Data Article

Heavy metal tolerance in Scopelophila cataractae: Transcriptomic and epigenetic datasets

M. Teresa Boquete a,b,c,*, Marc W. Schmid d, Niels C.A.M. Wagemaker e, Sarah B. Carey f, Stuart F. McDaniel f, Christina L. Richards c,g, Conchita Alonso b

a Area of Ecology, Department of Functional Biology, Universidade de Santiago de Compostela, Lope Gómez de Marzoa s/n, 15705, Spain
b Department of Evolutionary Ecology, Estación Biológica de Doñana, CSIC, Americo Vespucio 26, 41092 Sevilla, Spain
c Department of Integrative Biology, University of South Florida, 4202 East Fowler Ave, 33620, Tampa, FL, USA
d MWSchmid GmbH, Hauptstrasse 34, CH-8750 Glarus, Switzerland
e Faculty of Science, Raboud University, Heyendaalseweg 135, 6525, AJ Nijmegen, the Netherlands
f Department of Biology, University of Florida, 876 Newell Dr, 32611, Gainesville, FL, USA
g Plant Evolutionary Ecology group, University of Tübingen, Auf der Morgenstelle 5, D-72076 Tübingen, Germany

Article info

Article history:
Received 12 September 2022
Revised 17 October 2022
Accepted 22 October 2022
Available online 28 October 2022

Keywords:
Abiotic stress
Bryophytes
Copper moss
DNA methylation
Gene expression
Intraspecific phenotypic variation

Abstract

Studying how different plant groups deal with heavy metal exposure is crucial to improve our understanding of the diversity of molecular mechanisms involved in plant stress response. Here, we used RNA sequencing (RNA-seq) and epigenotyping by sequencing (epiGBS) to assess gene expression and DNA methylation changes respectively in plants from four populations of the metallophyte moss Scopelophila cataractae treated with Cd or Cu in the laboratory. We built RNA-seq and epiGBS sequencing libraries from control and treated samples from each population and sequenced them using Illumina HiSeq 3000 (PE-150 bp) and Illumina HiSeq X-Ten System (PE-150 bp) respectively. For the RNA-seq data, we performed a read quality filter, mapped the reads to

DOI of original article: 10.1016/j.envexpbot.2022.104970
* Corresponding author.
E-mail address: teresa.boquete@usc.es (M.T. Boquete).
Social media: @Tere_BS (M.T. Boquete), @sarahbcarey (S.B. Carey), @mcdaniellab (S.F. McDaniel), @EcoEpig (C.L. Richards)

https://doi.org/10.1016/j.dib.2022.108710
2352-3409/© 2022 Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
Specifications Table

| Subject                  | Plant Science: General |
|--------------------------|------------------------|
| Specific subject area    | Transcriptomic and epigenetic changes during heavy metal exposure; abiotic stress response in plants |
| Type of data             | Raw fastq files |
|                          | Processed fasta files |
| How the data were acquired | Samples: extraction of total RNA and genomic DNA from control and treated plants. |
|                          | Data: RNA sequencing (RNA-seq) with Illumina HiSeq 3000 (PE-150 bp) and epigenotyping by sequencing (epiGBS) with Illumina HiSeq X-Ten System (PE-150 bp). |
|                          | Software RNA-seq: Trimmomatic v 0.36, Trinity v 2.8.4, CD-HIT v 4.6.6, Bowtie2 v 2.2.9, BUSCO v 3.1.0, RSEM, OmicsBox |
|                          | Software epiGBS: epiGBS custom pipeline (https://doi.org/10.5281/zenodo.7040291), RepeatMasker v 4.0, DIAMOND v 0.8.22. |
| Data format              | Raw data: Illumina FASTQ files for epiGBS (paired end reads), and for RNASeq (individual samples' Illumina paired end reads after quality trimming), and barcode file (barcodes.tsv) for epiGBS data submitted to the Sequence Read Archive (SRA) of NCBI under project number PRJNA790924. Processed data: reference contigs (de_novo_contigs.fasta.gz), annotation of the reference contigs (mergedAnnot.csv.gz), SNPs (snps.vcf.gz), methylation data (methylation.txt.gz), de novo transcriptome assembly (de_novo_transcriptome.fasta.gz), and de novo transcriptome assembly annotation (de_novo_transcriptome_annotation_blast&IPS.txt) |
|                          | Analysed data: supplementary tables (SupplementaryMaterials_Boquete_et_al_DiB.tar.gz) |
| Description of data collection | Gametophore tissue was collected from control, Cd-treated and Cu-treated plants of four populations of S. cataractae (3 biological replicates per treatment and population) and subjected to genomic DNA extraction. A subset of control and Cu-treated plants from two of the populations were subjected to total RNA extraction. RNA-seq and epiGBS libraries were sequenced on Illumina HiSeq 3000 and Illumina HiSeq X-Ten System respectively to obtain 150 bp paired-end reads. |

