Antrag auf Förderung im Rahmen der Ausschreibung “Forschungsstart” der Carl-Zeiss-Stiftung

Projekt “*EasyVectorOmics*”

Prof. Dr. Asis Hallab, Technische Hochschule Bingen, Mai 2023

1. Stammdaten

a. Daten zum/zur Antragsteller:in

- Vor- und Familienname: Asis Hallab
- Institution: Technische Hochschule Bingen
- ggf. private Kontaktdaten (sofern präferierter Korrespondenzweg)
  Bitte verwenden Sie meine Kontaktdaten an der Technischen Hochschule Bingen
- Geschlecht (m/w/d) männlich

b. Bezeichnung des Forschungsvorhabens/Antragstitel sowie Kurztitel bzw. Akronym (max. 30 Zeichen)

EasyVectorOmics

c. Benennung der Disziplinen, in denen das Forschungsvorhaben angesiedelt ist (nach DFG-Fachsystematik)

201-07 Bioinformatik und Theoretische Biologie

d. Allgemeinverständliche Kurzbeschreibung des Vorhabens für die Verwendung in der Öffentlichkeitsarbeit (max. 250 Zeichen).

EasyVectorOmics quantifiziert Genexpression mittels effizienter Vektoranalyse, um verwandte Gene zu vergleichen und somit unbekannte Gen-Funktionen zu finden. Ein intuitives grafisches Interface im Browser ermöglicht die Analyse und Visualisierung.

e. Finanzdaten: beantragte Mittel bei der Carl-Zeiss-Stiftung (Gesamt-, Personal-, Sachkosten und Investitionen)

f. Beantragter Förderzeitraum (Start- und Enddatum)

1. November 2023 bis 31. Oktober 2025

2. Abstract

Die Quantifizierung der Genexpression ist ein wesentliches Instrument in den Biowissenschaften, das z. B. zur Bewertung der Trockenheitsresistenz wichtiger Nutzpflanzen [1] oder zum Verständnis von Krebserkrankungen [2] eingesetzt wird. Um herauszufinden, welche molekularen Funktionen bei einer genetischen Reaktion entscheidend sind, werden hierbei Gene identifiziert, die eine signifikante Zunahme oder Abnahme der Expression aufweisen, wenn man z. B. gesundes mit krankem Gewebe oder Pflanzen, die unter idealer Bewässerung gewachsen sind, mit Pflanzen vergleicht, die unter Trockenstress gewachsen sind. So werden wichtige genetische Reaktionen oder Profile als signifikante Veränderungen in der Expression identifiziert.

Unsere Analyse des Genoms des behaarten Schaumkrauts [3] hat gezeigt, dass eine Vektoranalyse der Genexpression, im Kontext von Genfamilien, wichtige Einblicke in die Evolution neuer Funktionen liefert. Mit Hilfe dieser Analyse wurden z.B. wichtige Transkriptionsfaktoren gefunden, die eine Schlüsselrolle bei der Ausbildung bestimmter Pflanzenmerkmale wie Blatt-Ausbuchtungen oder faltiger Samenoberfläche spielen. Das Beispiel des Genomprojekts des behaarten Schaumkrauts zeigt, dass meine neuartige Vektoranalysemethode für den Vergleich der Genexpression zwischen verwandten Genen, also Genfamilien, geeignet ist.

Basierend auf den Erfahrungen aus interdisziplinären Pflanzenzüchtungsprojekten haben wir den Gene Expression Plotter (GXP) [3] entwickelt, der es Wissenschaftlern ermöglicht, die Genexpression auf einfache Weise zu analysieren und zu visualisieren, indem er eine Vielzahl verschiedener interaktiver wissenschaftlicher Plots zur Unterstützung der Hypothesenbildung und -überprüfung bietet.

EasyVectorOmics wird (a) die für die vergleichende Bitterkresse-Genomanalyse entwickelte Methode verallgemeinern und mit Hilfe effizienter linearer Algebra-Algorithmen für den Einsatz in vergleichenden Genomik-Projekten implementieren. Das Tool wird in den Gene Expression Plotter mit einer intuitiven Benutzeroberfläche integriert und visualisiert die Ergebnisse interaktiv, um detaillierte Informationen über die Ergebnisse zu liefern. Eine Kommandozeilenschnittstelle wird auch für Hochdurchsatz-Analyse-Pipelines verfügbar sein.

In der Region Bingen gibt es mehrere pharmazeutische Unternehmen und Forschungskliniken, deren Projekte von den in EasyVectorOmics implementierten neuen Methoden profitieren und durch Kollaboration zur Weiterentwicklung führen können.

References:

[1] López, C. M., Alseekh, S., Torralbo, F., Martínez Rivas, F. J., Fernie, A. R., Amil-Ruiz, F., & Alamillo, J. M. (2023). Transcriptomic and metabolomic analysis reveals that symbiotic nitrogen fixation enhances drought resistance in common bean. *Journal of Experimental Botany*, erad083.

[2] Lovero, D., D'Oronzo, S., Palmirotta, R., Cafforio, P., Brown, J., Wood, S., ... & Silvestris, F. (2022). Correlation between targeted RNAseq signature of breast cancer CTCs and onset of bone-only metastases. British Journal of Cancer, 126(3), 419-429.

[3] Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R. D., … **Hallab, A.** ... & Tsiantis, M. (2016). The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nature Plants, 2(11), 1-7.

