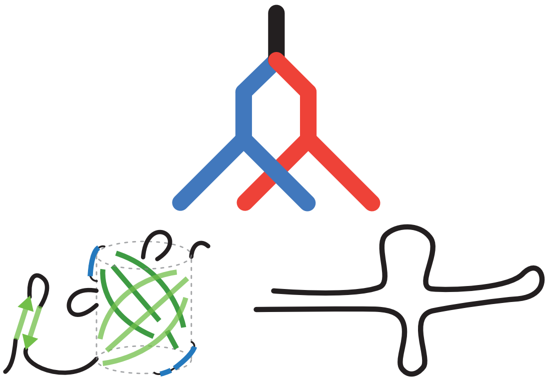
Applications of Comparative Genomics:

Dissemination and Phylogeny of Coding and Non-Coding Gene Families



DISSERTATION

zur

Erlangung des Doktorgrades

der Naturwissenschaften

(Dr. rer. nat.)

Dem Fachbereich Pharmazie

der Philipps-Universität Marburg

vorgelegt von

M. Sc. **Paul Moritz Johannes Klemm**

aus Berlin

Marburg an der Lahn im August 2023

Die Untersuchungen zur vorliegenden Arbeit wurden von Mai 2018 bis Juli 2023 unter der Betreuung von Prof. Dr. Roland Hartmann und Dr. Marcus Lechner in Marburg im Institut für Pharmazeutische Chemie und im Zentrum für Synthetische Mikrobiologie (SYNMIKRO) durchgeführt.

Erstgutachter/in: **Prof. Dr. Roland Hartmann**
Zweitgutachter/in: **Dr. Marcus Lechner**

Eingereicht am 8.8.2023
Tag der mündlichen Prüfung am 19.9.2023
Hochschulkennziffer: 1180

Erklärung

Ich versichere, dass ich meine Dissertation

“*Applications of Comparative Genomics: Dissemination and Phylogeny of Coding and Non-Coding Gene Families*”

selbständig ohne unerlaubte Hilfe angefertigt und mich dabei keiner anderen als der von mir ausdrücklich bezeichneten Quellen bedient habe. Alle vollständig oder sinngemäß übernommenen Zitate sind als solche gekennzeichnet.

Die Dissertation wurde in der jetzigen oder einer ähnlichen Form noch bei keiner anderen Hochschule eingereicht und hat noch keinen sonstigen Prüfungszwecken gedient.

Marburg, den ....

....

(Paul Klemm)

Abstract

Comparative genomics is an interdisciplinary field of study comparing the genetic makeup across multiple species. It aims to understand the similarities and differences in the genomes of various organisms to gain insights into their evolutionary relationships, functional characteristics, and adaptations. Some of the key applications of comparative genomics include phylogenetic reconstruction, where researchers construct evolutionary trees to visualize the evolutionary history of species or genes, and orthology predictions, where homologous genes with shared ancestry and similar functions are identified across different organisms. The highlighted work includes two biologically motivated projects that leverage bioinformatic tools from comparative genomics. Furthermore, advancements in sequencing technologies have revolutionized genomics by generating vast amounts of genomic data. On the one side, this data flood provides unprecedented opportunities for comparative genomics, allowing researchers to explore genomic diversity on a large scale. However, the sheer volume of data also poses significant challenges on the other side in terms of data processing and storage. The third project addresses this challenge of coping with the ever-increasing flood of genomic data by revising a critical tool of the field.

In the first project was focused on investigating the Kiwellin protein family in plants, which plays a critical role in plant-pathogen interactions. The research aimed to understand the structural features of this protein family and distinguish it from closely related Barwin-like proteins. The outcomes of this project were published in the article titled "*Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family*", introducing a systematic nomenclature that revealed three distinct sub-classes within the Kiwellin family. Additionally, a meta-analysis of publicly available transcriptome data revealed specific responses of Kiwellins in different plant tissues and cultivars, as well as their responses to biotic and abiotic stresses. This hints at the fact that this protein family may act as a general communication molecule in plants. This research provides a valuable foundation for further investigations into plant-microbe interactions.

The second project centered around the small non-coding RNA known as 6S RNA, which is associated with stress-coping mechanisms in bacteria. Among the bacteria, the diverse group of lactic acid bacteria (LAB) plays a significant role in the food industry, serving as starter cultures for industrial fermentation processes or as probiotics among others. However, some LAB can also act as pathogens, posing a potential threat. The primary objective was to identify this non-coding RNA and characterize its features in LAB. The outcomes of this project were presented in the publication "*Insights into 6S RNA in lactic acid bacteria (LAB)*". The research involved various methodologies, including secondary structure-guided alignments, synteny classifications, phylogenetic reconstruction, and a guide for identifying 6S RNAs.

The findings from this work offer valuable insights for optimizing fermentation processes, developing growth stage markers, or designing putative antibiotic supplements.

The third project revolved around the orthology prediction tool, `Proteinortho`, which holds significant importance in comparative genomics, particularly in relation to the two previous projects. Orthologs are homologous genes that evolved from a speciation event and are believed to have retained similar functions across different species. The inference of orthologs is a critical step in multiple applications, such as genome annotation, phylogenetic analysis, and supertree analysis. Due to the rapid increase in genomic data already mentioned above, it is necessary to constantly optimize the tools for data processing. In this project, we performed an algorithmic update of `Proteinortho`, with a specific emphasis on enhancing its primary stages: sequence comparison and clustering. The results of this project can be found in the article "*Proteinortho6: Accelerating graph-based detection of (co-)orthologs in large-scale analyses*". Our improvements significantly enhanced the overall performance and scalability of the tool for current datasets and available computational resources. Additionally, the update increased the tool's availability, interoperability, and usability, making it more accessible for researchers in the field of comparative genomics.

In summary, the presented projects help to paint a clearer picture of two important biological entities with direct industrial applications and highlight improvements to an established tool that is essential to the field of comparative genomics.

Zusammenfassung

Die vergleichende Genomik ist ein interdisziplinäres Forschungsgebiet, in dem die genetische Zusammensetzung mehrerer Arten verglichen wird. Sie zielt darauf ab, die Ähnlichkeiten und Unterschiede in den Genomen verschiedener Organismen zu verstehen, um Erkenntnisse über ihre evolutionären Beziehungen, funktionellen Merkmale und Anpassungen zu gewinnen. Zu den wichtigsten Anwendungen der vergleichenden Genomik gehören die phylogenetische Rekonstruktion, bei der Wissenschaftler Evolutionsbäume konstruieren, um die Evolutionsgeschichte von Arten oder Genen zu veranschaulichen, und Orthologiebestimmungen, bei denen homologe Gene mit gemeinsamer Abstammung und ähnlichen Funktionen in verschiedenen Organismen identifiziert werden. In dieser Arbeit werden zwei biologisch motivierte Projekte hervorgehoben, bei denen bioinformatische Werkzeuge aus der vergleichenden Genomik zum Einsatz kommen. Außerdem führen Fortschritte in der Sequenzierungstechnologie dazu, dass enorme Mengen an genomischen Daten erzeugt werden. Einerseits bietet diese Datenflut Wissenschaftlern in der vergleichenden Genomik beispiellose Möglichkeiten, um die genomische Vielfalt in großem Maßstab zu erforschen. Auf der anderen Seite stellt die immense Menge an Daten jedoch auch eine große Herausforderung für die Datenverarbeitung und -speicherung dar.

Das erste Projekt dieser Arbeit konzentrierte sich auf die Untersuchung der Kiwellin-Proteinfamilie in Pflanzen, die eine entscheidende Rolle bei der Interaktion zwischen Pflanzen und Krankheitserregern spielt. Ziel des Projektes war es, die strukturellen Merkmale dieser Proteinfamilie zu verstehen und sie von den eng verwandten Barwin-like Proteinen zu unterscheiden. Die Ergebnisse wurden in der Publikation "*Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family*" veröffentlicht, in der eine systematische Nomenklatur entwickelt wurde, die drei verschiedene Unterklassen innerhalb der Kiwellin-Familie aufzeigen konnte. Darüber hinaus ergab eine Meta-Analyse öffentlich zugänglicher Transkriptomdaten spezifische Reaktionen von Kiwellinen in verschiedenen Pflanzengeweben und -sorten sowie ihre Reaktionen auf biotische und abiotische Stressfaktoren. Dies deutet darauf hin, dass diese Proteinfamilie möglicherweise als ein allgemeines Kommunikationsmolekül in Pflanzen fungiert. Diese Forschung bietet eine wertvolle Grundlage für weitere Untersuchungen der Interaktionen zwischen Pflanzen und Mikroben.

Das zweite Projekt befasste sich mit der kleinen nicht-kodierenden RNS, der so genannten 6S RNS, die mit Stressbewältigungsmechanismen in Bakterien in Verbindung gebracht wird. Innerhalb der Bakterien spielt die vielfältige Gruppe der Milchsäurebakterien (LAB) eine wichtige Rolle in der Lebensmittelindustrie, in der sie u.a. als Starterkulturen für Fermentationsprozesse oder als Probiotika fungieren. Einige LAB können jedoch auch als Krankheitserreger wirken und stellen somit eine potenzielle Bedrohung dar. Das Hauptziel dieses Projekts bestand darin, die 6S RNS zu identifizieren und ihre Eigenschaften in LAB zu charakterisieren. Die Ergebnisse wurden in der Veröffentlichung "*Insights into 6S RNA*

in lactic acid bacteria (LAB)" publiziert. Die Forschung umfasste verschiedene Methoden, darunter sekundärstrukturgeleitete Alignments, Synteny-Klassifizierungen, phylogenetische Rekonstruktion und einen Leitfaden zur Identifizierung der 6S RNS. Die Erkenntnisse aus dieser Arbeit bieten wertvolle Einsichten für die Optimierung von Fermentationsprozessen, die Entwicklung von Markern für das Wachstumsstadium oder die Entwicklung möglicher Antibiotikazusätze.

