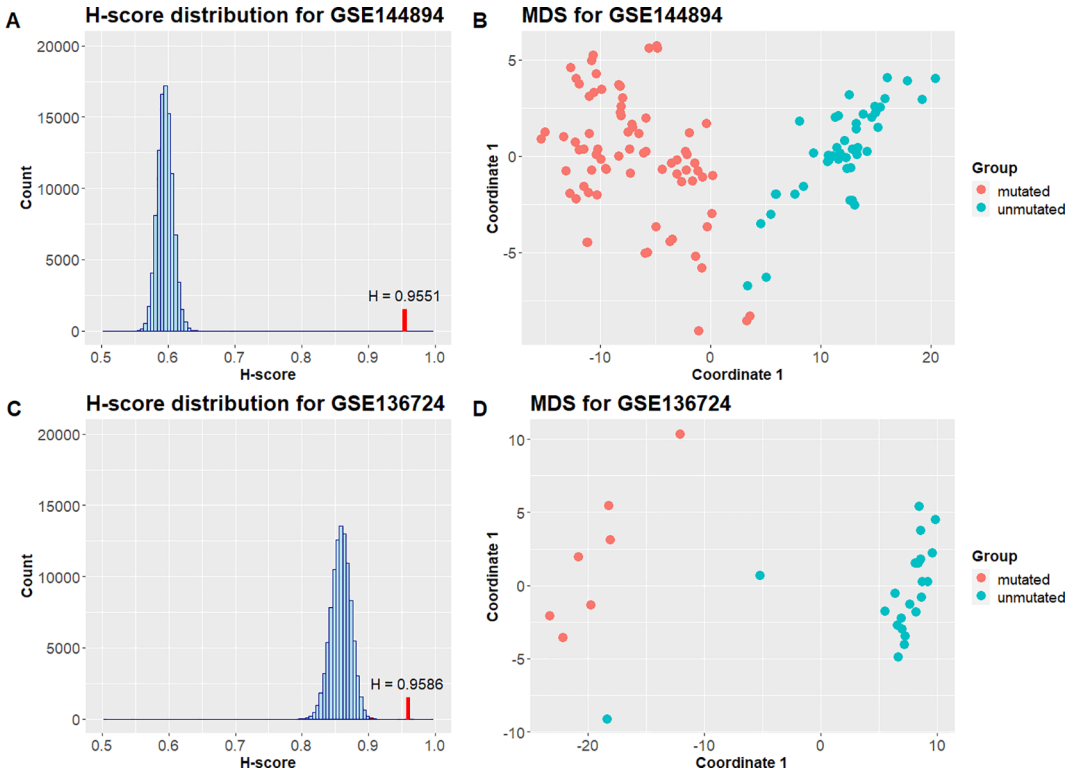

Figure 5. A: H-score distribution for GSE144894 dataset (blue) and H-score corresponding to the validated signature (red) B: MDS plot for GSE144894 samples based on the differential methylation signature C: H-score distribution for GSE136724 dataset (blue) and H-score corresponding to the validated signature (red) B: MDS plot for GSE136724 samples from untreated group based on the differential methylation signature.

Data availability

Underlying data

NCBI Gene Expression Omnibus: Alternatively processed and compiled RNA-Sequencing and clinical data for thousands of samples from The Cancer Genome Atlas, https://identifiers.org/ncbiprotein:GSE62944

NCBI Gene Expression Omnibus: Modeling precision treatment of breast cancer, https://identifiers.org/ncbiprotein:GSE48216

NCBI Gene Expression Omnibus: Spatial proximity to fibroblasts impacts molecular features and therapeutic sensitivity of breast cancer cells influencing clinical outcomes, https://identifiers.org/ncbiprotein:GSE80333

NCBI Gene Expression Omnibus: Next Generation Sequencing Analysis of Mycfl/fl and MycIE, ERT2 intestinal transcriptomes, https://identifiers.org/ncbiprotein:GSE155460

NCBI Gene Expression Omnibus: DNA methylation of chronic lymphocytic leukemia with differential response to chemotherapy, https://identifiers.org/ncbiprotein:GSE136724

NCBI Gene Expression Omnibus: A dataset of sequential DNA methylation profiles (2 timepoints) of 10 patients with chronic lymphocytic leukemia, https://identifiers.org/ncbiprotein:GSE143411

NCBI Gene Expression Omnibus: CLL intraclonal fractions exhibit established and recently-acquired patterns of DNA methylation [ME], https://identifiers.org/ncbiprotein:GSE144894

Extended data

Analysis code

Analysis code available from: https://github.com/lab-medvedeva/Hobotnica-main

Archived analysis code as at time of publication: https://doi.org/10.5281/zenodo.5656814

License: GNU General Public License v2.0

Acknowledgements

We thank Frank Emmert-Streib, Leslie Cope and Elana Fertig for fruitful discussions.