(continued on next page)
| Data source location | • Institution: Estación Biológica de Doñana (EBD-CSIC)  
|                      | • City/Town/Region: Sevilla, Andalucía  
|                      | • Country: Spain  
|                      | • Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: 17N 572397E 3951654N  
| Data accessibility   | Raw sequencing data and barcodes file can be accessed from NCBI SRA (BioProject ID: PRJNA790924)  
|                      | https://www.ncbi.nlm.nih.gov/bioproject/PRJNA790924  
|                      | epigBS custom scripts: can be found on zenodo (https://doi.org/10.5281/zenodo.7040291) [1]  
|                      | Processed data: reference contigs, annotation of the reference contigs, SNPs, methylation data, de novo transcriptome assembly, and de novo transcriptome assembly annotation are available on zenodo (https://doi.org/10.5281/zenodo.6695117) [2]  
|                      | Analysed data: supplementary tables are available on zenodo (https://doi.org/10.5281/zenodo.6695117) [2]  
| Related research article | Boquete, M.T., Schmid, M.W., Wagemaker, N.C.A.M., Carey, S.B., McDaniel, S.F., Richards, C.L., Alonso, C. Molecular basis of intraspecific differentiation for heavy metal tolerance in the copper moss Scopelophila cataractae. Environ. Exp. Bot. 201 (2022) 104970. https://doi.org/10.1016/j.envexpbot.2022.104970  

**Value of the Data**

• The transcriptomic data provide a detailed comparative view of gene expression changes in plants of the metallophyte moss Scopelophila cataractae with different Cu tolerance levels. The molecular mechanisms involved in intraspecific differentiation for Cu tolerance can thus be inferred.

• The epigenetic data provides a unique source of information about changes in DNA methylation in response to abiotic stress in bryophytes.

• Both datasets could be used by plant scientists to perform comparative studies of the evolution of plant abiotic stress responses.

• These data could also contribute to the identification of target genes for the development of biotechnological tools in the field of phytoremediation.

1. **Data Description**

An overview of the raw epigenetic and transcriptomic data available on the Sequence Read Archive (SRA) can be found in Table 1. The epigenetic dataset was obtained by extracting genomic DNA from control, Cd- and Cu-treated plants from four populations of the metallophyte moss Scopelophila cataractae that were collected in the field where they were exposed to different levels of heavy metals. The transcriptomic dataset was obtained from total RNA extracted from a subset of the previous sample set that included control and Cu-treated plants from two of the four populations (one tolerant and one sensitive population of S. cataractae). The plants were treated in the laboratory for 30 days, then DNA and RNA were extracted, and the sequencing libraries were prepared as described in Section 2.

In total, 26,638 transcripts of the de novo transcriptome assembly were successfully annotated. The GO terms related to all molecular functions represented in the list of differentially expressed transcripts (DETs) in response to Cu in two of the studied populations (Sc3 and Sc4) is shown in Fig. 1.

Finally, Fig. 2 shows all the significantly over-represented GO terms in response to Cu exposure between the annotated list of differentially expressed transcripts (test set) and the full list of annotated transcripts (reference set) for each population of the S. cataractae (Sc3 and Sc4).
Table 1
Sample names and additional information of the raw epigenetic (epiGBS sequencing reads) and transcriptomic (RNA-seq) datasets obtained from gametophore samples of the moss Scopelophila cataractae subjected to control (plants watered with deionized - DI - water), Cd, and/or Cu treatments (plants watered with DI water containing either 0.1 mM Cd or 1 mM Cu) available on the Sequence Read Archive (SRA). Multiplexed raw fastq files (R1 and R2 corresponding to paired-end sequencing) correspond to the epigenetic data; individual samples’ R1 and R2 file correspond to the transcriptomic data.