Das dritte Projekt drehte sich um das Tool zur Vorhersage von Orthologien, `Proteinortho`, das in der vergleichenden Genomik von großer Bedeutung ist, insbesondere in Bezug auf die beiden o.a. Projekte. Orthologe sind homologe Gene, die sich aus einem Speziationsereignis entwickelt haben und von denen man annimmt, dass sie ähnliche Funktionen artenübergreifend beibehalten haben. Die Bestimmung von Orthologen ist ein entscheidender Schritt bei zahlreichen Anwendungen, wie z.B. der Genomannotation, der phylogenetischen Analyse und der Supertree-Analyse. Aufgrund der bereits oben angesprochenen raschen Zunahme genomischer Daten ist es erforderlich, die Tools der Datenverabeitung stetig zu optimieren. In diesem Projekt haben wir eine algorithmische Aktualisierung von `Proteinortho` durchgeführt, mit besonderem Schwerpunkt auf der Verbesserung seiner primären Phasen: dem Sequenzvergleich und dem Clustering. Die Ergebnisse dieses Projekts sind in dem Artikel "*Proteinortho6: Accelerating graph-based detection of (co-)orthologs in large-scale analyses*" zu finden. Unsere Verbesserungen haben die Gesamtleistung und Skalierbarkeit des Tools für aktuelle Datensätze und verfügbare Rechenressourcen erheblich verbessert. Darüber hinaus wurden durch diese Aktualisierung die Verfügbarkeit, Interoperabilität und Benutzerfreundlichkeit des Tools verbessert, wodurch es für Wissenschaftler im Bereich der vergleichenden Genomik leichter handhabbar wird.

Zusammenfassend lässt sich sagen, dass die vorgestellten Projekte dazu beitragen, ein klareres Bild der 6S RNS und der Kiwellinfamilie im Hinblick auf potentielle industrielle Anwendungen zu zeichnen und Verbesserungen an einem etablierten Tool aus dem Forschungsbereich der vergleichenden Genomik hervorzuheben.

Contribution to Publications

⭑: included in this thesis.

Published

⭑ **Klemm, P.**, Christ, M., Altegoer, F., Freitag, J., Bange, G., & Lechner, M. Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family. *Frontiers in plant science* 13 (2022): 4832. doi: 10.3389/fpls.2022.1034708

> I was responsible for conducting the bioinformatic analyses. Collaboratively with MC and FA, I conducted an evaluation and validation of structural aspects, with MC specifically focusing on the manual assessment of structures. Furthermore, I conceptualized and executed the phylogenetic analysis, which included the supermatrix and reconciliation analysis. My involvement extended to the acquisition and analysis of RNA-seq data sets for the meta-analysis. Together with MC, I undertook the biological evaluation of these datasets. Subsequently, MC and I collaborated closely in the drafting of the manuscript. Furthermore, I created all figures and tables to elucidate our findings.

⭑ Cataldo, P. G., **Klemm, P.**, Thüring, M., Saavedra, L., Hebert, E. M., Hartmann, R. K., & Lechner, M. (2021). Insights into 6S RNA in lactic acid bacteria (LAB). *BMC Genomic Data*, 22, 1-15. doi: 10.1186/s12863-021-00983-2

> In this research endeavor, both ML and I served as the primary analysts for bioinformatic analysis. My specific areas of focus included conducting the phylogenetic analysis, synteny investigation, and the identification cre sites. Furthermore, I produced the majority of the figures and compiling the appendix.

- Bach, S., Demper, J. C., **Klemm, P.**, Schlereth, J., Lechner, M., Schoen, A., Kämpfer, L., Weber, F., Becker, S., Biedenkopf, N., & Hartmann, R. K. (2021). Identification and characterization of short leader and trailer RNAs synthesized by the Ebola virus RNA polymerase. PLoS pathogens, 17(10), e1010002. doi: 10.1371/journal.ppat.1010002 *(not related to thesis)*

> My roles encompassed data curation, formal analysis, and methodology development. Specifically, I curated, prepared and analysed the sequencing data (RNA-seq). Additionally, I produced figures for visualizing the data to enhance its interpretability.

- Obermann, W., Friedrich, A., Madhugiri, R., **Klemm, P.**, Mengel, J. P., Hain, T., Pleschka, S., Wendel, H.-G., Hartmann, R. K., Schiffmann, S., Ziebuhr, J., Müller, C., & Grünweller, A. (2022). Rocaglates as Antivirals: Comparing the Effects on Viral Resistance, Anti-Coronaviral Activity, RNA-Clamping on eIF4A and Immune Cell Toxicity. Viruses, 14(3), 519. doi: 10.3390/v14030519 *(not related to thesis)*

My key contribution involved conducting a comparison of sequencing data derived from the serial passaging experiment. Additionally, I generated the corresponding figure that visually encapsulated the results.

Submitted

- ⭐ **Klemm, P.**, Stadler, P. F., & Lechner, M. (expected 2023). Proteinortho6: Pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. Frontiers in Bioinformatics. In revision (Frontiers in Bioinformatics).

  I conceptualized and implement all improvements to `Proteinortho`. Additionally, I designed and executed the experimental setup, including the benchmarking aspects such as the quantification of running time, memory consumption, scalability, and the Quest for Orthologs benchmark assessment. Furthermore, I drafted the initial version of the manuscript and I was responsible for the creation of all figures and tables.

- Meier, D., Rauch, C., Wagner, M., **Klemm, P.**, Blumenkamp, P., Müller, R., Ellenberger, E., Karia, K. M., Vecchione, S., Serrania, J., Lechner, M., Fritz, G., Goesmann, A., & Becker, A. (expected 2023). A MoClo-compatible toolbox of ECF sigma factor-based regulatory switches for proteobacterial chassis. In revision (BioDesign Research). *(not related to thesis)*

  I conducted an assessment to determine the feasibility of predicting ECF sigma factor induced crosstalk *in silico*. For this, I processed and integrated the different sequencing datasets (RNA-seq, Cappable-seq). This included the generation of the corresponding figures. Additionally, I facilitated data distribution and sharing by providing and managing a GitLab repository, contributing to the reproducibility and transparency of our research efforts.

In Preparation

- Damm, K., Lechner, M., Helmecke, D., **Klemm, P.**, & Hartmann, R. K. 3'-tailing methods for ultrashort RNAs in RNA-seq applications at the example of 6S RNA-derived pRNAs. In preparation. *(not related to thesis)*

  I am collaborating closely with ML for the bioinformatic analyses in this project. Our combined efforts encompass designing, developing, and validating algorithmic solutions.

- Gößringer, M., **Klemm, P.**, Wäber, N. B., Schencking, I., Lechner, M., & Hartmann, R. K. Prokaryotic protein-only RNase P substitution in *Escherichia coli*: viability, processing defects and differences between isoenzymes. In preparation. *(not related to thesis)*

  My role in this project is the processing and analyzing the sequencing data (RNA-seq). Additionally, I integrate the results into figures and tables and provide a comprehensive supplementary documentation.

- Michel, C., **Klemm, P.**, Pauck, K., Türe, E., Wang, Y., Nist, A., Stiewe, T., Fritz, L., Raifer, H., Zhang, Y., Hühn, S., Giel, G., Grünewald, A., Weiland, P., Otto, C., Brendel, C., Metzelder, S., Neubauer, A., Ernst, T., Saussele, S., Miethe, S., Bange, G., Huber, M., Garn, H., Hochhaus, A., Lechner, M., & Burchert, A. Unraveling Mechanisms of Successful Treatment-Free Remission in Chronic Myeloid Leukemia: Insights from Translational Substudies in the German CML Study IX (Endure). In preparation. *(not related to thesis)*

  I am currently the primary data analyst for this project, which encompasses tasks such as conceptualization, *in silico* analyses, interpretation, and discussions. My responsibilities involve the processing and analysis of single-cell sequencing datasets, which are sourced from the BD Rhapsody^{TM} pipeline.

Talks and Poster Presentations

Talk 2019 in Bled, Slovenia: TBI Winterseminar.

Need for Speed: Proteinortho Nitro

Poster 2019 in Marburg, Germany: Spotlight on Methods in Microbiology.

Proteinortho - A tool to detect orthologous genes within different species

Talk 2020 and 2021 in Marburg, Germany: Symposium on Interdisciplinary Bioinformatics and Biomedical Data Science.

’20 *Large-scale Orthology Prediction*
’21 *Kiwellin proteins as mediators of plant-pathogen and -symbiont interaction*

Talk 2022 in Hirschegg, Austria: Pharmaceutical Chemistry Winterseminar.

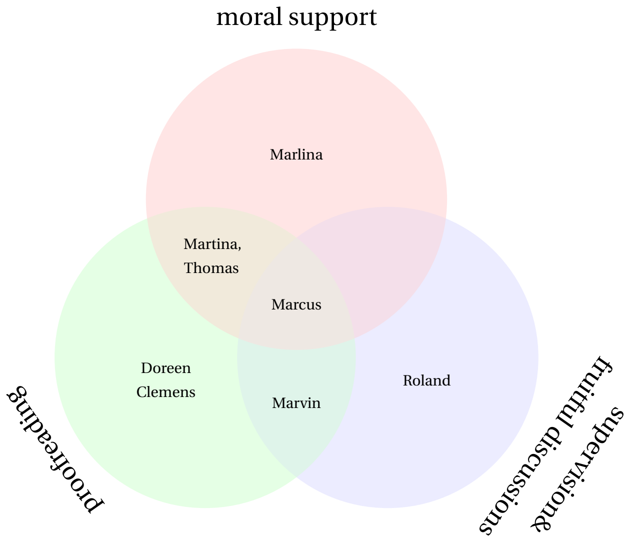
Insights into 6S RNA in lactic acid bacteria (LAB)

Poster 2023 in Sesimbra, Portugal: Cell2Cell Workshop.