[4] Eiteneuer, C., Velasco, D., Atemia, J., Wang, D., Schwacke, R., Wahl, V., ... & **Hallab, A.** (2022). GXP: Analyze and plot plant omics data in web browsers. Plants, 11(6), 745.

3. Zielsetzung des Forschungsvorhabens

a. Ausgangssituation und Motivation, ggf. Beschreibung der wissenschaftlichen Vorarbeiten

Genetic responses and genetic profiles are widely studied in a variety of life sciences research projects, e.g. in agricultural sciences to understand how a plant crop responds to stresses like drought [1], lack of nutrients, or soil contaminants, or in cancer research to understand key molecular functions employed by malignant cells to promote uncontrolled growth [2]. A collection of methods, software tools, and pipelines [3-7] exist to efficiently process the results of sequencing respective RNA samples, quantify gene expression, and compare control with stressed or diseased samples with the goal of identifying differentially expressed genes, i.e. those genes whose expression either significantly increases or decreases. While these tools efficiently assess changes in expression of a single gene, the analysis of gene expression in the context of related genes, i.e. gene families, has not yet been developed. In our study of the hairy bittercress model plant genome [8] I developed a vector analysis method that identifies significant changes of gene expression within a family of related genes. We used these results to find genes that are key in the formation of plant traits like leaf-form (lobe formation) or wrinkled seed surface. The vector-space analysis was developed to overcome the limitation of linear model based or negative binomial distribution based analysis of gene expression, which are not suitable for the comparison of different genes, but are used to compare

expression of any gene in different tissues (like diseased and healthy) or under different conditions (like optimal watering and drought stress). In this vector analysis normalized quantified gene expression is projected into a vector space, where each axis represents quantified gene expression within a sampled tissue or condition. Subsequently, The expression of genes belonging to a family is analyzed and deviations from related genes are detected. In doing so, and integrating results from comparative genomics the functions of duplicated genes in terms of changes in gene expression can be efficiently identified. For the above results in the hairy bittercress genome for example, we measured gene expression within five different tissues (root, stem, leaves, and developing seedling) both sampling hairy bittercress and its close relative *Arabidopsis thaliana*. Subsequently we analyzed gene families with genes found to be unique (duplicated) in the hairy bittercress genome, many of which conserved the molecular function with their relatives, but showed differential expression in different tissues. We analyzed euclidean distances and angles to identify genes who significantly changed their expression and switched tissues.

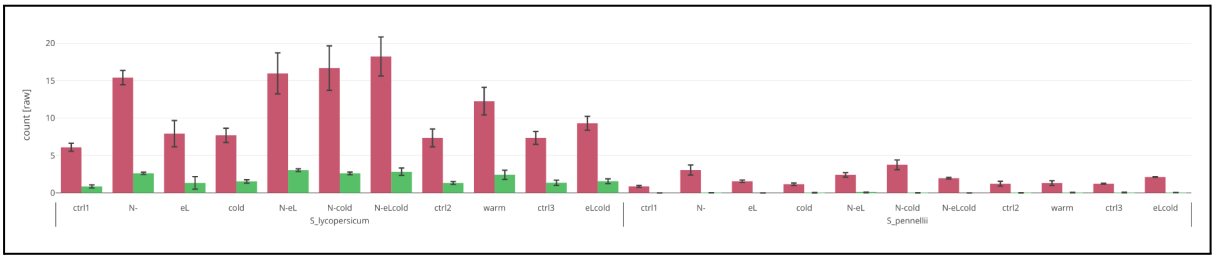
In another interdisciplinary project we analyzed gene expression in tomato and potato, and found that in order to enable our geneticist and developmental biologist partners to analyze the results of our quantification and differential gene expression analysis they needed to be able to quickly browse the results for a set of candidate genes of their interest. Here, different visualizations were required to explore expression of such genes of interest (see figure 1) or compare expression profiles of different samples e.g. with a scatter plot of principal components or a heatmap plot visualizing correlation between samples (see figure 2). We thus developed the browser based omics analysis and visualization tool “Gene Expression Plotter” (GXP) [9].

The goal of EasyVectorOmics is to develop and implement the vector based analysis of quantitative gene expression and other quantitative omics data to be used in life sciences projects e.g. to identify interesting drug targets or understand diseased tissue in greater depth. Currently, to our knowledge, there is no method available to compare expression of related genes, particularly with the focus on significant change in expression comparing related genes.

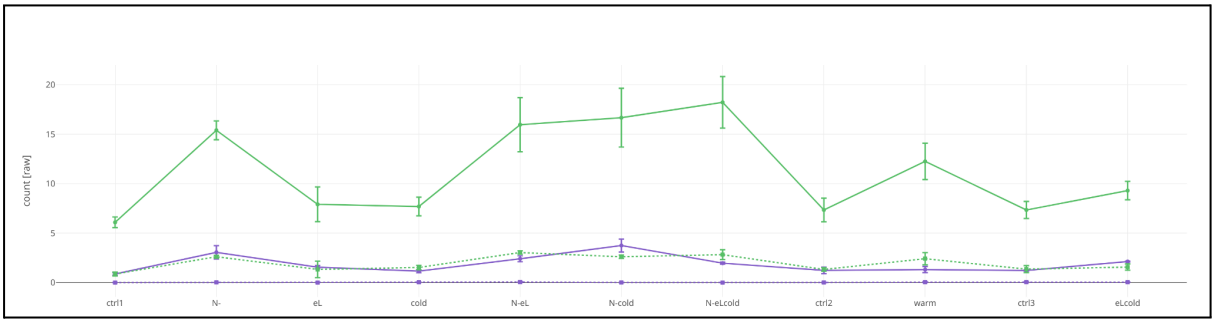
The method is to be tested and its efficiency, accuracy, and usability validated using carefully selected public reference data from plant and animal models. Subsequently the prototype is implemented in a high performance compiled programming language like Rust, C/C++, or Fortran. This enables the usage in high performance and high throughput analysis as typically carried out e.g. in medical or pharmaceutical context. This code will then be compiled (translated) to Web Assembly [10] which enables its integration into GXP and the browser where specific visualizations of the results will be included.