| SRA sample name | Population | Treatment | SRA library ID | File type | Filename (R1)     | Filename (R2)     |
|-----------------|------------|-----------|----------------|-----------|-------------------|-------------------|
| Scopelophila_cataractae_epiGBS | Sc1, Sc2, Sc3, Sc4 | Control, Cd-treated, Cu-treated | Sc_epiGBS | fastq | Sc_EpiGBS_R1.fq.gz | Sc_EpiGBS_R2.fq.gz |
| 10              | Sc3        | Control   | Sc3_C1_RNAseq  | fastq     | 10_R1_paired.fq.gz | 10_R2_paired.fq.gz |
| 21              | Sc3        | Control   | Sc3_C2_RNAseq  | fastq     | 21_R1_paired.fq.gz | 21_R2_paired.fq.gz |
| 29              | Sc3        | Control   | Sc3_C3_RNAseq  | fastq     | 29_R1_paired.fq.gz | 29_R2_paired.fq.gz |
| 16              | Sc4        | Control   | Sc4_C1_RNAseq  | fastq     | 16_R1_paired.fq.gz | 16_R2_paired.fq.gz |
| 30              | Sc4        | Control   | Sc4_C2_RNAseq  | fastq     | 30_R1_paired.fq.gz | 30_R2_paired.fq.gz |
| 38              | Sc4        | Control   | Sc4_C3_RNAseq  | fastq     | 38_R1_paired.fq.gz | 38_R2_paired.fq.gz |
| 22              | Sc3        | Cu-treated| Sc3_Cu1_RNAseq| fastq     | 22_R1_paired.fq.gz | 22_R2_paired.fq.gz |
| 33              | Sc3        | Cu-treated| Sc3_Cu2_RNAseq| fastq     | 33_R1_paired.fq.gz | 33_R2_paired.fq.gz |
| 34              | Sc3        | Cu-treated| Sc3_Cu3_RNAseq| fastq     | 34_R1_paired.fq.gz | 34_R2_paired.fq.gz |
| 9               | Sc4        | Cu-treated| Sc4_Cu1_RNAseq| fastq     | 9_R1_paired.fq.gz  | 9_R2_paired.fq.gz  |
| 23              | Sc4        | Cu-treated| Sc4_Cu2_RNAseq| fastq     | 23_R1_paired.fq.gz | 23_R2_paired.fq.gz |
| 28              | Sc4        | Cu-treated| Sc4_Cu3_RNAseq| fastq     | 28_R1_paired.fq.gz | 28_R2_paired.fq.gz |
Fig. 1. Pie chart of gene ontology (GO) terms related to all molecular functions represented in the list of differentially expressed transcripts (DETs) in response to Cu in Sc3 and Sc4 respectively. The number of DETs within each GO and the percentage from the total number of DETs are presented within brackets.
Fig. 2. Word cloud representation of gene ontology (GO) terms related to all molecular functions significantly over-represented in the list of differentially expressed transcripts (DETs) in response to Cu in Sc3 (A; n=22 GOs) and Sc4 (B; n=4 GOs). The size of the words is proportional to the frequency of the GO term in the DETs list. GOs inside grey rectangles show under-represented terms in the DETs’ list.

2. Experimental Design, Materials and Methods

2.1. Sample collection and laboratory treatments

We used plants from four populations of Scopelophila cataractae collected in the field where they were naturally growing under different levels of heavy metal contamination. These plants were clonally propagated in a growth chamber in the Department of Integrative Biology at the University of South Florida (Tampa, FL, USA) and treated with water (controls), water enriched with 0.1 mM Cd (Cd-treated) or water enriched with 1 mM Cu (Cu-treated) - 4-5 replicates per population and treatment - as described in [3]. After one month, we harvested all plants from the different treatments and split each sample into several aliquots: one to perform growth measurements [3], one to extract total RNA, and one to extract genomic DNA. The latter two aliquots were flash frozen in liquid nitrogen and stored at -80°C prior to extractions.