Unraveling Mechanisms of Successful Treatment-Free Remission in Chronic Myeloid Leukemia

Acknowledgements

I want to extend my heartfelt appreciation to Prof. Roland Hartmann and Dr. Marcus Lechner, my supervisors, for granting me the invaluable chance to carry out the research for this thesis and for being a constant source of guidance and encouragement throughout the journey. Furthermore, I would like to thank my committee, Prof. Moritz Bünemann, Prof. Arnold Grünweller, and Prof. Stefan Jansen, for reading and evaluating my thesis. Moreover, I like to thank Clemens, Marvin, and Doreen for proofreading and valuable discussion. Next, I would like to thank my working group and friends for scientific and non-scientific discussions: Rebecca Feyh, Amri Schlüter, Jana Demper, Jana Wiegard, Katja Hütte, Marietta Thüring, Wiebke Obermann, Clara Müller, Mathäus Drabek, Nadine Weber, Simone Bach, Sweetha Ganapathy, Clemens Elias Thölken, Michael Vockenhuber, Bayan Alanati, and Fareha Masood. Throughout my early scientific career, I have been incredibly fortunate to receive unwavering support and guidance from remarkable individuals: Schikora, Matze, Prof. Karlhans and Nicole Endlich. Last but not least, I would like to thank my parents, Martina and Thomas, and my fiancee, Marlina, for providing me with continuous encouragement throughout the process of writing this thesis. This work would not have been possible without you all.



Contents

Abbreviations

6S RNA	
sRNA	small RNA
RNAP	RNA polymerase
pRNA	product RNA
LNA	Locked Nucleic Acid
LAB	lactic acid bacteria
B. subtilis	*Bacillus subtilis*
E. coli	*Escherichia coli*
Kiwellins	
kDa	kilo Dalton
DPBB	double-psi-β-barrel
BL	Barwin-like
Kwl	Kiwellin
Cmu	chorismate mutase
U. maydis	*Ustilago maydis*
Proteinortho	
BLAST	Basic Local Alignment Search Tool
RBAH	reciprocal best alignment heuristic
RBH	reciprocal best hit
k	thousand
α**,** `conn`	algebraic connectivity clustering parameter
`e`	E-value cutoff
f**,** `sim`	similarity threshold
QfO	Quest for Orthologs
CC	connected component
ARI	adjusted Rand index
l2FC	log2 fold change
h	hours
GB	Gigabyte
`RMSD`	root-mean-square deviation
`MA`	matching atoms
Phylogenetics	
UPGMA	unweighted pair group method with arithmetic mean
MP	maximum parsimony
ML	maximum likelihood
HGT	Horizontal gene transfer
LBA	long branch attraction
DTL	model with duplication, transfer and loss events

I

Introduction

Comparative genomics is a relatively young field that involves comparing genomes or proteomes of different species to identify similarities in their genetic makeup. In 1977, Sanger and Coulson pioneered a sequencing technique called the chain-termination method, which revolutionized the field of genetics by enabling the deciphering of genetic material (Sanger, 1977). Since then, sequencing techniques have advanced significantly, not only in terms of efficiency but also in cost reduction (Sboner, 2011; Stephens, 2015).

These advancements have led to an exponential growth of sequencing data, as demonstrated in Fig. 1 with the `NCBI GenBank` database (Benson, 2012). This flood of genomic data presents new opportunities for extensive analysis on one side. Still, it poses various challenges in processing this massive amount of information on the other side. The analysis that mainly benefits from this includes phylogenetic reconstruction, synteny evaluation, multiple sequence alignment, and orthology predictions. This work aims to address these challenges in the context of a non-coding RNA (6S) and a protein family (Kiwellin) and show computational improvements to orthology prediction. A comprehensive understanding of the field's intricacies and illumination of potential solutions are aimed to be provided from a biological and especially a computational point of view.

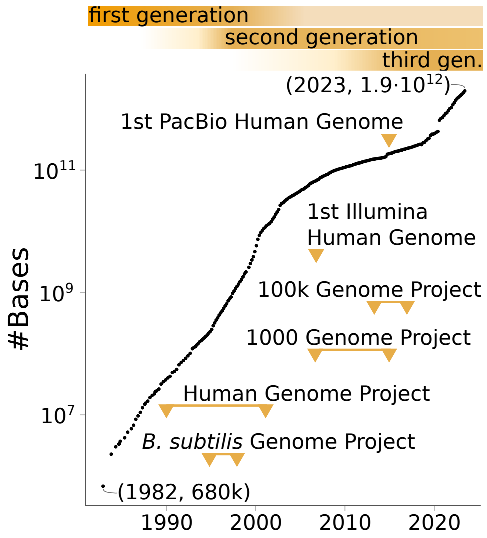


Figure 1: Number of bases in `NCBI GenBank` over the years (Benson, 2012) with selected milestones[1] (Stephens, 2015).

This chapter provides a comprehensive overview of the central biological systems, with a specific focus on the 6S RNA and the Kiwellin protein family. Additionally, key bioinformatic concepts such as homology, orthology, including the tool `Proteinortho`, and phylogenetics will be introduced to the reader in detail. Lastly, a short overview of the mathematical

[1] 1995-1997: *B. subtilis* Genome Project (Kunst, 1997), 1990-2001: Human Genome Project (first generation) (I. Consortium, 2001), 2008: Illumina human genome (second generation) (Bentley, 2008), 2008-2015: 1000 Genome Project (1. G. P. Consortium, 2015), 2015: PacBio human genome (third generation) (Chaisson, 2015), 2013-2018: 100.000 Genome Project (Turnbull, 2018),

foundation is given, explaining the concept of spectral clustering. Each section will summarize the theoretical underpinnings of the concepts, offering a concise explanation of their fundamental principles. Furthermore, selected implementations and state-of-the-art approaches will be highlighted to showcase the practical applications of these concepts. This chapter establishes a basis for the readers and provides the motivation for the research questions of the different projects presented in the next chapter.

Moving on to the second chapter, the three research articles are presented, each addressing the research questions introduced in the first chapter. The results and discussions of the articles lay the foundation for the following chapter, which includes overarching and topic-specific conclusions and outlooks. The formulated hypotheses in this chapter can potentially serve as starting points for future researchers interested in delving further into these intriguing topics.

The Appendix of this dissertation comprises additional results and discussions obtained from the three articles, delving deeper into the research investigations. For further details on the results, including programs and datasets utilized, as well as the figures produced during this study, readers can refer to the accompanying supplementary repository[2]. Furthermore, the supplementary files of the three articles are also available in the same repository, providing a comprehensive set of materials to support and validate the findings presented in this dissertation.

[2] https://gitlab.uni-marburg.de/synmikro/ag-lechner/paul-klemm-dissertation-supplement

1.1 Biological Systems

1.1.1 The 6S RNA

Small non-coding RNAs (sRNAs) are increasingly becoming the focus of scientific research[3]. One prominent member is the 6S RNA (also known as SsrS RNA), that was first discovered by Hindley in 1967 in *E. coli* (Hindley, 1967).

The function of 6S RNA was initially elusive, as knock-out or over-expression mutants of the 6S RNA did not produce significant phenotypes (Wassarman, 2000). Nowadays, it is known as an important regulatory unit of the transcriptional apparatus in the context of environmental stresses such as oxidative stress or nutrient deficiency (Wehner, 2014; Burenina, 2022) and being sporulation associated (Cavanagh, 2013).

Similar to proteins, the structure of non-coding RNAs is crucial to their function. 6S RNA is between 160 and 200 nucleotides long and canonically takes the shape of a rod with an enlarged inner loop called central bulge flanked by two helical arms on both sides as shown in Fig. 2 (Wehner, 2014). This unique structure mimics B-form DNA and resembles an open promoter complex (Chen, 2017). This enables the 6S RNA to bind with the bacterial DNA-dependant RNA polymerase (RNAP) holoenzyme, which is saturated with the housekeeping sigma factor[4] (Wassarman, 2000; Barrick, 2005).

In *E. coli* and *B. subtilis*, the 6S RNA accumulates in the exponential phase and peaks in the stationary phase where nutrients become limited (Beckmann, 2011; Wassarman, 2000). The RNAP uses the 6S RNA as a template for the transcription and produces short abortive product RNAs called pRNAs (Wassarman, 2000). Therefore, under stresses like this nutritional deprivation in the stationary phase, the 6S RNA competes with regular DNA promoters of the housekeeping sigma factor as it reduces the availability of this holoenzyme. This interaction suppresses housekeeping-associated gene expression and aids the formation of other holoenzymes with different alternative sigma factors (Wehner, 2014).

As the bacteria transition to a new exponential growth phase, for example, by increasing the nutrient concentration, the pRNAs increase in lengths (Beckmann, 2011). Once they reach a length of around 14 nucleotides, a 6S RNA:pRNA hybrid is formed, causing a conformational change that ultimately leads to the release of the RNAP holoenzyme and the degradation of the 6S RNA (Beckmann, 2012; Steuten, 2014). This release of the RNAP induces housekeeping sigma factor associated gene expression. Therefore, the 6S RNA directly encodes a self-regulatory element in the form of the pRNAs in different lengths.

While most of the early studies were focused on *E. coli* and *B. subtilis*, a comprehensive phylogenetic analysis in 2014 revealed the widespread presence of the 6S RNA across bacteria, including extremophiles like the hyperthermophilic *Aquifex aeolicus* that was found, for example, in hot springs in the Yellowstone National Park (Willkomm, 2005). A single 6S RNA

[3] PubMed results for "small RNA": 1991: 24, 2010: 447, 2021: 974 p.a.