To get user feedback and enhance the spectrum of results the tool then will be disseminated with regional companies and clinics, e.g. the Universitätsklinik Mainz where a student of mine is analyzing gene expression in cancer patients. This will enable the future development of more methods and interactive visualizations of interdisciplinary quantitative omics data.

A)



B)



C)

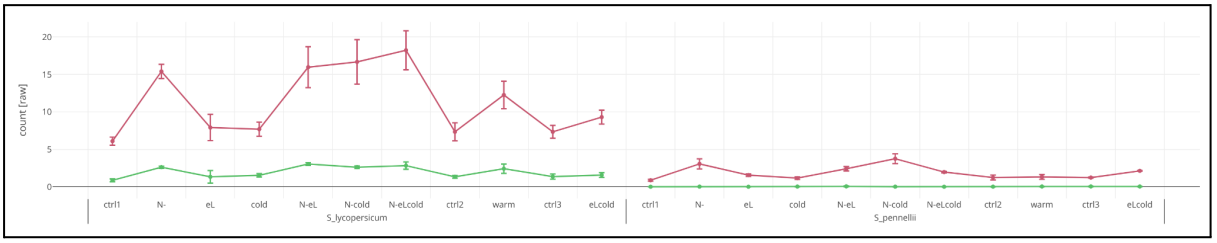
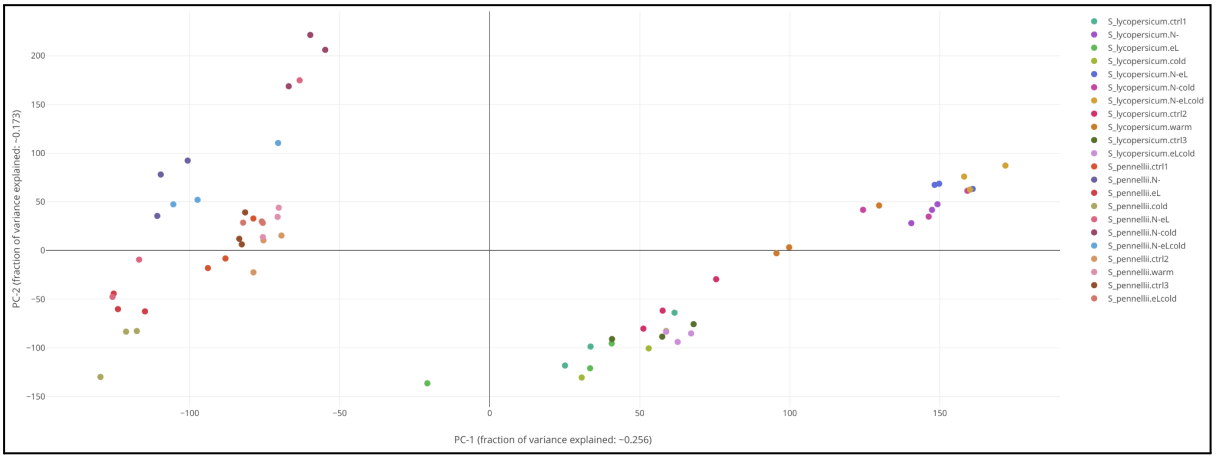


Figure 1: Example interactive plot in our tool “Gene Expression Plotter” (GXP) Comparing gene expression of two tomato genes in different stress conditions (cold, heat, intensive light, nitrogen deficient soil, and combinations thereof) [9]. The plot is displayed in the browser and interactive, distribution statistic values are shown when hovering over points with the mouse, zooming is supported, and genes can be switched off or on. A) is a barplot, B) a stacked line plot, and C) a separate line plot.

A)



B)