2.2. Genomic DNA extraction, epiGBS library preparation, sequencing, and filtering

We isolated genomic DNA (gDNA) from 36 samples (three biological replicates from each of control, Cd-treated, and Cu-treated plants) using the cetyltrimethylammonium bromide (CTAB) DNA extraction protocol for recalcitrant plant tissues (https://www.protocols.io/view/high-quality-dna-extraction-protocol-from-recalcitr–i8jchun). Then, we checked the quality of the gDNA with the NanoDrop (Nanodrop™ 8000 Spectrophotometer; Thermo Scientific), and quantified its concentration using the Qubit 3.0 Fluorometric dsDNA BR assay kit (Q32851; Life Technologies). We prepared the epiGBS libraries following [4] with 400 ng of gDNA per sample as starting material. First, we digested the gDNA with the restriction enzyme PstI; second, we ligated methylated, non-phosphorylated barcoded adapters to both ends of the digested fragments; third, we concentrated the library using the NucleoSpin™ Gel and PCR Clean-up Kit (12303368, Macherey-Nagel™) and performed a fragment size selection with 0.8x SPRI beads (A63880, Agencourt AMPure XP Beckman Coulter); fourth, we performed nick translation, bisulfite conversion of the DNA using the EZ Lightning methylation kit (Zymo Research), and amplified
the library with the KAPA HIFI Uracil+ Hotstart Ready Mix (Roche). Finally, we sequenced the libraries at Novogene (HK) Company Limited in Hong Kong on the Illumina HiSeq X-Ten System (PE-150 bp).

Using a custom pipeline, developed by [4] and modified as in https://doi.org/10.5281/zenodo.7040291, we demultiplexed the sequencing reads, quality trimmed them, and used these pooled reads to create a de novo reference genome assembly. Then, we individually mapped the reads from each sample to the assembly and performed strand-specific nucleotide (single nucleotide polymorphisms, SNPs) and methylation (single methylation polymorphisms, SMPs) variant calling. We recovered 85.8 million reads after demultiplexing with an average number of reads per sample of 2.5 million (ranging between 0.6-4.5 million).

The unfiltered datasets resulting from the epiGBS pipeline consisted of 279,103 SNPs and 22.3 million SMPs across all samples. We filtered these datasets as follows: (1) we removed SNPs and SMPs without a minimum coverage of 3 and a maximum coverage equal to the 99th percentile of the read coverage distribution, in at least one replicate sample per group. In total 58,773 SNPs and 290,547 SMPs were kept; (2) we identified low quality samples, i.e. samples lacking more than 20% of the SNPs or more than 25% of the SMPs; and (3) we removed these samples from the experimental design, used the original (unfiltered) SNP and SMP dataset and removed SNPs and SMPs without a minimum coverage of 10 and a maximum coverage equal to the 99th percentile of the read coverage distribution in at least one replicate sample per group (i.e. population x treatment combination). The last filtering step resulted in a final working dataset of 52,513 SNPs and 239,728 SMPs across 30 samples comprising 2-3 replicates per group.

2.3. Total RNA extraction, library preparation, sequencing, and filtering

We isolated total RNA from a subset of 12 samples, including three biological replicates from each of control and Cu-treated plants from two populations that had been exposed to different levels of heavy metals in the field and showed contrasting responses to Cu exposure in the laboratory, with the RNeasy Plant Mini Kit (74904, Qiagen) following the manufacturer instructions with small modifications. First, we added 450 µl of disruption buffer (RTL) containing 1% BME to each frozen sample, vortexted them vigorously, and ground them in a TissueLyser. Then, we followed the manufacturer’s protocol including the on-column DNase I digestion to remove DNA from all samples (Rnase-free DNase set, 79254, Qiagen). Then, we measured the concentration of total RNA in each sample using the Qubit RNA BR assay kit (Q10210, Invitrogen) and assessed the quality with the RNA integrity number (RIN) from the 2200 Tape Station (Agilent).

We prepared the cDNA libraries with the NEBNext® Ultra™ II Directional RNA Library Prep with Sample Purification Beads (E7765S, NEB) following the manufacturer’s protocol with 1 µg of total RNA per sample as input material. Briefly, we isolated messenger RNA (mRNA) from total RNA using the NEBNext® Poly(A) mRNA Magnetic Isolation Module kit (E7490S, NEB). We indexed each library with a combination of two of the NEBNext® Multiplex Oligos for Illumina® (Dual Index Primers Set 1) (E7600, NEB), quantified the libraries using the Qubit dsDNA assay kit, pooled all libraries using equimolar concentrations, and checked their quality with the 2200 Tape Station (Agilent). Finally, we sequenced the libraries at the University of Florida Interdisciplinary Center for Biotechnology Research on the Illumina HiSeq 3000 (PE-150 bp).