[4] e.g. σ_{70} for *E. coli* and σ_A for *B. subtilis*

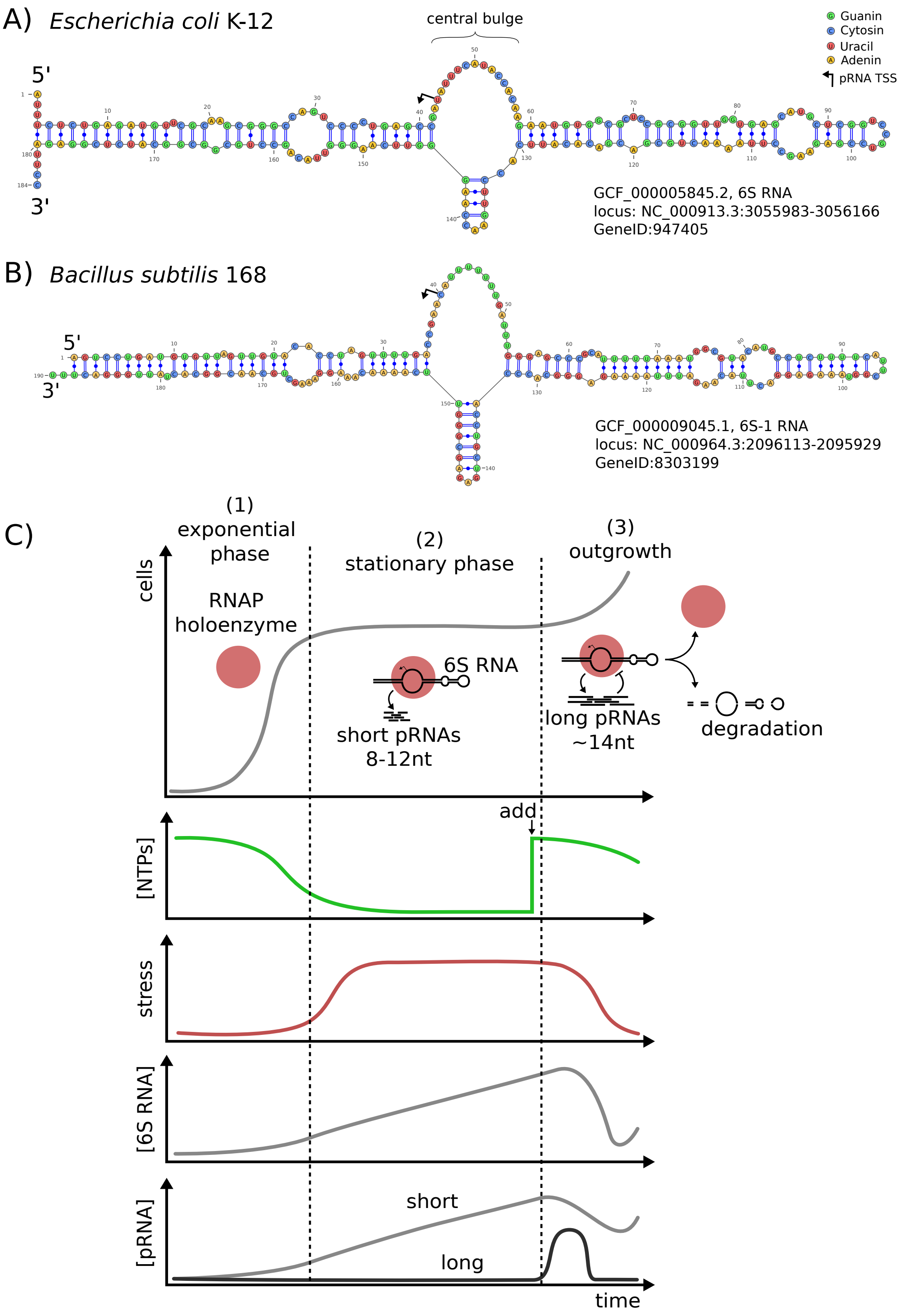


Figure 2: `RNAfold` structure predictions of the 6S RNA of *Escherichia coli* strain K-12 **(A)** and the 6S-1 RNA of *Bacillus subtilis* strain 168 **(B)**. More details on the two structure predictions can be found in the supplementary repository. Schematic model of the function of the 6S RNA **(C)**. Three typical growth stages (separated by dashed lines) modeled after experimental results (Beckmann, 2011; Beckmann, 2012): The exponential phase (1) with high but depleting nutrition concentration and exponentially rising numbers of cells. The stationary phase (2) starts when nutrients are sparse, the number of cells plateaus, and the stress level peaks. A second exponential or outgrowth phase (3) is induced by adding further nutrients (black arrow, add). The 6S RNA concentration peaks in the late stationary phase. Short pRNAs accumulate from the early exponential phase to the late stationary phase. Long pRNAs burst in expression shortly after the induced outgrowth, suppressing the 6S RNA:RNAP complex, which leads to a degradation of the 6S RNA.

molecule has been identified in most bacterial taxa, with some exceptions (Wehner, 2014). For instance, for *Bacillus subtilis*, two paralogs exists: the 6S-1 RNA[5], corresponds to the canonical 6S RNA, and the 6S-2 RNA[6], that is specifically involved in the regulation of biofilm formation in wild-type strains (Thüring, 2021). It is to be noted that the structural rearrangement-driven regulatory mechanism of the pRNAs is conserved in bacteria as well (Beckmann, 2012).

The 6S RNA is found in many different types of bacteria, which is also true for the lactic acid bacteria (LAB), which is of key interest to the food sector. LAB are a diverse group of Gram-positive, acid-tolerating bacteria of the order of Lactobacillales, which can be subdivided into the following six families: *Aerococcaceae*, *Carnobacteriaceae*, *Enterococcaceae*, *Lactobacillaceae*, *Leuconostocaceae*, and *Streptococcaceae*. One common metabolic characteristic among LAB is their ability to produce lactic acid in carbohydrate fermentation (Leroy, 2004). Because of the ability to ferment milk and other food products, LAB are of key interest in the food industry, such as starter cultures. Examples of products include yogurt, cheese, natto, and kimchi. Furthermore, some LAB strains are used as probiotics probiotics (Leroy, 2004). Many of the LAB are Generally Recognized as Safe (GRAS) with exceptions of opportunistic pathogens found mainly in the genera of *Streptococcus* and *Enterococcus* (Mattila-Sandholm, 1999).

LAB are exposed to diverse stresses, both within the gastrointestinal tract or in industrial environments (Smid, 2010), where the regulatory function of the 6S RNA comes into play. Therefore, identifying and classifying 6S RNA in LAB are crucial for future research and have direct applications. Preliminary studies showed that 6S RNA knockout strains metabolize nutrients faster compared to the corresponding wild type strain (Cavanagh, 2012). By exploring the regulatory mechanisms of 6S RNA in LAB, researchers can potentially optimize LAB strains for better performance in food fermentation, storage, and other relevant applications. Furthermore, in this context, it is possible to investigate the potential for increased production of secondary metabolites, such as surfactants, in 6S RNA knockout strains.

The objective of this work is to explore the 6S RNA in LAB. This includes identifying 6S RNA, conducting phylogenetic analysis to understand its evolutionary relationships, and investigating its genomic context.

[5] *bsrA*, GeneID:8303199

[6] *bsrB*, GeneID:8303197

1.1.2 The Kiwellin Protein Family 🥝

The world population is expected to surpass 10 billion until the year 2050 (Roser, 2013; Hickey, 2019). This constant growth induces an increasing relevance of crop diseases for the global food security (Van Dijk, 2021; Fenu, 2021). In combination with globalization, climate change adds to this problem, enabling pathogens to spread to new regions (Bebber, 2013). Furthermore, the increasing global temperatures induce drought stress, weakening resistance systems of plants (Bebber, 2013). Together this gives pathogens that evolved over millions of years in an arms race against plants more and more an advantage (Maor, 2005). Recent results have highlighted Kiwellins in maize plants as a weapon in the fight against a pathogenic fungus (Han, 2019; Altegoer, 2020).

Corn smut is a disease in maize (*Zea mays* 🌽) caused by the infection with the biotrophic fungi *Ustilago maydis*. The fungus secretes virulence factors called effectors to counteract the plant defense systems like proteases (Misas Villamil, 2019). Another example of these effectors is the chorismate mutase Cmu1 of *U. maydis* that targets the salicylic acid pathway of the plant (Djamei, 2011). This is achieved by converting the plant chorismate to prephenate and thus reducing its availability as a component for the salicylic acid pathway, resulting in a suppressed plant defence response (Djamei, 2011). The effector mimics the plant's own chorismate mutase but lacks the allosteric regulation present in the plant enzyme. Two Kiwellins of *Z. mays* were shown to specifically bind to fungi Cmu1 at the active site and thus counteracting this mechanism (Han, 2019; Altegoer, 2020). Fig. 3B depicts this interaction schematically. This discovery represents the first instance where Kiwellins have been found to have an additional function besides human allergens identified in kiwifruit (*Actinidia spp.* 🥝) (Tamburrini, 2005; Fine, 1981; Wang, 2019)

Kiwellins are around 20 kDa or 190 amino acids of size and are cysteine-rich, contributing to plenty of disulfide bridges and exceptional stability. Moreover, a signal peptide in most Kiwellins suggests that they can be secreted from the cell via the conventional pathway. More insights were gained from the published crystal structures (Han, 2019; Hamiaux, 2014), that revealed different domains as shown in Fig. 3A. The primary characteristic features are a short β-hairpin motif at the N-terminus in combination with a double-psi-β-barrel fold (DPBB) that is composed of three α-helices and six parallel and antiparallel connected β-strands. The DPBB of Kiwellins is shares high similarities to the class of Barwin-like proteins (BL).