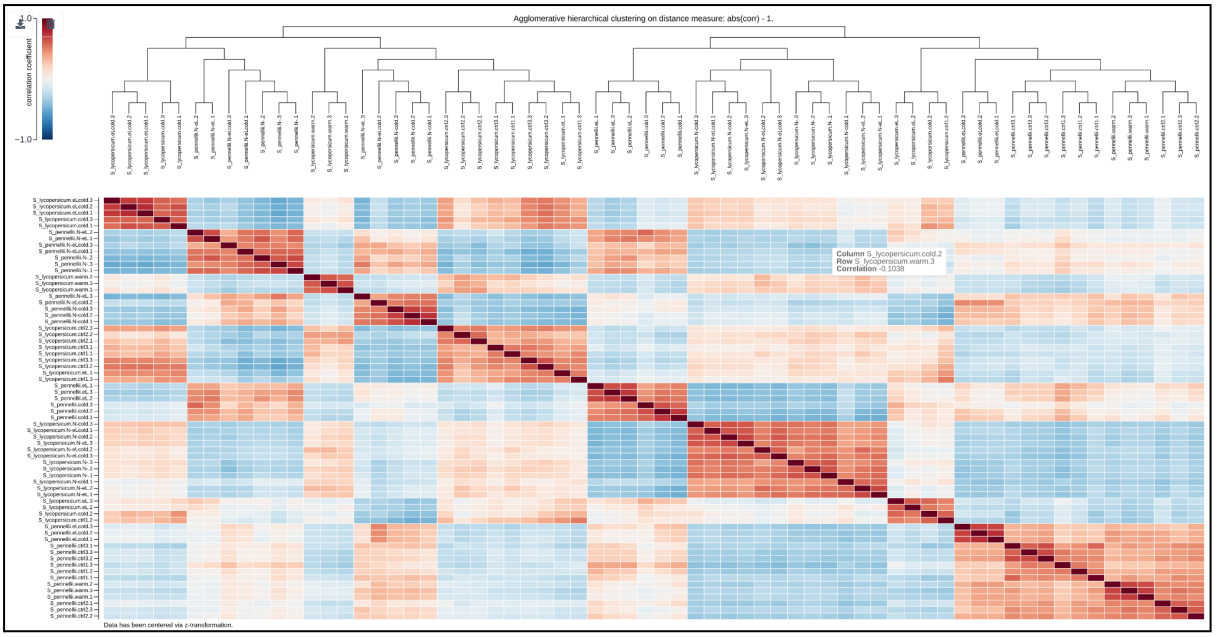


Figure 2: Example interactive plot in our tool “Gene Expression Plotter” (GXP) comparing gene expression profiles of different tissue samples taken from plants subjected to different forms of stress [9]. The interactive plots enable the scientist to assess how alike or dislike the plant genetic responses to these stresses are. The plot is interactive, supports zoom, investigation of values plotted, and switching on and off samples. A) shows a scatter plot of principal components, and B) shows a heatmap and correlation based clustering result.

b. Angestrebte wissenschaftliche Ziele und Alleinstellungsmerkmale des beantragten Vorhabens

The goal of EasyVectorOmics is to mature, test, and extend a novel vector algebra based analysis developed during a comparative genomics research project on hairy bittercress [8]. The method efficiently compares quantitative gene expression data between related genes, i.e. within gene families, and identifies for example genes that evolved recently and whose expression play a key causal role in traits of interest. To our knowledge such a method to efficiently and robustly explore quantitative omic data like gene expression data does not exist for the comparison of related genes.

First the method is to be extracted from the scripts developed for the hairy bittercress genome project, prototyped in a scripting programming language, and tested. Suitable test data is selected from animal (human) and plant research. The method is thus tested on a wider range of organisms.

Next, the method is implemented in a high performance compiled language like C/C++, Rust, or Fortran so it can be both used in high performance and high throughput research pipelines. It will be compiled to WebAssembly and integrated in our tool “Gene Expression Plotter” (GXP) with an intuitive graphical user interface

and specific interactive visualizations of the results, thus enabling the use by non informatician scientists.

c. Kurze Darstellung, wie die Mittel der Carl-Zeiss-Stiftung zur Entlastung der/des Neuberufenen genutzt werden sollen, insbesondere wenn der/die Antragstellende bereits eine Forschungsprofessur mit reduziertem Lehrdeputat innehat.

The funding will be used to employ a scientific programmer to carry out the above analyses and implementation work. A reduction of teaching load is not planned.

References:

[1] López, C. M., Alseekh, S., Torralbo, F., Martínez Rivas, F. J., Fernie, A. R., Amil-Ruiz, F., & Alamillo, J. M. (2023). Transcriptomic and metabolomic analysis reveals that symbiotic nitrogen fixation enhances drought resistance in common bean. *Journal of Experimental Botany*, erad083.

[2] Lovero, D., D'Oronzo, S., Palmirotta, R., Cafforio, P., Brown, J., Wood, S., ... & Silvestris, F. (2022). Correlation between targeted RNAseq signature of breast cancer CTCs and onset of bone-only metastases. British Journal of Cancer, 126(3), 419-429.

[3] Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357-359.
[4] Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nature methods, 12(4), 357-360.

[5] Van De Geijn, B., McVicker, G., Gilad, Y., & Pritchard, J. K. (2015). WASP: allele-specific software for robust molecular quantitative trait locus discovery. Nature methods, 12(11), 1061-1063.

[6] Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14(4), 417-419.

[7] Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology, 15(2), 1-17.

[8] Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R. D., … **Hallab, A.** ... & Tsiantis, M. (2016). The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nature Plants, 2(11), 1-7.

[9] Eiteneuer, C., Velasco, D., Atemia, J., Wang, D., Schwacke, R., Wahl, V., ... & **Hallab, A.** (2022). GXP: Analyze and plot plant omics data in web browsers. Plants, 11(6), 745.

[10] Haas, A., Rossberg, A., Schuff, D. L., Titzer, B. L., Holman, M., Gohman, D., ... & Bastien, J. F. (2017, June). Bringing the web up to speed with WebAssembly. In Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation (pp. 185-200).