We received individual raw sequencing reads files for each sample and quality trimmed them by removing adaptors and low-quality bases with Trimmatic v 0.36 [5] with a 4-base sliding window and quality threshold of 25. The reads from each individual sample were pooled and used for de novo transcriptome reconstruction with Trinity v 2.8.4 [6], following the protocol by [7]. Then, we clustered highly redundant transcripts, i.e. transcripts with > 95% sequence similarity, using CD-HIT v 4.6.6 [8], and selected the longest isoform per gene using Trinity’s custom script (get_longest_isoform_seq_per_trinity_gene.pl).
2.4. De novo transcriptome quality assessment and annotation

We assessed the quality of the de novo transcriptome assembly by mapping the trimmed reads back to the assembly with Bowtie2 v 2.2.9 [9] and evaluated its completeness by searching for orthologues with BUSCO v 3.1.0 [10] using the viridiplantae-odb10 as a reference database (E-value cutoff for the blast alignments: 1e−08).

We used the pipeline available within the bioinformatics platform OmicsBox [11,12] to annotate the de novo transcriptome.

2.5. Data analysis

All analyses were performed in R v.3.5.1 [13] running under R Studio v.1.2.5019 [14].

First, we tested for genetic differentiation among populations of S. cataractae by performing an analysis of molecular variance (AMOVA) with the function poppr.amova in the poppr R package [15] using the SNP data. We assessed the significance of the model using a randomization test with 9999 permutations on the output of the AMOVA (function randtest from the ade4 package [16]).

Second, we tested the effect of the heavy metal treatments on DNA methylation by performing a distance-based redundancy analysis (dbRDA [17]) on the matrix of pairwise epigenetic distances created from the SMP dataset following the formula presented by [18]. We run two separate dbRDA models, one testing the effect of Cu (epigenetic distance ~ Population * Cu treatment) and one testing the effect of Cd (epigenetic distance ~ Population * Cd treatment).

Third, we identified differentially methylated cytosines positions (DMPs) as cytosines with significant differences in their mean methylation level between control and Cu-treated, and control and Cd-treated plants using the R package DSS [19]. First, we modelled the methylation frequency at each cytosine position within each group using a beta-binomial distribution with arcsine link function and the formula “~ 0 + group” (function DMLfit.multifactor). Then, we performed Wald tests to detect differential methylation between groups at each position using the function DMLtest.multifactor which reports adjusted p-values by the Benjamini-Hochberg method (i.e. FDR). We considered cytosines to be differentially methylated (i.e. DMPs) when FDR ≤ 0.05 and methylation change was ≥ 10%.

Fourth, we identified differentially expressed transcripts (DETs) between control and Cu-treated plants for each population as follows: (1) we estimated transcript abundance within each individual sample using RSEM [7] wrapped by scripts included in Trinity (align_and_estimate_abundance.pl); (2) we used edgeR v3.24.3 [20] to filter low count transcripts by removing those with less than 2 cpm (counts per million; i.e. less than 10 counts per transcript) in at least 3 samples; and (3) we fitted the model using the function glmQLFit (option robust = TRUE) and tested for DETs using the function glmTreat. Transcripts were considered differentially expressed when FDR < 0.001 and expression change was ≥ 4-fold (log2FC ≥ 2).

Finally, we performed a Fisher’s exact test with the FatiGO package [21] on the list of DETs from each population to find differences in the fraction of transcripts annotated with a specific Gene Ontology (GO) term between the DETs list and the full list of annotated transcripts (reference set). We applied an FDR cutoff of 0.01, and used the “Reduce to most specific” option within OmicsBox to remove the more general, less informative, GO terms.

Ethics Statements

This research did not involve human or animal subjects nor data collected from social media platforms.
Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data Availability

Raw sequencing data and barcodes file for epiGBS (Original data) (NCBI SRA).
Processed and analyzed data (Original data) (Zenodo).
epiGBS custom scripts (Original data) (Zenodo).

CRediT Author Statement

M. Teresa Boquete: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Writing – original draft, Writing – review & editing; Marc W. Schmid: Data curation, Formal analysis, Writing – review & editing; Niels C.A.M. Wagemaker: Methodology; Sarah B. Carey: Methodology, Writing – review & editing; Stuart F. McDaniel: Methodology, Writing – review & editing; Christina L. Richards: Conceptualization, Supervision, Investigation, Methodology, Writing – review & editing; Conchita Alonso: Conceptualization, Supervision, Investigation, Methodology, Writing – review & editing.