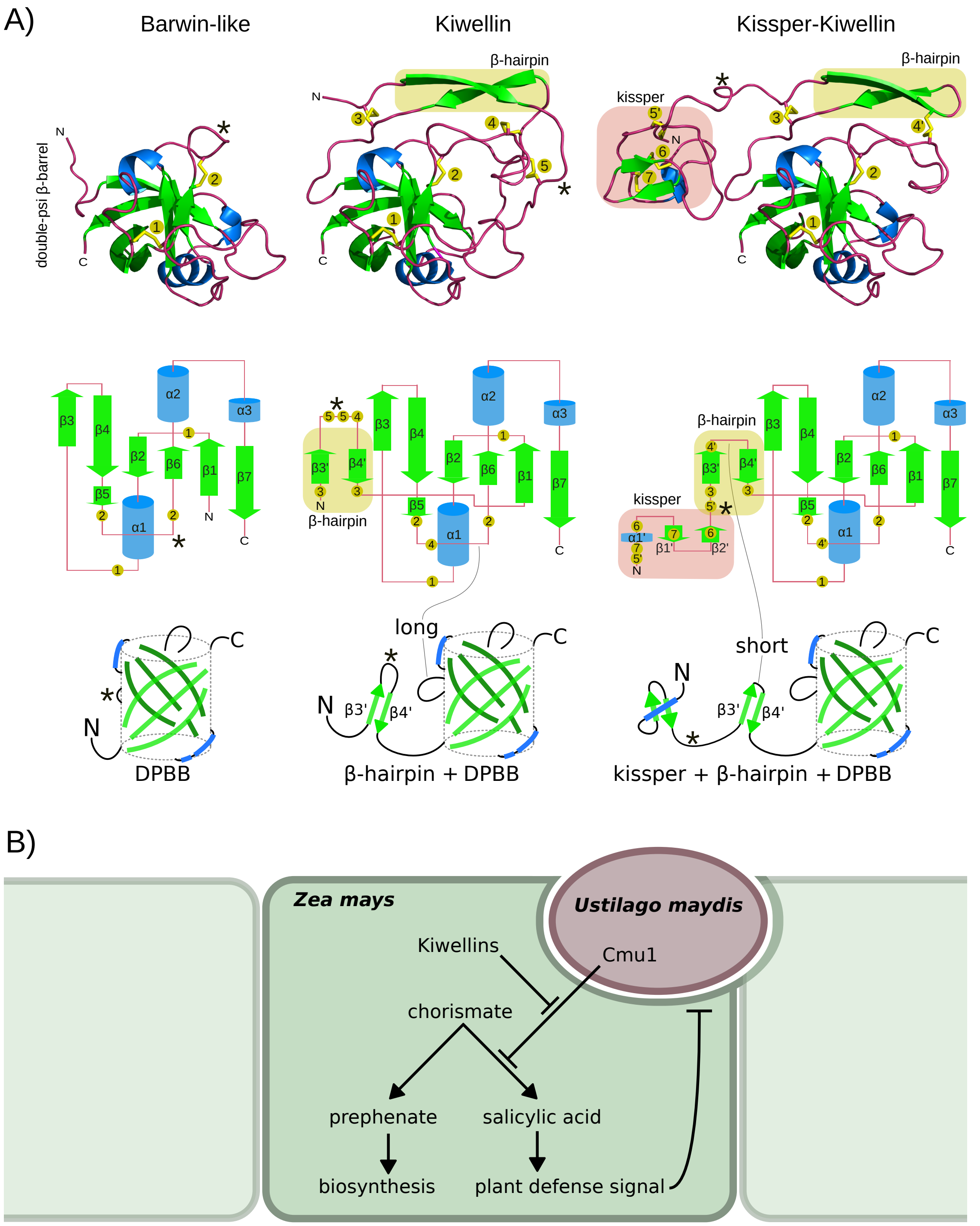


Figure 3: From left to right **(A)**: Barwin-like (BL), Kiwellin, and Kissper-Kiwellin proteins structure prediction of consensus sequences with different levels of abstractions from top to bottom (3D model, planar projection, comic abstraction). Green arrow: β sheets, blue barrel: α helices, red/black lines: loop regions, yellow box: the β-hairpin, red box: the kissper domain, yellow circles: cysteine residues, $\ast$: variable length loop region. Two characteristic features are highlighted: long: an internal loop region that is predominantly short in BL and long in Kiwellins, short: a short loop in of the β-hairpin in Kissper-Kiwellins. Adapted from (Klemm, 2022). Schematic model of Cmu1 mediated pathogenesis in *Zea mays* infected by *Ustilago maydis* **(B)**. Cmu1 suppresses the salicylic acid mediated plant defense system, and Kiwellins counteract this mechanism by binding to the Cmu1. Green: Maize cell, red: *Ustilago maydis* cell. Adapted from (Djamei, 2011; Han, 2017).

This similarity led to frequent confusion between the Kiwellin and BL protein families (Jaswal, 2021; Han, 2019; Blum, 2021). The BL protein family is widespread, found in bacteria, plants, and fungi, with various functions like sugar binding, cleavage activity, and pathogenesis-related activity (Dabravolski, 2021; Scherer, 2010). Another closely related subfamily is the Kissper-Kiwellins which contain another N-terminus extension called the kissper domain, which has been reported to exhibit pore-forming activity (Tuppo, 2008).

Driven by the findings in maize, where Kiwellins counteract a specific pathogenic effector, this work aims to classify and investigate the dissemination of the whole Kiwellin family. Based on this, a phylogeny can be constructed and investigated, resulting in a nomenclature. Furthermore, the research seeks to examine whether there are distinct subclasses and differentiation characteristics and how they are distributed across different taxonomic groups. A meta-analysis aims to shed light of how Kiwellin proteins are expressed in different situations. The study aims to foster new understandings of Kiwellins with putative agricultural implications by addressing these aspects.

1.2 Bioinformatic Concepts

To simplify terminology, the term “protein” refers in the following to the amino acid representation and “gene” analogously to the nucleotide representation; in addition, genes and proteins can be freely interchanged in this context.

1.2.1 Homology, Orthology, and co.

Homology is a concept that generally captures the “sameness” of structures, genes, or proteins, and its definition has evolved over time (Petsko, 2001; Panchen, 2007; Wake, 2007; Koonin, 2001; Jensen, 2001; Boyden, 1969; Pearson, 2013; Wake, 1994; Fitch, 2000). Charles Darwin gave homology an evolutionary perspective and associated it with similarity due to common ancestry (Darwin, 1859). Fitch provided a formal definition, where homology refers to characters that descend with divergence from a common ancestral state (Fitch, 1970; Fitch, 2000). Characters refer to any feature of sequence (nucleotide, amino acids), structure, morphology, or behavior. Genes are considered homologs when a significant fraction of residues show homology in the ancestral species beyond what would be expected by chance alone (Fitch, 1970).

It is to be noted that the term homology is heavily overloaded and comes in various flavors in different fields of study, e.g. structural/morphological, functional, behavior, cladistic, and many more. Furthermore, it has been recycled and used in pre-Darwinian times, with a focus on similarity rather than evolutionary relationships. In this work, the post-Darwinian point of view of homology is used in congruence with the definition of Fitch.

Fitch further refined homology into two sub-classes (Fitch, 1970), which are orthologs and paralogs. Orthologs (ortho = exact) are homologs that originate in a speciation event, where two species diverge. They correspond to the genes whose evolutionary relationship directly corresponds with the phylogeny of the species where the genes reside. On the other hand, paralogs (para = beside) originate from duplication events.

For example, the two *6S RNA* genes of *B. subtilis*[7] represent paralogs while the *6S RNA* genes from *B. subtilis* and *E. coli*[8] represent orthologs. Due to the interplay of duplication and speciation events, the relationship of orthologs can be considered hierarchical and varies in complexity: one-to-one, one-to-many, or even many-to-many. The aforementioned example would represent a one-to-many relationship. Groups of orthologs with lineage-specific duplicates are called co-orthologs.

There are some common *a priori* assumptions and conjectures that are essential for the interpretation and inference of homologs and orthologs:

Orthology function conjecture. Homology is often used in a functional context, although the Fitch definition does not explicitly mention functional relationships. Furthermore, orthologs

[7] *bsrA*/GeneID:8303199, *bsrB*/GeneID:8303197

[8] ssrS/GeneID:947405

are conceptually expected to conserve functions, while paralogs are more functionally diverse (Gabaldón, 2013; Pearson, 2013). The idea is based on the assumption that duplicated genes are subject to low selective pressure as the original gene already fulfills the necessary function. Mutations can accumulate without a decrease in fitness which can lead to a loss of the gene or otherwise result in a new function (neofunctionalization) or specialized function (subfunctionalization) Ohno (1970). Orthologs, on the other hand, originate in the event of high evolutionary pressure (speciation event) and should thus be more conserved. It is worth noting that this conjecture already fails for many co-orthologs where only one lineage-specific duplicated gene retains the function of the defining speciation event. Furthermore, similar structures can have different functions in different organisms (Fitch, 2000). Although this conjecture is controversially discussed (Gerlt, 2000; Nehrt, 2011; Petsko, 2001) in general this conjecture appears to hold (Gabaldón, 2013).

Orthology is transitive. Transitivity implies that if gene A is orthologous to gene B, and gene B to gene C, then gene A will be considered orthologous to gene C. This principle serves as the foundation for forming orthology groups like $\{A, B, C\}$. It is important to note that, conceptually, this assumption has been critiqued as inaccurate (Johnson, 2007). From a practical standpoint, it is often more feasible to aggregate orthologs into groups for analysis purposes. This approach allows for a more straightforward analysis compared to a list of individual orthology relationships.

Orthologs are among the most similar sequences. While this hypothesis holds as a statistical trend, there are cases where it fails due to factors like horizontal gene transfer (Koski, 2001; Gabaldón, 2013; Altenhoff, 2009; Hulsen, 2006; Wolf, 2012). This conjecture is at the heart of the reciprocal best alignment heuristic described in the next section.

In summary, the ambiguity surrounding the concept of homology arises from its historical development, the different definitions used in various fields, and the changing perspectives on its meaning. Despite these challenges and assumptions, homology remains a fundamental concept in biology, particularly in comparative genomics.

1.2.2 (Co-)Orthology Detection

Identifying orthologs within a set of sequences from different species is known as orthology inference, which can be classified into two methodologies: the tree-based and graph-based approaches.