4. Forschungsvision der/des Antragstellenden

a. Darstellung der Forschungsvision bzw. des langfristigen Forschungsziels der/des Antragsteller:in

Modern sequencing and high performance chromatography technologies enable life science research to integrate interdisciplinary assessments of an organism's state and response to stimuli like stress or disease. In this, each method is capable of answering different questions, genome sequencing and comparative genomics identify key genes or families that played important roles in the evolution of the studied species [1], genomic variants of patients that develop a disease and those who do not in spite of infection can identify alleles crucial for immunity [2] and thus open routes to the development of new treatments. Gene expression profiles, i.e. transcriptome sequencing or transcriptomics, e.g. enable the assessment of success of chemotherapy on treated cancers in that treatment returns the tissue to healthy gene expression profiles [3]. The analysis of metabolites of important crops like maize show how plants react to stress and how this may affect the chemical composition of the harvest, especially in terms of key nutritional compounds [4]. Importantly the interdisciplinary integration of these different research methods yields more results than the simple sum of the respective parts. For example, does the integration of genomic variant analysis with metabolomics under drought stress conditions in maize reveal novel variants [5] that can be used for future plant breeding directed at providing crop variants more resistant to the expected challenges imposed by global climate change.
In many of these interdisciplinary research projects the multi omics data is integrated in a somewhat manual approach. This naturally is time consuming and lacks the benefits of a standardized, well tested and robust method yielding reproducibility of results. Furthermore, interdisciplinary research projects require that the results of one expert group are explored by another group of different expertise. Thus, easy to use interactive tools are required that support interactive visualization, contextualization, and on the fly analysis, preferably without the need to install any software.
It is my vision to develop novel methods to carry out integrated multi-omics analyses, and implement them in the form of highly usable software tools. The experience that

led me to the development of our “Gene Expression Plotter” (GXP) tool for the on the fly analysis and interactive visualization of multi-omics data showed that lessons learned from professional software engineering are best carried over into the development of such tools. The scientists that will be using the tool are to be regularly queried about usage experience, new feature requests, and novel ideas for interactive visualizations, that only come to light while using a tool and wishing for features not yet implemented. My vision is to continue generating novel tools for comparative genomics and multi omics analyses, apply high standard software engineering standards, and make the resulting tools available both for the use in high throughput pipelines as well as in intuitive graphical applications like our Gene Expression Plotter.

b. Zusammenfassung, welche Rolle das beantragte Vorhaben bei der Verfolgung dieser Forschungsvision spielt

EasyVectorOmics will implement the first step in realizing the above vision. We will extract existing source code of a method I successfully developed and applied in the high impact hairy bittercress genome project, prototype it, identify suitable test data for verification of broader usability, implement a high performance version using compiled programming languages for usage in high throughput pipelines, and finally integrate EasyVectorOmics into our browser based Gene Expression Plotter with an intuitive interface and interactive visualization of results.
Because the method has been used successfully in a genome project and exists in a first version the likelihood of failure of EasyVectorOmics is low. The project thus provides an excellent first step for launching new research in the department of applied bioinformatics at the university of applied sciences in Bingen.
Furthermore, already ongoing student projects and Bachelor thesis research in transcriptome analysis of disease integrates well with the researched topics and enables the establishment of an interdisciplinary network for the future development of novel methods, tools, and interactive visualizations.

c. Konkrete Ansätze für die Einwerbung weiterer Drittmittel und Darstellung von relevanten weiteren Forschungsfragen, die sich aus dem Antrag ergeben können

I received 337,874.00 Euro funding (grant number 031B0921) from the Federal Ministry of Education and Research (BMBF) for my ongoing international research project on wild maize relatives with partners from Mexico and the institute for bioinformatics at the Forschungszentrum in Jülich. In it, we analyze a large collection of different wild maize taxa with an multi-omics approach to identify key genetic variants that provide the basis for future maize breeding directed at providing maize variants suited to withstand the challenges resulting from climate change.
The project is entering its midterm and we will soon be analyzing the genome and transcriptome data produced in the first third of the project’s duration. We are

currently preparing the publication of a genomic variant analysis in which we integrated population genetics and protein function analysis to identify how climatic growing conditions drove genome evolution and selected certain plant traits. In the upcoming comparative multi omics analysis on full genome and transcriptome data we will develop new prototypic methods to understand, integrate, and analyze the data, very much alike to the hairy bittercress genome project. Resulting from these experiences new questions and methods for the analysis of integrated genomic variant and phenotypic data will arise, which will become the basis for future research of new methods that already have a working prototype from an existing multi omics research project.
Based upon our current results we plan to submit a new research proposal for continuation of maize research using novel methods for integrated multi omics analysis. A possible platform will be the Bioeconomy international platform from the Federal Ministry of Education and Research (BMBF).

References:

[1] Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R. D., … **Hallab, A.** ... & Tsiantis, M. (2016). The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nature Plants, 2(11), 1-7.

[2] Díez-Fuertes, F., De La Torre-Tarazona, H. E., Calonge, E., Pernas, M., Bermejo, M., García-Pérez, J., ... & Alcamí, J. (2020). Association of a single nucleotide polymorphism in the ubxn6 gene with long-term non-progression phenotype in HIV-positive individuals. Clinical Microbiology and Infection, 26(1), 107-114.