Acknowledgments

We thank Dr. Carlos M. Herrera for his advice throughout the development of the project; we also thank Dr. Jonathan A. Shaw and Dr. Blanka Aguero for their help with field sample localization and collection, Luiza Silva Simoes and Olivia Santiago for help during laboratory work at USF, and Dr. Mariano Alvarez, Dr. Marta Robertson, Dr. Sabrina McNew, Dr. Jeannie Mounger, and Sandra Voors for continuous discussions and feedback about the epiGBS data analyses. This research has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 704141-BryOomics. Dr. M. Teresa Boquete has been supported during the development of this project by the Juan de la Cierva-Incorporación program from the Spanish Ministry of Science, Innovation and Universities (IJC2018-035018-I) and is currently supported by the Maria Zambrano program from the Spanish Ministry of Science, Innovation and Universities.

Supplementary Materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2022.108710.

References

[1] M. Schmid, M.W.Schmid/epiGBS\_Nov\_2017\_fixed: epiGBS\_2017\_11\_fixed, (n.d.), doi:10.5281/zenodo.7040291.
[2] M.T. Boquete, M.W. Schmid, C.A.M. Wagemaker, S.B. Carey, S.F. McDaniel, C.L. Richards, C. Alonso, Non-genetically-based intraspecific differentiation for heavy metal tolerance in the copper moss Scopelophila cataractae, (2021).
[3] M.T. Boquete, I. Lang, M. Weidinger, C.L. Richards, C. Alonso, Patterns and mechanisms of heavy metal accumulation and tolerance in two terrestrial moss species with contrasting habitat specialization, Environ. Exp. Bot. 182 (2021) 104336, doi:10.1016/j.envexpbot.2020.104336.
[4] T.P. van Gurp, N.C.A.M. Wagemaker, B. Wouters, P. Vergeer, J.N.J. Ouborg, K.J.F. Verhoeven, epiGBS: reference-free reduced representation bisulfite sequencing, Nat. Methods 13 (2016) 322–324, doi:10.1038/nmeth.3763.
[5] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 214–2120, doi:10.1093/bioinformatics/btu170.

[6] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnikre, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, doi:10.1038/nbt.1883.

[7] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. LeDuc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–1512, doi:10.1038/nprot.2013.084.

[8] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152, doi:10.1093/bioinformatics/bts565.

[9] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (2012) 357–359, doi:10.1038/nmeth.1923.

[10] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.

[11] S. Götz, J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, M. Robles, M. Talón, J. Dopazo, A. Conesa, High-throughput functional annotation and data mining with the Blast2GO suite, Nucleic Acids Res. 36 (2008) 3420–3435, doi:10.1093/nar/gkn176.

[12] OmicsBox – Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019, (2019). https://www.biobam.com/omicsbox/.

[13] R Core TeamR: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2018 https://www.r-project.org/.

[14] RStudio TeamRStudio: Integrated Development for R, RStudio, PBC, Boston, MA, 2019 https://www.rstudio.com/.

[15] Z.N. Kamvar, J.F. Tabima, N.J. Grünwald, Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction, PeerJ 2 (2014) e281, doi:10.7717/peerj.281.

[16] S. Dray, A.-B. Dufour, The ade4 Package: implementing the duality diagram for ecologists, J Stat Softw 22 (2007) 1–20, doi:10.18637/jss.v022.i04.

[17] P. Legendre, M.J. Anderson, Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments, Ecol. Monogr 69 (1999) 1–24 10.1890/0012-9615(1999)069[0001:DBRATM]2.0.CO;2.

[18] Y. Wang, M. Qian, P. Ruan, A.E. Teschendorff, S. Wang, Detection of epigenetic field defects using a weighted epigenetic distance-based method, Nucleic. Acids. Res. 47 (2019) e6–e6, doi:10.1093/nar/gky882.

[19] H. Feng, K.N. Conneely, H. Wu, A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data, Nucleic. Acids. Res. 42 (2014) e69, doi:10.1093/nar/gku154.

[20] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (2010) 139–140, doi:10.1093/bioinformatics/btp616.

[21] F. Al-Shahrour, R. Díaz-Uriarte, J. Dopazo, FatigO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, Bioinformatics 20 (2004) 578–580, doi:10.1093/bioinformatics/btg455.