References

1. Parker JS, Mullins M, Cheang MCU, *et al.*: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J. Clin. Oncol.* 2009; **27**(8): 1160–1167.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Cardoso F, van't Veer LJ, Bogaerts J, *et al.*: **70-gene signature as an aid to treatment decisions in earlystage breast cancer.** *N. Engl. J. Med.* 2016; **375**(8): 717–729.
Publisher Full Text

3. Subramanian A, Tamayo P, Mootha VK, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc. Natl. Acad. Sci.* 2005; **102**(43): 15545–15550.
PubMed Abstract | Publisher Full Text | Free Full Text

4. Liu C, Jing S, Yang F, *et al.*: **Compound signature detection on lincs l1000 big data.** *Mol. BioSyst.* 2015; **11**(3): 714–722.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Rand WM: **Objective criteria for the evaluation of clustering methods.** *J. Am. Stat. Assoc.* 1971; **66**(336): 846–850.

6. Abbas OA: **Comparisons between data clustering algorithms.** *Int. Arab. J. Inf. Technol.* 2008; **5**(3).

7. Chen G, Jaradat SA, Banerjee N, *et al.*: **Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data.** *Stat. Sin.* 2002; 241–262.

8. Kafieh R, Mehridehnavi A: **A comprehensive comparison of different clustering methods for reliability analysis of microarray data.** *J. Med. Signals Sens.* 2013; **3**(1): 22.

9. Rahman M, Jackson LK, Evan Johnson W, *et al.*: **Alternative preprocessing of rna-sequencing data in the cancer genome atlas leads to improved analysis results.** *Bioinformatics.* 2015; **31**(22): 3666–3672.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Liberzon A, Subramanian A, Pinchback R, *et al.*: **Molecular signatures database (MSigDB) 3.0.** *Bioinformatics.* 05 2011; **27**(12): 1739–1740. ISSN 1367-4803.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Parker JS, Mullins M, Cheang MCU, *et al.*: **Supervised risk predictor of breast cancer based on intrinsic subtypes.** *J. Clin. Oncol.* 2009; **27**(8): 1160–1167.
PubMed Abstract | Publisher Full Text

12. Varley KE, Gertz J, Roberts BS, *et al.*: **Recurrent read-through fusion transcripts in breast cancer.** *Breast Cancer Res. Treat.* 2014; **146**(2): 287–297.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Marusyk A, Tabassum DP, Janiszewska M, *et al.*: **Spatial proximity to fibroblasts impacts molecular features and therapeutic sensitivity of breast cancer cells influencing clinical outcomes.** *Cancer Res.* 2016; **76**(22): 6495–6506.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Daemen A, Griffith OL, Heiser LM, *et al.*: **Modeling precision treatment of breast cancer.** *Genome Biol.* 2013; **14**(10): R110–R114.
Publisher Full Text

15. Costello JC, Heiser LM, Georgii E, *et al.*: **A community effort to assess and improve drug sensitivity prediction algorithms.** *Nat. Biotechnol.* 2014; **32**(12): 1202–1212.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Luo Y, Yang S, Wu X, *et al.*: **Intestinal MYC modulates obesity-related metabolic dysfunction.** *Nat. Metab.* July 2021; **3**(7): 923–939.
PubMed Abstract | Publisher Full Text

17. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for rna-seq data with deseq2.** *Genome Biol.* 2014; **15**(12): 1–21.
Publisher Full Text

18. Robinson MD, McCarthy DJ, Smyth GK: **edger: a bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J. R. Stat. Soc. Series B* 1995; **57**(1): 289–300.

20. Kulis M, Heath S, Bibikova M, *et al.*: **Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia.** *Nat. Genet.* Nov 2012; **44**(11): 1236–1242.

21. Yosifov DY, Bloehdorn J, Döhner H, *et al.*: **DNA methylation of chronic lymphocytic leukemia with differential response to chemotherapy.** *Sci. Data.* 05 2020; **7**(1): 133.

22. Zapatka M, Tausch E, Öztürk LM, *et al.*: **Clonal evolution in chronic lymphocytic leukemia is scant in relapsed but accelerated in refractory cases after chemo(immune) therapy.** *Haematologica.* 03 2022; **107**(3): 604–614.
Publisher Full Text

23. Bartholdy BA, Wang X, Yan XJ, *et al.*: **CLL intraclonal fractions exhibit established and recently acquired patterns of DNA methylation.** *Blood Adv.* 03 2020; **4**(5): 893–905.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 2

Reviewer Report 26 September 2022

https://doi.org/10.5256/f1000research.134369.r147651

© **2022 Malinverni R.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Roberto Malinverni**
Cancer and Leukemia Epigenetics and Biology Program, Josep Carreras Leukemia Research Institute (IJC), Badalona, Spain

In my opinion the authors answer to all my criticism. In this version of the article, they correct some imprecision in the graphs and add figures that help to understand the method. I think that now is ready for indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetics. R-developrer

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 20 September 2022

https://doi.org/10.5256/f1000research.134369.r147650

© **2022 Tripathi S.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Shailesh Tripathi**
[1] Production and Operations Management, University of Applied Sciences Upper Austria, Linz, Austria
[2] FH Austria, Steyr, Austria

I have gone through the responses provided by the authors and the updates of the paper. The author has addressed the main questions and revised the manuscript. I have no further questions.