Graph-based methods rely mainly on similarity measurements as a proxy for the underlying evolutionary relationship. Typically these methods begin with the *reciprocal best alignment heuristic* (RBAH) (Bork, 1998). Using a homology search algorithm like `BLAST`, lists of homologs can be inferred between two species, each acting as the database or query. If a gene in one species is the best match for another gene in the other species and vice versa, it is considered a reciprocal best hit (RBH) or bidirectional best hit. This procedure is then repeated for any pair of species, ultimately resulting in a similarity graph, where nodes represent genes and edges correspond to putative orthologs, the so-called reciprocal best hit graph or RBH graph. With the transitive conjecture, the connected components form groups of putative co-orthologs.

It is known that these graphs suffer from the small world phenomenon (Milgram, 1967): with an increasing number of species or genes, the connected components grow disproportional in size. Resulting in massive groups covering more and more of the graph and, in turn, becoming less informative. This is demonstrated with real-world data in section 1.3.1. Subsequently, clustering algorithms are usually employed to prune the graph to a digestible size and reduce false positives. Examples are `OMA` (Altenhoff, 2021), `SonicParanoid` (Cosentino, 2019), `OrthoMCL` (Li, 2003), and `Proteinortho` (Lechner, 2011).

On the other hand, tree-based methods aim to reconstruct the evolutionary gene tree using sequence alignments together with the underlying species tree. Orthology predictions are based on this reconciliation procedure (the alignment of the species and gene tree). Conceptionally are these methods closer to the evolutionary definition of Fitch. An example that combines both approaches is `OrthoFinder2` (Emms, 2018), which initially employs a graph-based method to estimate putative orthologs and subsequently refines this prediction with a phylogenetic analysis. The main drawbacks of the tree-based approach are its higher computational costs and the challenges in inferring and reconciling the species tree.

`Proteinortho`

A widely adopted orthology inference tool in the scientific community, with more than 1000 citations and 100k downloads on `Bioconda`[9], is `Proteinortho` (Lechner, 2011). `Proteinortho` has been used to address various research questions, including the identification of conserved proteins for constructing species trees (Klemm, 2022; Peter, 2018) or the discovery of antibiotic resistance genes between resistant and susceptible strains in the human gut (Bisanz, 2018). Furthermore, it is utilized in programs such as `funannotate`, an automatic pipeline for genome annotation (Palmer, 2022).

`Proteinortho` follows the graph-based methodology and a schematic workflow is depicted in Fig. 4. In the first step, databases are generated for all input species. The next step involves an all-versus-all blast search, where each species is compared against all other species using a homology search program such as `diamond` (Buchfink, 2015) or `BLAST` (Altschul, 1990)[10]. The resulting outputs are then filtered based on a minimum E-value[11], minimum coverage[12] among others. Next, the pairwise results are combined to construct a graph using a modified version of the RBAH that is relaxed by a factor to allow for sub-optimal hits, the so-called adaptive RBAH (Lechner, 2011). The parameter controlling false positives in the adaptive RBAH is the similarity threshold f[13]. By setting the cutoff to 100%, only the canonical RBH is retained, while a value of 0 allows any reciprocal hit to be included in the output graph.

[9] https://anaconda.org/bioconda/proteinortho

[10] parameter: `-p`, default: `diamond`

[11] parameter: `-e`, default: 10^{-5}

[12] parameter: `-cov`, default: 50%

[13] parameter: `-sim`, default: 95%

In the final step, spectral clustering is applied to break down the graph into smaller groups. The number and size of these groups can be controlled by adjusting the algebraic connectivity cutoff α[14]. A higher threshold leads to further partitioning of the graph leading to more but smaller groups.

In this work, the advancements introduced in the updated version of `Proteinortho6` will be compared to its predecessor and other competing tools in the field. Additionally, the influence of various parameters that control the quality as well as the resource consumption will be investigated. Finally, alternative `BLAST` methods such as `diamond` will be explored, the spectral clustering approach will be re-evaluated, and the accuracy of `Proteinortho` in the QfO (Quest for Orthologs) benchmarking framework will be assessed (Altenhoff, 2016).

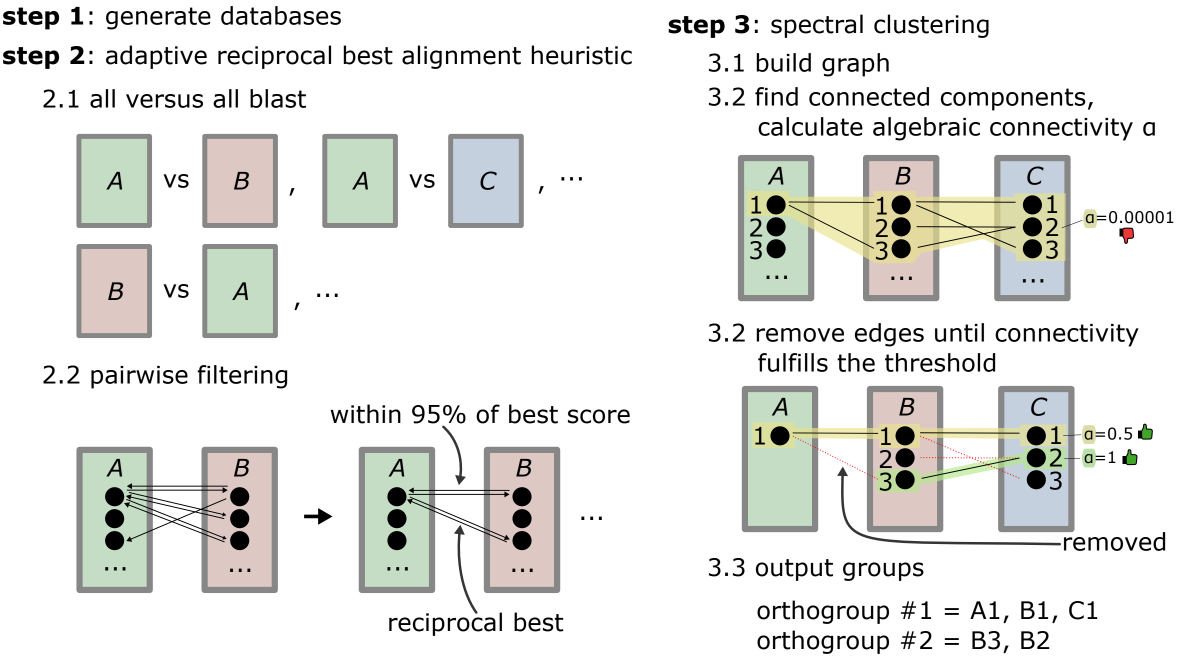

step 1: generate databases
step 2: adaptive reciprocal best alignment heuristic
2.1 all versus all blast
A vs B , A vs C , ...
B vs A , ...
2.2 pairwise filtering
within 95% of best score
A
B
A
B
...
reciprocal best
step 3: spectral clustering
3.1 build graph
3.2 find connected components, calculate algebraic connectivity α
A
B
C
1
2
3
α=0.00001
3.2 remove edges until connectivity fulfills the threshold
α=0.5
α=1
removed
3.3 output groups
orthogroup #1 = A1, B1, C1
orthogroup #2 = B3, B2

Figure 4: Schematic workflow of `Proteinortho`. Colored boxes: species, black circle: genes or proteins, black arrow: blast hit, red dotted line: removed edge. Step 1: generating databases. Step 2: adaptive RBAH using the specified homology search tool and filtering thresholds, Step 3: combine pairwise results into an undirected graph, identify connected components, calculate the algebraic connectivity, and remove edges until connectivity is sufficiently high.

1.2.3 Phylogenetic Tree Inference

In 1837, Charles Darwin sketched a simple evolutionary tree of life in his notebook B with the words "*I think*" scribbled on top, see Fig. 5 (Darwin, 1837). Since then, evolutionary trees have become a crucial tool in the field of evolutionary biology. Darwin's central idea was that genes evolve through different events and accumulate changes along their way, a concept he referred to as *descent with modification* or *natural selection* (Darwin, 1859). This idea laid the foundation for representing gene families in the form of undirected trees, called phylogenetic trees. However, it is important to note that there are cases where this logic fails, such as if horizontal gene transfer or hybridization occurs, which are inherently directed events. This

[14] parameter: `-conn`, default: 0.1

renders trees unable to fully represent the true evolutionary history relationship. Nevertheless, this concept remains groundbreaking for understanding the complex relationships in nature.

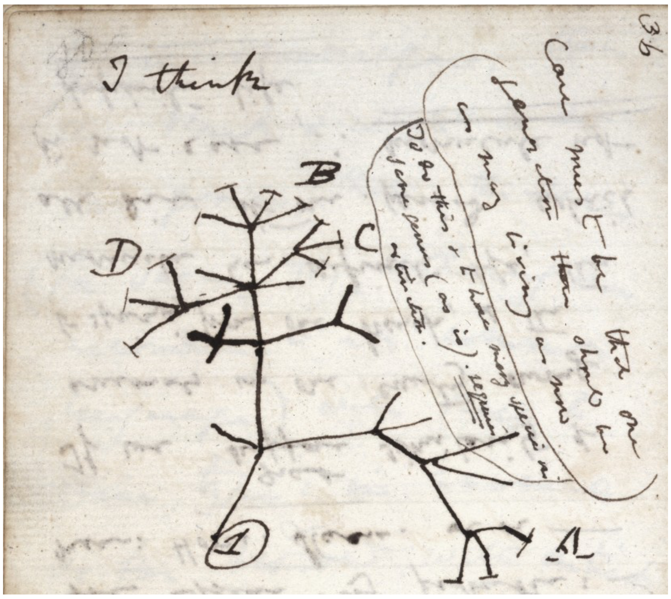

36
I think
B
C
D
A
1
Case must be that one generation then should be as many living as now. To do this & to have many species in same genus (as is). requires extinction.

Figure 5: Excerpt of Charles Darwin's notebook B, the circled one may indicate the origin of life. (Darwin, 1837)

Today, the field of phylogenetics focuses on reconstructing and assessing past lineages through evolutionary trees. Reconstructing the evolutionary history of a given set of sequences is a challenging problem only sequences from extant species can be used as input. There are two primary methodologies for phylogenetic inference: distance-based and character-based. Both of these methods are heuristics because the problem of finding the optimal tree is computationally infeasible.