[3] Charles, M. P., Therese, S., Michael, B. E., Matt, V. D. R., Stefanie, S. J., Christian, A. R., ... & David, B. (2000). Molecular portraits of human breast tumours. Nature, 406(6797), 747-752.

[4] Alvarez, S., Marsh, E. L., Schroeder, S. G., & Schachtman, D. P. (2008). Metabolomic and proteomic changes in the xylem sap of maize under drought. Plant, Cell & Environment, 31(3), 325-340.

[5] Zhang, F., Wu, J., Sade, N., Wu, S., Egbaria, A., Fernie, A. R., ... & Dai, M. (2021). Genomic basis underlying the metabolome-mediated drought adaptation of maize. Genome Biology, 22(1), 1-26.

5. Transfer und Vernetzung

a. Potentiale für eine Vernetzung mit Wirtschaft, insbesondere in der Region der Hochschule

In the region of Bingen we find several companies carrying out pharmaceutical and medical research, most notably Bioscientia, analyzing patient samples among others with sequencing methods, Biontech, who developed a most efficient vaccination against the SARS-CoV-2 virus, and Boehringer which is the biggest researching pharmaceutical company in Germany.
All these companies apply interdisciplinary multi omics research to find new treatments and treatment targets. Thus, new opportunities arise to jointly identify requirements, find solutions and new methods, and implement them with feedback from scientific users within the context of industrial research and development.

b. Soweit bereits absehbar: Zusammenfassung der Anknüpfungspunkte für eine Vernetzung innerhalb der Hochschule (Kooperation mit anderen Professuren und/oder Fachbereichen)

Currently, I collaborate with Prof. Dr. Maik Lehmann and Prof. Dr. Kerstin Troidl in a research project in which we assess mutation rates and the effect of identified mutations on viral proteins of SARS-CoV-2 populations grown in cell culture and sequenced in regular intervals. In this, we developed a pipeline to identify genomic variants by aligning reads produced by Illumina sequencing to the Wuhan reference genome and subsequently analyzed the potential effect of these mutations on viral protein function. In this, a Bachelor thesis project applied deep learning protein tertiary structure prediction to find proteins that show a tertiary structure significantly different to the reference.

Similar interdisciplinary multi omics projects are expected to arise in the future in which the EasyVectorOmics method or related analyses will be applied in collaborative research projects.

6. Arbeitsprogramm

Work package one (WP 1) - Prototyping and reproduction of previous results

In the hairy bittercress comparative genomics research project I developed a vector algebra based method to compare data on quantitative gene expression of related

proteins (gene families). The method successfully identified genes linked with key plant traits of interest like leaf form and seed surface.
This vector algebra analysis was implemented in R in a package developed for the analysis for the hairy bittercress genome. The package is tightly integrated with project specific code and data.

(Milestone 1.1) The vector algebra code is to be extracted and repackaged in a generalized module capable of being used with any comparative genomics data. The code is developed using standard software engineering paradigms like test driven development in which each software component is checked whether it produces the expected results and reacts correctly for unexpected input values (edge cases).

Different vector algebraic analysis functions for quantitative transcriptomic data are to be implemented or extracted from the existing source code.

(a) First data normalization, which is implemented as z-transformation (centering by subtracting gene mean expression and subsequent scaling by division with gene expression standard deviation). Optionally normalization involves logarithmic scaling (the default). Typically three biological replicates are sequenced to account for biological noise. The mean values between these replicates are used as a single representative of gene expression in each sampled tissue or experimental condition, respectively. Thus, for each gene a single normalized gene expression vector is generated, in which each entry shows this gene's mean expression in the respective experimental condition (e.g. drought treated optimally watered plants) or type of tissue (e.g. cancer or healthy) that vector entry refers to.

(b) Next, normalized gene expression of related genes is projected into n-dimensional euclidean vector space. Each tissue sample or experimental condition (see a) is used as a single axis. Thus genes point to a certain position in a multidimensional vector space.

(c) For each group of related genes several vector algebraic tests are carried out. E.g. a gene family is separated into conserved (ancestral) genes and derived (duplicated) related genes. Subsequently, tests are carried out whether the derived genes occupy a different region in expression vector space or not (see figure 3.A).
In doing so, changes in tissue specificity or tissue preference can be assessed. The details are explained in the caption of figure 3.

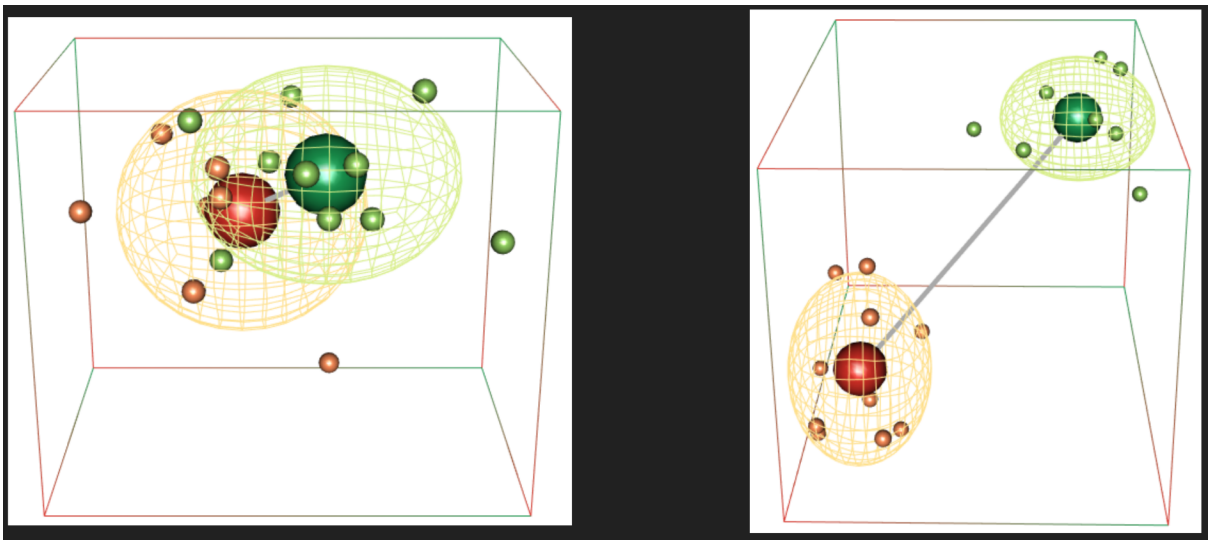
(d) Program a function to assess significant distances in gene expression space (see figure 3.A).