For the final acceptance, I approve it.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data science, Machine learning, network analysis, computational biology, gene expression data analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 05 January 2022

https://doi.org/10.5256/f1000research.78645.r102284

© 2022 Tripathi S. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Shailesh Tripathi
[1] Production and Operations Management, University of Applied Sciences Upper Austria, Linz, Austria
[2] FH Austria, Steyr, Austria

The authors present an approach called Hobotonica for quantitatively evaluating (by assigning H score) MFS quality for given sample labels. This approach could be useful for analyzing samples, for e.g., quality comparison, filtering out poor quality samples, and comparing different phenotypical conditions and experiments. It is important that the authors should discuss a reasonable H-score interpretation in terms of various implications of data quality/outcome related to experimental conditions, sample size, data preprocessing, and the complexity of biological systems reflecting the non-trivial correlation structure.

I highlight some of the recommendations to be discussed in the paper:

1. The author should add simulation studies providing a realistic understanding and interpretation of the H score.

2. How is the current approach different from the clustering-based approach where the optimized number of clusters are compared to sample labels using rand-index (where a high rand score means the clustering solution and the sample labels are in agreement) or other measures.

3. The analysis should consider experimental conditions (data derived from multiple experiments representing the same phenotype), data preprocessing methods, sample size, and gene expression data covariance structure.

4. How does H-score vary with relation to the number of phenotype conditions and number of MFS. The authors should add analysis and interpretation of results:
    - when MFS is differentially expressed genes.
    - when MFS is randomly selected.
    - When MFS is a predefined set (e.g., GO pathway).

5. The author should add accurate descriptions of all the notations used.

6. Add a definition of H-score.

Is the rationale for developing the new method (or application) clearly explained?
Partly

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Data science, Machine learning, network analysis, computational biology, gene expression data analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 May 2022

Alexey Stupnikov, Moscow Institute of Physics and Technology, Moscow, Russian Federation

- **The author should add simulation studies providing a realistic understanding and interpretation of the H score.**

The Reviewer raises an important problem of parametric and nonparametric statistics. The

nature of H-score distribution indeed was not discussed in detail. Yet, its distribution was not the focus of this study, for distribution of H-scores for randomly selected Molecular Feature Sets is only employed to compute empirical p-values. This procedure is nonparametric and, therefore, is not affected by the nature of H-score distribution. We agree the nonparametric nature of the statistic needs to be mentioned more explicitly in the manuscript, and add a paragraph to the Discussion section:

"*The non-parametric statistic used in the approach not only allows for MFS of various types (differential, predefined, etc.) and data modalities (expression, methylation, etc.), but also for different structure of contrasted samples groups (sample size, preprocessing methods, etc.)*"

- **How is the current approach different from the clustering-based approach where the optimized number of clusters are compared to sample labels using rand-index (where a high rand score means the clustering solution and the sample labels are in agreement) or other measures.**

We agree the difference of our approach from clustering-based methods was not mentioned explicitly. We have added a paragraph to the Methods section discussing the difference of our approach.

"The proposed approach is different from existing metrics, such as often used Rand index and other clustering-based measures, as it allows one to avoid clustering procedure, that itself may be carried out with various approaches and parameters. In contract to clustering-based methods, H-scores directly reflects the sample stratification quality."

- **The analysis should consider experimental conditions (data derived from multiple experiments representing the same phenotype), data preprocessing methods, sample size, and gene expression data covariance structure.**

The dataset details the reviewer mentions are crucial for understanding a particular dataset's structure, and it is correct that they may affect some types of analysis. However, since the nature of the statistic is nonparametric, the details do not affect our approach. We agree this point needs to be stressed explicitly in the manuscript, and add a paragraph in the Discussion section:

"*The non-parametric statistic used in the approach not only allows for MFS of various types (differential, predefined, etc.) and data modalities (expression, methylation, etc.), but also for different structure of contrasted samples groups (sample size, preprocessing methods, etc.)*"

The exploration of H-score distribution in relation to experimental conditions mentioned above certainly is an interesting fundamental question. However, it does not affect the practical implementation we introduce. Therefore, we believe that such study is beyond the scope of a current paper.