Distance-based methods simplify the alignment by constructing a distance matrix using metrics such as the number of non-nonidentical matches or the Levenshtein edit distance (Levenshtein, 1966). In the second step, a hierarchical clustering of the rows and columns is used to construct a tree out of this distance matrix. Two general methods that implement such inference methods are UPGMA (unweighted pair group method with arithmetic mean) (Michener, 1957) and neighbor-joining (Saitou, 1987). A specialized example that clusters non-coding RNAs based on their secondary structure is `RNAclust` (Engelhardt, 2010).

Character-based methods, on the other hand, improve the quality of output at the expense of being more computationally expensive. Instead of using a simple distance metric, these methods analyze the full differences between the sequences in question. Different methodologies of character-based methods employ different mathematical frameworks, such as maximum parsimony and maximum likelihood.

The main concept of maximum parsimony (MP) is to find the solution that explains the evidence while minimizing the complexity of the model, favoring the simplest evolutionary solution. This concept is inspired by the philosophical principle of Occam's razor: *Numquam* *ponenda est pluralitas sine necessitate* (plurality must never be posited without necessity). Despite the idealized simplicity, the stringent assumptions do not always correspond to the best approximation of the true evolutionary history (Felsenstein, 1978). It was shown that parsimony approaches are entangled with the molecular clock assumption, which assumes a constant rate of mutations across all lineages as in the UPGMA algorithm (Felsenstein, 1978; Kapli, 2020). This assumption is often violated in real-world datasets, leading to incorrect tree topologies (Felsenstein, 1978).

To address these issues, the maximum likelihood (ML) method increases the model complexity at the expense of higher computational costs. At its core, ML approaches rate trees by their likelihood to generate the observed data and thereby seek to maximize this likelihood given some set of parameters, such as tree topology or branch lengths. When modifying these parameters, one can compare the resulting likelihood and search for a tree maximizing this function. Nowadays, ML methods are preferred over MP methods as the parameters can be arbitrarily complex to fit the needs of the data. Popular implementations of ML are `RAxML` (Stamatakis, 2014) and `IQ-Tree` (Nguyen, 2015). `IQ-Tree` differs from the approach of `RAxML` by shifting the objective from a global optimization problem to a local one. The local structures are so-called quartets that are trees from four sequences and thus represent only a tiny fraction of the data. By optimizing a set of quartets and combining the information, the final phylogenetic tree is constructed (Strimmer, 1996).

The authors of `IQ-Tree` demonstrated that their tool could produce higher likelihoods than `RAxML` (Nguyen, 2015). The newer variation `RAxML-NG` (Kozlov, 2019) significantly improved upon `RAxML` but shows mixed results in comparison with `IQ-Tree`. In summary, both `IQ-Tree` and `RAxML-NG` are well-suited tools for phylogenetic inference. However, in this study, `IQ-Tree` was chosen as the preferred option due to its superior usability and user-friendly features.

The following sections are aimed to give a basic understanding of common phylogenetic biases and two analyses that are used throughout this work.

Phylogenetic Errors

In analyzing phylogenetic trees it is crucial to understand the types of errors that can occur and how to handle them. There are two major types of errors: stochastic errors, which arise from random noise in the data or method used, and systematic errors, which result from incorrect model assumptions or the underestimation of data complexity. Systematic errors often stem from the assumption of homogeneity in the model, whereas real-world data exhibit higher complexity, leading to incorrect results. While some errors can be easily avoided, for example, by integrating them into the model, others may be far more challenging to identify and rectify. Despite this, it remains crucial to consider them in phylogenetic analysis. The following list some of the most common systematic errors in phylogenetic tree analysis (Kapli, 2020; Felsenstein, 1978; Philippe, 2005):

Horizontal gene transfer (HGT) describes the non-sexual transfer of genetic material between organisms, primarily observed in bacteria. Another closely related phenomenon, *hybridization*, involves the combination of genomes of different species, which is particularly notable in plants. Traditional phylogenetic trees struggle to represent these events accurately since they introduce cycles and directionality, contradicting the *descent with modification* hypothesis. As a result, the presence of these events in the data conflicts with the tree inference. Recent research has focused on developing (semi-directed) phylogenetic networks instead of trees to account for these phenomena, like `PhyNEST` (Kong, 2022).

Heterogeneity of rates across lineages (or branch-length heterogeneity) is caused by unrelated fast-evolving taxa that are wrongly grouped together as a result. The phenomenon, known as long branch attraction (LBA), can be mitigated by integrating the branch length into the model or removing regions of high evolutionary rates. Typically, a combination is applied, starting with alignment trimming, followed by phylogenetic inference that accounts for branch lengths, such as with ML approaches. Methods that are especially prone to LBA are UPGMA and MP methods (Felsenstein, 1978).

Heterogeneity of character compositions describes the problem of incorrectly grouping based on similar base compositions. This error can be accounted for in the model, the so-called mixture models (Nguyen, 2015), albeit at an increased computational cost. Other common methods to handle this error include removing outliers using a composition test, such as `IQ-Tree` provides a chi-square homogeneity test on observed character frequency against the total frequency of the alignment (Nguyen, 2015). Alternatively, character states can be pooled together to counteract this bias, such as Dayhoffs six categories for amino acids based on chemical properties (Dayhoff, 1978). It is to be noted that this procedure is criticized for performing poorly as too much information is lost in the process (Hernandez, 2021).

Rate heterogeneity of sites describes the phenomenon that different parts of a gene evolve at different rates. To improve accuracy, models that account for this error, such as the free-rate model or the gamma-mixture model, are used in combination with a substitution model (Yang, 1995). Prominent phylogenetic inference tools like `RAxML` and `IQ-Tree` directly implement these models.

Incomplete lineage sorting or hemiplasy can be observed in gene families with high genetic diversity and large population size. It is caused by ancestral polymorphism that can lead to gene-tree-species-tree incongruence, where the inferred gene tree does not match the associated species tree. This error can be exacerbated in scenarios where multiple speciation events follow each other in quick succession near the ancestral species, combined with long branches at the tips. In this case, it is difficult to infer the true topology as the internal edges exhibit a complexity that is masked by mutations of the long branches while the ancestral polymorphism confounds the signals. It is worth noting that the result of incomplete lineage sorting can sometimes be mistakenly interpreted as a horizontal gene transfer event, as the results can be similar (Kapli, 2020).

Supermatrix Analysis

Supermatrix or Supertree analysis is a commonly employed approach for inferring species trees from a given set of species. First, an orthology analysis is conducted using tools like `Proteinortho`. Next, the core-groups, which cover all species, are aligned individually and subsequently concatenated into a so-called supermatrix (Kapli, 2020). The supermatrix is then used to infer a single phylogenetic tree reflecting the evolutionary history of the species. This approach is more robust to outliers than using a single gene family like 16S RNA to infer the species tree and thus enhances the reliability of the resulting phylogeny (Kapli, 2020). Nevertheless, finding concise groups of orthologs covering all species is challenging. A workflow can utilize `Proteinortho`, to generate the set of ortholog core-groups, `muscle`, to align the groups, and `IQ-Tree`, to infer a tree from the supermatrix.

Furthermore, the supermatrix can be combined with an existing topology, such as the `Open Tree of Life: Synthetic Tree` (OpenTree, 2021), which is derived from published trees. The tree inference tool then uses this tree as a fixed topology and infers the branch lengths from the supermatrix.

Reconciliation Analysis

Reconciliation is a powerful method used to combine the evolutionary scenario of a gene family with the corresponding species tree, as shown in the example in Fig. 6. The reconciliation method seeks to rearrange the gene tree to fit the species tree by introducing new events to the tree and simultaneously minimizing the associated cost of these events (Menet, 2022). For example, the DTL model incorporates the events of **d**uplication, horizontal gene **t**ransfer, and gene **l**oss, in addition to speciation events (Menet, 2022). Speciation events are represented by nodes, where the gene tree and species tree align, while duplication events are depicted by nodes of the gene tree placed on the edges of the species tree. HGT events connect different branches of the species tree, and loss events are usually marked with an `X` as shown in Fig. 6. One tool for reconciliation analysis is `GeneRax` (Morel, 2020), which implements different evolutionary models like the DTL. Typically, this type of analysis builds on top of the classical gene tree inference in combination with a supermatrix analysis that is used to generate a species tree.

By combining information from the species tree and the gene of interest, a reconciliation analysis enables the interpretation of the internal nodes in the gene tree regarding the introduced events like speciation or duplication. These events can be used to refine orthology groups as it is utilized in `OrthoFinder2` (Emms, 2018). Furthermore, the reconciliation can be used to generate a nomenclature of the genes, as described in the following. The ancient speciation events hold limited information as they result in disjoint sets of species of the proteins below. On the other hand, primal duplication events may be associated with functional divergence, which divides the genes into distinct classes (Ohno, 1970). For example, in Fig. 6, the two early duplication events define three classes (a1 and c1), (a2 and b1) and (b2, c2, and d1), in which the gene d2 is not directly assignable. Consequently, this information can be used to establish a nomenclature for the protein or gene family under investigation.

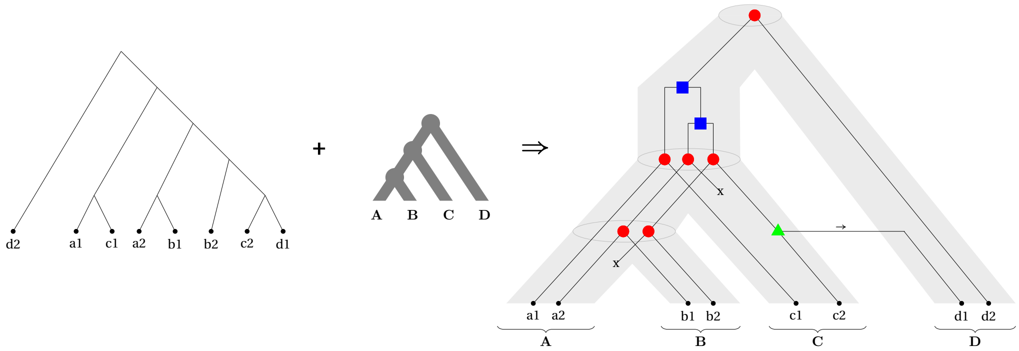


Figure 6: Schematic representation of a reconciliation. The evolutionary history of the genes {a1,a2,b1,b2,c1,c2,d1,d2} (left) embedded in the species tree (middle) of the corresponding species {A,B,C,D}. Gene losses are marked with X. Red circle: speciation event, blue square: duplication event, green triangle: HGT event.

