ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 21(9): 3162-3164 (2025)

www.bioinformation.net
Volume 21(9)

Research Article

Received September 1, 2025; Revised September 30, 2025; Accepted September 30, 2025, Published September 30, 2025
DOI: 10.6026/973206300213162

SJIF 2025 (Scientific Journal Impact Factor for 2025) = 8.478
2022 Impact Factor (2023 Clarivate Inc. release) is 1.9

Edited by P Kangueane
Citation: Nuc *et al.* Bioinformation 21(9): 3162-3164 (2025)

MASHt: Software for statistical analysis of transcriptomes' qualitative features in factorial experiments

Maria Nuc[1], Michał Stanoch[1], Hanna Ćwiek-Kupczyńska[1,2], Barbara Naganowska[1] & Paweł Krajewski[1,*]

[1]Institute of Plant Genetics, Polish Academy of Sciences, Strzeszyńska 34, 60-479 Poznań, Poland; [2]Luxembourg Centre for Systems Biomedicine, University du Luxembourg, 1 Bd du Jazz, 4370 Esch-Belval, Luxembourg ; *Corresponding author

Availability:
The source code and explanations are available at https://github.com/mikrzol/masht and in ZENODO at https://dx.doi.org/10.5281/zenodo.13382550.

Affiliation URL:
www.igr.poznan.pl

Author contacts:
Maria Nuc - E-mail: mnuc@igr.poznan.pl
Michał Stanoch - E-mail: msta@igr.poznan.pl
Hanna Ćwiek-Kupczyńska - E-mail: hanna@cwiek-kupczynska.com

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 21(9): 3162-3164 (2025)

Barbara Naganowska - E-mail: bnag@igr.poznan.pl
Paweł Krajewski - E-mail: pkra@igr.poznan.pl

Abstract:
In Next-Generation Sequencing (NGS) data analysis, attention is given more often to the variation in gene expression levels in several experimental variants than to the qualitative differences between transcriptomes. We present the MASHt toolkit for analyzing qualitative variation in sequencing data from multifactorial experiments. The analysis involves computing pairwise Mash distances between datasets, conducting principal coordinate analysis of the resulting distance matrix, and applying univariate and multivariate analyses of variance to the principal coordinates. MASHt supports computations on data subsets, such as those annotated by a specific ontology. We demonstrate the use of MASHt with an example experiment examining the impact of temperature on the transcriptomes of different barley genotypes.

Keywords: transcriptomics, factorial experiments, sequence variation, software tools

Background:
Transcriptomics most often focuses on differences in NGS-derived transcript quantities between samples representing various genotypes or conditions, and numerous tools exist for such differential gene expression analysis **[1]**. However, transcriptomic data are less frequently used to identify qualitative differences between sequences characteristic of different organisms, cell types, or conditions, which encompass the variability in the presence/absence patterns of transcripts, alternative splicing, or structural variations; these may arise from genomic variations, genetic mosaicism or chimerism, chromatin structure differences, or post-transcriptional modifications **[2]**. Additionally, biotic and abiotic environmental factors may affect transcript maturation mechanisms **[3]**. In simple case/control designs, finding specific polymorphisms may be sufficient for the inference, while in experiments involving many (possibly interacting) factors, the analysis should be based on statistical models appropriate for studying the variation between treatments relative to the background variation, considering the experimental design **[4]**. The proposed MASHt toolkit identifies significant qualitative differences in sequencing data from multifactorial experiments using linear models; it utilizes Mash distances **[5]**, which measure qualitative differences by identifying common and unique k-mers between datasets. Then, it conducts principal coordinate analysis (PCoA; **[6]**) on the resulting distance matrix to transform qualitative differences into quantitative variables and applies univariate and multivariate analyses of variance to the principal coordinates. MASHt supports the analysis of datasets annotated with a chosen ontology, facilitating functional interpretation of qualitative differences **[7]**. Therefore, it is of interest to describe the design and development of software for statistical analysis of transcriptomes' qualitative features in factorial experiments.