(e) Program a function to assess significant changes in tissue specificity (see figure 3.B).

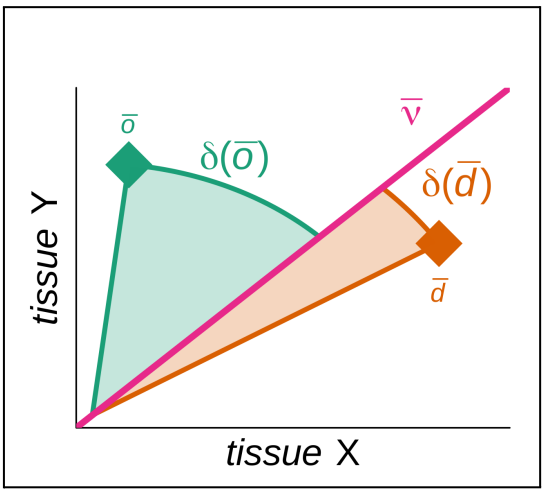
(f) Program a function to assess significant changes in tissue preference (see figure 3.C).

(Milestone 1.2) The so produced generalized module is tested with the data published with the hairy cress genome in order to ensure reproducibility of results.

A)



B)



C)

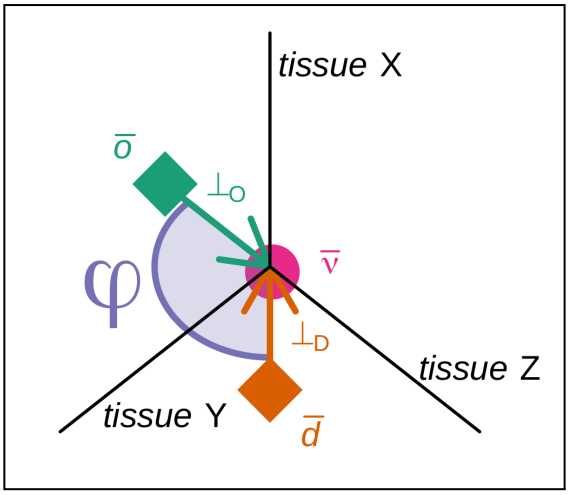


Figure 3: Different vector space algebra based analyses carried out on gene expression data. Note that each point in the multi-dimensional expression space represents a gene. Axis represent tissue samples or experimental conditions. A) shows two scenarios of gene expression within a given gene family. In the first case the conserved (green) members occupy the same expression space as do the derived (duplicated, orange) genes. No significant difference is discernible between the two group's respective mean expressions (bigger spheres). The difference is not significant because the regions covered by the respective standard deviations overlap. The second example shows a significant deviation of derived (duplicated) gene expression from the ancestral one. In this the gene expression regions covered by the respective ancestral and duplicated standard deviations do not overlap. B) shows a two dimensional simple example of how tissue specificity of gene expression is assessed in terms of the angle (delta) to the diagonal (*v*). The bigger this angle (max 90 degrees) the more specific this gene's expression only occurs within a given tissue. The smaller the angle the more equal a gene's expression in both tissues. C) shows how with maintained general tissue specificity a switch between tissues is measured. A three dimensional expression plot is shown in which the observer looks down upon the coordinate system along the diagonal (represented by the central pink circle *v*). Both ancestral and duplicated genes show the same degree of tissue specificity (angle to *v*), however there is a switch to in which tissues the respective genes are expressed. Ancestral genes are expressed equally in tissues X and Y, while duplicated genes are no longer expressed in tissue X but are equally expressed in tissues Y and Z. This switch in tissue preference is assessed by measuring the angle phi which is the amount of rotation around the central diagonal (v) from one gene expression vector (green) to the other (orange).

Work package two (WP 2) - Identification of suitable test data

(Milestone 2.1) Public data stemming from scientific articles about transcriptomic research of human disease, e.g. cancer or viral infection, is searched for a suitable test dataset. In this, the data preferably comprises a large sample size and covers several experimental conditions and tissue samples. The literature is searched for

results on non single copy genes, i.e. genes from gene families, in which one member displays a significantly different expression pattern linked with a different genetic response and thus function. Preferably, these proteins still maintain the molecular function of their close relatives.