- **How does H-score vary with relation to the number of phenotype conditions and number of MFS. The authors should add analysis and interpretation of results:**
  - **when MFS is differentially expressed genes.**
  - **when MFS is randomly selected.**
  - **When MFS is a predefined set (e.g., GO pathway).**

The nature of MFS indeed may be quite different. For this reason we have performed and

discussed the analysis of following MFS types:
- differentially expressed genes
- randomly selected genes
- MysigDB gene sets (predefined genesets)

In the revised version of manuscript differentially methylated MFS are also considered. In this way, we feel the analysis and the interpretation the Reviewer suggested can be found in the manuscript.

- **The author should add accurate descriptions of all the notations used.**

To address Reviewer's comment we carefully checked the manuscript to ensure all introduced notations are defined and explained, and added a subsection to the Methods section for more detailed description of H-score.

- **Add a definition of H-score.**

We agree the definition of the H-score statistic and notation we introduce should be expanded. To address Reviewer's comment, we have added a subsection to the Methods section where we define H-score with more details.

Competing Interests: No competing interests were disclosed.

Reviewer Report 20 December 2021

https://doi.org/10.5256/f1000research.78645.r102280

© 2021 Malinverni R. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Roberto Malinverni
Cancer and Leukemia Epigenetics and Biology Program, Josep Carreras Leukemia Research Institute (IJC), Badalona, Spain

In this short article the authors present an R package called Hobotnica, whose purpose is to evaluate the goodness with which different methodologies can stratify the results presented as Molecular Feature Sets (MFS). With MFS the authors point to all those types of data as a result of different *-omics* techniques (such as expression, methylation, Mutation / single nucleotide variation calling or pathways analysis).

Major comments

1. The authors present three examples in which it is demonstrated how this approach is able to evaluate the effectiveness of MFS stratification, but the examples considered are all based on expression data. To verify the statements presented in the article, it would be useful to test the methodology on different data (for example methylation arrays). The approach chosen for this evaluation is based on the calculation and comparison of Distance

Matrices (DM).

2. The example of figure 3 evaluates two different standard approaches for the analysis of RNAseq using Hobotnica and the H0 value as discriminant. It can be appreciated in this figure how the stratification quality of Deseq2 is decidedly more efficient than both random genes and top variant genes. Surprisingly, however, the H0 value calculated using the top 100 genes collected with edgeR is very similar to that calculated using random genes, this confused me. Authors should explain this similarity more in depth.

The data presented in this article do not seem to convince satisfactorily. The quality evaluation power obtained by applying Hobotnica does not seem to correspond to the premises made. While not in fact a slate on the methodology, my advice is to review the examples and try to improve in benchmarking, adding different types of data.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Partly

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Epigenetics. R-developrer

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 23 May 2022

Alexey Stupnikov, Moscow Institute of Physics and Technology, Moscow, Russian Federation

- **The authors present three examples in which it is demonstrated how this approach is able to evaluate the effectiveness of MFS stratification, but the**

examples considered are all based on expression data. To verify the statements presented in the article, it would be useful to test the methodology on different data (for example methylation arrays). The approach chosen for this evaluation is based on the calculation and comparison of Distance Matrices (DM).

We thank the Reviewer for this suggestion. Indeed, the method we propose is applicable for Molecular Feature Sets evaluation of different nature, yet in the manuscript we demonstrated its work only for Expression based data.To improve the manuscript in a way the Reviewer suggested , we have performed additional analysis on Methylation based data for several datasets. We added this case study to the Validation and Results sections and comments to the Discussion section.

- **The example of figure 3 evaluates two different standard approaches for the analysis of RNAseq using Hobotnica and the H0 value as discriminant. It can be appreciated in this figure how the stratification quality of Deseq2 is decidedly more efficient than both random genes and top variant genes. Surprisingly, however, the H0 value calculated using the top 100 genes collected with edgeR is very similar to that calculated using random genes, this confused me. Authors should explain this similarity more in depth.**

The point the Reviewer mentions indeed raises concern. After thorough examination we identified a data analysis related problem that caused the wrong genes subset process and resulted in the incorrect depiction in the presented barplot. We have corrected this issue. The rest of our findings were not changed during our reevaluation and all the results hold. The results are depicted and Fig.5 in the current manuscript version.

- **The data presented in this article do not seem to convince satisfactorily. The quality evaluation power obtained by applying Hobotnica does not seem to correspond to the premises made. While not in fact a slate on the methodology, my advice is to review the examples and try to improve in benchmarking, adding different types of data.**

To address the Reviewer's comments on the evaluation made in the manuscript we conducted additional analysis and validation carried out for methylation data modality on several datasets, and improved and fixed validation for differential expression analysis. Now we feel that the argument and claims we make regarding the H-score are compelling and plausible.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research