Program input:
The input consists of sets of nucleotide sequences. The package includes three modules—mash, stats, blaster—each automating a specific step of the analysis.

Mash:
Compute Mash distances. Many types and formats of input data are allowed, including fasta and fastq, containing long (e.g., scaffolds or full-length transcripts) or short (e.g., sets of NGS reads) sequences.

Stats:
It performs PCoA to represent the data structure in a low-dimensional space on the basis of calculated distances. The subsequent ANOVA/MANOVA tests, based on an input file containing the description of the experimental design and linear model formula, assess whether experimental factors or their interactions account for the observed variability.

Blaster:
Splits observed sequences into groups based on their mapping to a reference set. The inputs for this module include a file containing reference sequences annotated with features of interest (*e.g.*, GO or GOSlim terms).

Program output:
To illustrate the MASHt approach and output, we used RNA-seq data obtained in the experiment performed to examine the impact of heat stress on the barley transcriptome **[8]**. Three genotypes were investigated ('Bowman' cultivar, BW, and its two near-isogenic lines, BW827 and BW828) under two treatments: optimal temperature (8 °C/16 °C night/day from sowing to the end of tillering) and high temperature (28 °C night/day in the same period). Samples were collected at two timepoints—the start of the tillering stage and ten days later—in three biological replicates. The data were assembled into scaffolds using SOAPdenovo2 **[9]**. The primary outputs obtained from MASHt were the distance matrix between all pairs of sets of scaffolds, PCoA results with low-dimensional plots **(Figure 1A)** and ANOVA/MANOVA results (**Table 1**). The analysis showed that day, temperature, and their interaction had statistically significant effects on the distribution of the data points **(Figure 1A)**. In general, points associated with the optimal temperature have lower values of PC1 than those associated with high temperature, and points associated with the first day of tillering have lower values of PC2 than those associated with the tenth day. To perform the part of the analysis related to functional annotation, the reference set of barley transcript isoforms, annotated with GOSlim terms, was downloaded from the Ensembl Plants database. Each of the target datasets was

divided into subsets containing scaffolds linked to a particular term, and the analysis was performed for these subsets (with Bonferroni correction relative to the number of terms). The results of this analysis are presented (**Figure 1B**) in the GOSlim ontology graph.

Table 1: Results of ANOVA and MANOVA performed on the principal coordinates of barley scaffold datasets

Source of variation	P value from ANOVA for principal coordinates				P value from MANOVA
	PC1	PC2	PC3	PC4	
Genotype (G)	0.9794	0.289	0.76	0.5929	0.4717
Day (D)	0.555	0.024	5.23E-09	0.0001	4.41E-12
Temperature (T)	1.31E-05	0.25	0.4274	0.9548	6.72E-05
G x D	0.2278	0.714	0.1281	0.6033	0.4961
G x T	0.2357	0.399	0.0522	0.802	0.1521
D x T	0.0657	0.31	8.12E-09	4.06E-05	2.23E-09