(Milestone 2.2) The test-data from Milestone 2.1 is used with the prototype implementation (Milestone 1.1) to assess how well the method is usable within the context of transcriptomics research projects. The published results are used to assess the type one and type two error rates, i.e. to measure sensitivity, specificity, and thus assess accuracy and robustness of the method.

Work package three (WP 3) - Reimplementation in an efficient compiled programming language

(Milestone 3.1) The prototype implementation of the method developed and tested in work packages one and two is now reimplemented in a highly efficient programming language like Rust, C/C++, or Fortran. Popular and efficient linear algebra programming libraries are used like the GNU scientific library, Boost, LAPACK or similar. The code is developed using current software engineering quality standards like test driven development (see milestone 1.1). The tests from milestone 2.2 are integrated in the automated tests to ensure correctness and robustness of the new implementations. A well documented (man-entry, see link 1 below) and GNU Linux standard command line interface (see link 2 below) is developed so that the analysis can be used in high throughput multi omics pipelines.

Links:
[1] https://pubs.opengroup.org/onlinepubs/9699919799/utilities/man.html
[2] https://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap12.html

Work package four (WP 4) - Integration of EasyVectorOmics into Gene Expression Plotter

(Milestone 4.1) The efficient implementation from milestone 3.1 is compiled to WebAssembly and integrated into our Gene Expression Plotter browser based analysis and visualization program (see link 1 below) as another analysis tool. An intuitive graphical interface is developed and specific interactive plots to visualize the results are integrated.

Links:
[1] https://usadellab.github.io/GeneExpressionPlots

7. Finanzierung

a. Grundausstattung der Hochschule

Die technische Hochschule Bingen hat zwei compute server zur Verfügung, die von ihrer Rechenkapazität ausreichen, um die Berechnungen, das Testen auf realen Daten, und die Entwicklung von EasyVectorOmics durchzuführen. Eine Kollaboration mit der Uni Mainz ermöglicht die Nutzung des Netzwerks aus Großrechnern der Uni Mainz.

Folgende compute server hat die angewandte Bioinformatik an der technischen Hochschule Bingen zur Verfügung:

- Dell PowerEdge R750 (Bestelldatum 28.10.2022)
    - 2* Intel Xeon Gold 6338 (jeweils 16 Cores/32 Threads)
    - 16* 64GB RDIMM, 3200MT/s (Insgesamt 1 TB RAM)
    - 2* M.2 480GB (RAID 1)
    - 5* 2,4TB 10K SAS
- Dell PowerEdge R710 (Bestelldatum 22.02.2011):
    - 2* Intel Xeon X5680 (jeweils 6 Cores/12 Threads)
    - 10* 8GB RAM
    - 6* 500GB DATA 7.2k 3,5" HD Hot Plug

b. Beantragte Förderung

Ein wissenschaftlicher Programmierer (“Techniker”) mit mindestens einem Bachelor of Science oder Master of Science Abschluss in Bioinformatik, Informatik, oder einer verwandten Disziplin, soll für die Laufzeit des Projekts beschäftigt werden, um die Programmierarbeiten gemäß des Arbeitsplans unter meiner Anleitung durchzuführen. Dies beinhaltet auch das Testen der Methode mittels publizierter und erwarteter Ergebnisse. Eine wissenschaftliche Hilfskraft (“HiWi”) soll für Organisatorisches und Unterstützung bei der Entwicklung und zum Testen in Teilzeit beschäftigt werden. Der wissenschaftliche Programmierer braucht einen Arbeitscomputer mit entsprechender Ausstattung (“Laptop, etc.”).

Dies ergibt Personal- und Arbeitsmaterial-kosten wie folgt:

	2023	2024	2025
Techniker (**E-11**)	11.000 €	68.000 €	59.000 €
HiWi	0 €	4.000 €	3.000 €
Laptop, etc.	5.000 €	0 €	0€

Anhang

Rekrutierungskonzept

Eine internationale Ausschreibung wird publiziert mit dem Ziel, einen wissenschaftlichen Programmierer mit fließenden Englischkenntnissen zu finden, der Erfahrung in der Programmierung von Vektoranalyse hat. Der/die Kandidat-in soll Kenntnisse im Umgang mit GNU/Linux Betriebssystem und der Nutzung von Großrechenanlagen haben. Der Bewerber muss Kenntnisse in der wissenschaftlichen Programmiersprache *R* (für das Arbeitspaket eins), sowie einer Hochsprache wie Rust, C/C++, oder Fortran (für das Arbeitspaket drei), und Kenntnisse in Javascript- und HTML/CSS-Entwicklung (für das Arbeitspaket vier) haben, um die Integration in Gene Expression Plotter und die Entwicklung eines intuitiven Benutzerinterface und geeigneter interaktiver wissenschaftlicher Plots umsetzen zu können. Erwartete Softskills umfassen gute Teamfähigkeit und eine gute Selbstorganisation bzw. selbstständiges Arbeiten.
Die Suche wird bewusst international durchgeführt, um die Chance einen geeigneten Kandidaten trotz Fachkräftemangel zu finden. Mit Erfolg haben wir über internationale Ausschreibungen schon in der Vergangenheit gute Mitarbeiter gefunden.

Weitere Dokumente

Bitte finden Sie folgend angehängt:

- Lebenslauf von Prof. Dr. Asis Hallab
- Meilensteinplan
- Finanzierungsplan