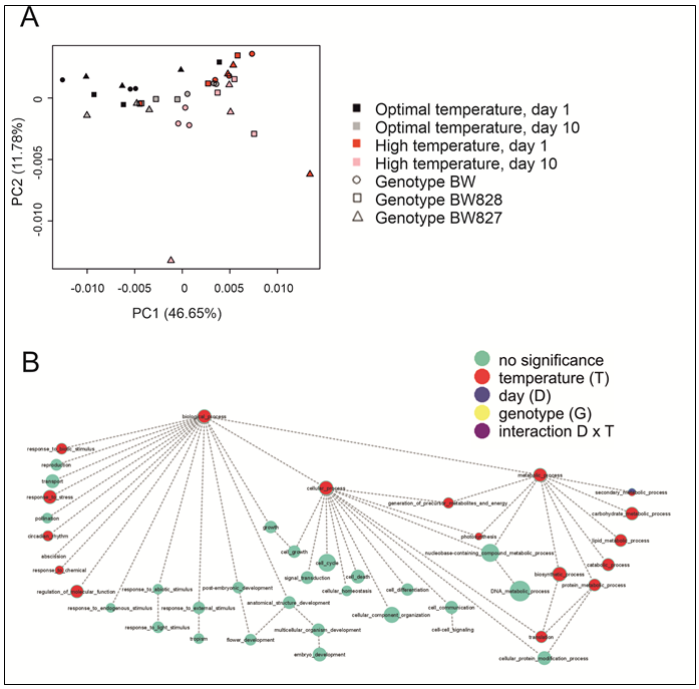

Figure 1: A. Location of sets of scaffolds in the coordinate system of PC1 and PC2 based on principal coordinate analysis of the Mash distance matrix. B. Visualization of MASHt results for barley scaffolds annotated by GOSlim terms describing biological processes (PC1) obtained using Cytoscape **[11]**. Nodes represent 'Biological process' GOSlim terms and edges represent their connections; the size of nodes represents the percentage of variation explained by PC1, and colors represent experimental factors affecting the distribution of datasets.

Caveat:

MASHt is implemented as a Python package; it uses the MASH executable **[5]** for computing distances and BLAST+ **[10]** for grouping the observations on the basis of annotation according to ontology terms. For good inference, experimental designs involving biological replicates are recommended. The results of the analysis—significant changes in nucleotide sequences between various experimental variants—can be interpreted by themselves to infer biological conclusions from the experiments or as additional explanatory information in other analyses, e.g., in quantitative differential expression analysis or in genetic variant analysis. Computations in MASHt are fast, so its availability should help with the incorporation of qualitative methods into sequence analysis. Although developed for RNA-seq data in multifactorial experiments, MASHt is also applicable to DNA-seq data with appropriate interpretative adjustments.

Acknowledgments:

We are grateful to Dr. Grzegorz Koczyk, Head of Biometry and Bioinformatics, IPG PAS (gkoc@igr.poznan.pl), for helpful suggestions. This work obtained partial financing from the National Science Center, Poland, project nos. 2016/22/M/NZ9/00251 and 2016/23/B/NZ9/02677. The development and testing of the software was conducted at the Poznan Supercomputing and Networking Center (https://www.psnc.pl).

Conflicts of interest: None

References :

[1] Rosati D *et al. Comput Struct Biotechnol J.* 2024 **23**:1154 [DOI: 10.1016/j.csbj.2024.02.018].
[2] Corbett AH, *Curr Opin Cell Biol*. 2018 **52**:96 [DOI: 10.1016/j.ceb.2018.02.011].
[3] Biniaz Y *et al. Plants*. 2022 **11**:502 [DOI: 10.3390/plants11040502].
[4] Scheffé H, *The Analysis of Variance*, 1999, WIley & Sons Ltd. [ISBN: 978-0-471-34505-3].
[5] Ondov BD *et al. Genome Biol.* 2016 **17**:132 [DOI: 10.1186/s13059-016-0997-x].
[6] Zuur AF *et al. Analysing Ecological Data*, 2007, Springer [ISBN: 978-0-387-87457-9].
[7] Ashburner M *et al. Nat Genet.* 2000 **25**:25 [DOI: 10.1038/75556].
[8] Mikołajczak K *et al. BMC Genomics.* 2022 **23**:177 [DOI: 10.1186/s12864-022-08410-1].
[9] Luo R *et al. Gigascience.* 2012 **27**:18 [DOI: 10.1186/2047-217X-1-18].
[10] Camacho C *et al. BMC Bioinformatics.* 2009 **10**:421 [DOI: 10.1186/1471-2105-10-421].
[11] Shannon P *et al. Genome Res.* 2003 **13**:2498 [DOI: 10.1101/gr.1239303].