1.3 Mathematical Framework

1.3.1 Spectral Clustering

The goal of graph clustering is to identify groups of nodes that are densely connected within the same group while having sparse connections to other groups. While this task may seem trivial for small graphs from a human perspective, it presents algorithmic challenges when considered from a computational standpoint.

Graph clustering is a powerful approach widely used in various applications, including image processing for object clustering and graph-based orthology inference. As described in chapter 1.2.2, the graph-based orthology inference first builds up a graph of putative orthologs using an all-versus-all BLAST approach. However, as the number of species increases, the connected components in the graph tend to grow rapidly, resulting in massively connected components. This is also known as the small world phenomenon (Milgram, 1967) and is demonstrated with some real-world examples in Fig. 7. Therefore, it becomes essential to prune the homology graph into smaller groups representing orthologs groups.

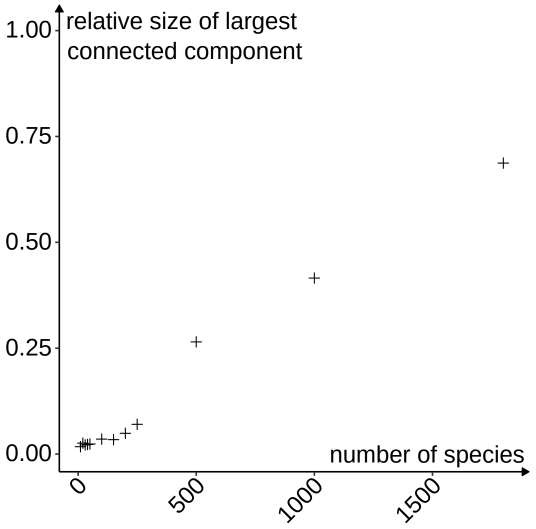


Figure 7: The size of the largest connected components relative to the total number of nodes from randomly selected bacterial proteomes of UniProt. The graphs were built using `Proteinortho` with default parameters. More details in the supplement of the third article 2.3.

Spectral clustering offers an effective method for addressing this problem. It leverages the characteristics, or spectrum, of the underlying adjacency matrix and employs techniques from linear algebra to dissect the graph. The main idea of spectral clustering can be traced back to Fiedler's work (Fiedler, 1973). This chapter is aimed to highlight the core concept behind this approach and provide a simplified motivation based on the works of Fiedler (1973), Shi (2000), and Miettinen (2017).

Min-Cut Problem

The Minimum-Cut problem provides a formalized way of expressing the graph clustering objective: Remove as few edges as possible from a graph to obtain two new connected subgraphs. Let V denote a set of nodes of a graph. One way to achieve this is by finding two sets of nodes, denoted as P and Q with $P \cup Q = V$, such that the number of edges between them is

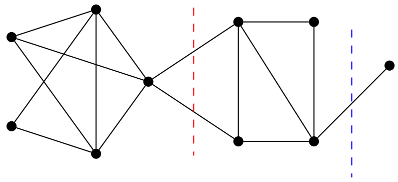


Figure 8: Example graph with 10 nodes. The most optimal split minimizing the number of edges removed is indicated in blue and the most balanced split in red (ratio-cut).

minimal. This can be formulated as an optimization objective:

$$\min_{P,Q} \text{cut}(P,Q) := \min_{P,Q} \sum_{i\in P, j\in Q} a_{i,j}$$

Here, $a_{i,j}$ is 1 if there is an edge between nodes i and j, and 0 otherwise.

While the cut objective initially seems suitable for clustering graphs, it is easy to find examples where the cuts could be more optimal. For instance, in Fig. 8, removing the isolated node is preferred over the split of the two highly connected components, as it involves removing only one edge instead of two. To address this, a normalized version of the cut objective is often used, incorporating a measure of the size of the resulting clusters. For example, the ratio-cut is defined as:

$$\text{ratio-cut}(P,Q) := \frac{\text{cut}(P,Q)}{|P|} + \frac{\text{cut}(P,Q)}{|Q|}$$

Here, $|P|$ denotes the number of nodes in the set P. With this modification, imbalanced splits are punished and in the example of Fig. 8 the isolated node is not split first. Another normalization approach is the norm-cut, defined as:

$$\text{norm-cut}(P,Q) := \frac{\text{cut}(P,Q)}{vol(P)} + \frac{\text{cut}(P,Q)}{vol(Q)}$$

where $vol(P)$ represents the sum of the node degrees for all nodes of P. Because solving the normalized cut problems directly is computationally complex, which is classified as NP-hard (Shi, 2000), it becomes essential to adopt a relaxed approach to make feasible algorithms possible.

From Ratio-Cut to Eigenproblem

First, the definition of the normalized cuts will be simplified by omitting Q, which is always implicitly given by the complement of P. For example, the ratio-cut can be reformulated as follows:

$$\frac{\text{cut}(P,Q)}{|P|} = \frac{\text{cut}(P,\overline{P})}{|P|}$$

where $\overline{P}$ represents all nodes except those in P or simply the nodes of Q. Note that the definition of the normalized-cut involves two terms that require reformulation. To simplify matters, we will omit the second term. The term $cut(P,\overline{P})$ still describes the number of edges between P and Q and can be reformulated as the number of edges of P with any node of the graph minus the number of edges within P or

$$\mathrm{cut}(P, \overline{P}) = \mathrm{cut}(P, Q) = \mathrm{cut}(P, V) - \mathrm{cut}(P, P)$$

where V denotes the set of all nodes. Let $\mathbf{x}$ be the vector that encodes the split between A and B, where $x_i = 1$ if $i \in P$ and 0 otherwise. The term $\mathrm{cut}(P, V)$ can be represented with the matrix of node degrees $\mathbf{D}$, where $D_{i,i} = degree(i)$ and 0 otherwise.

$$\mathrm{cut}(P, V) = \sum_{i \in P} \sum_{j \in V} a_{i,j} = \mathbf{x}^T \mathbf{D} \mathbf{x} \tag{I.1}$$

The second term $\mathrm{cut}(P, P)$ can be reformulate analogously

$$\mathrm{cut}(P, P) = \sum_{i \in P} \sum_{j \in P} a_{i,j} = \mathbf{x}^T \mathbf{A} \mathbf{x} \tag{I.2}$$

Here, $\mathbf{A}$ denotes the adjacency matrix. Combining I.2 and I.1, obtaining:

$$\mathrm{cut}(P, \overline{P}) = \mathbf{x}^T \mathbf{D} \mathbf{x} - \mathbf{x}^T \mathbf{A} \mathbf{x} = \mathbf{x}^T (\mathbf{D} - \mathbf{A}) \mathbf{x}$$

The L2-norm of $\mathbf{x}$, denoted by $||\mathbf{x}||$, corresponds to the Euclidean norm and therefore $||\mathbf{x}||^2$ corresponds the size of P. With this, the ratio-cut(P) can be rewritten:

$$\frac{\mathrm{cut}(P, \overline{P})}{|P|} = \frac{\mathbf{x}^T (\mathbf{D} - \mathbf{A}) \mathbf{x}}{||\mathbf{x}||^2}$$

In total, the clustering problem can be formulated as

$$\min_{x \text{ binary}} \frac{\mathbf{x}^T (\mathbf{D} - \mathbf{A}) \mathbf{x}}{||\mathbf{x}||^2}$$

Relaxation and Solution

The problem remains as difficult as before, as the *cut* statement is only reformulated. Therefore, by relaxing the constraint that the clustering vector $\mathbf{x}$ hold binary values (indicating which nodes belong to one side of the split P) the ratio-cut can be rewritten as the following:

$$\frac{\mathbf{x}^T (\mathbf{D} - \mathbf{A}) \mathbf{x}}{||\mathbf{x}||^2} = (\frac{\mathbf{x}}{||\mathbf{x}||})^T (\mathbf{D} - \mathbf{A}) \frac{\mathbf{x}}{||\mathbf{x}||}$$

This relaxation now allows us to rewrite the problem using a normalized length vector $\mathbf{u} := \frac{\mathbf{x}}{||\mathbf{x}||}$ and the Laplacian matrix $\mathbf{L} := \mathbf{D} - \mathbf{A}$, where $\mathbf{D}$ is the degree matrix and $\mathbf{A}$ is the adjacency matrix:

$$\mathbf{u}^T \mathbf{L} \mathbf{u}$$

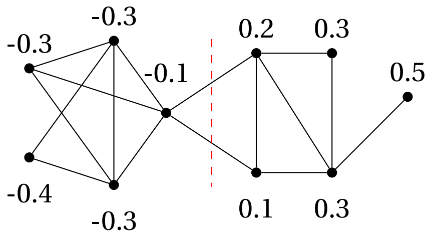


Figure 9: Example graph of Fig. 8, where each node is assigned the corresponding value of the Fiedler vector calculated using `Lapack ssyevr` (algebraic connectivity: 0.04). The entries with positive and negative entries define the two clusters.

Note that this would not be possible under the restriction of binary vectors. So, in summary:

$$\min_{\mathbf{u},\mathbf{u}^T\mathbf{u}=1} \mathbf{u}^T\mathbf{L}\mathbf{u} \tag{I.3}$$

The constraint comes from the definition of **u** (normal length).

Following Miettinen (2017), the formulated optimization problem of I.3 can be approximated with the Lagrangian Relaxation, that integrates the constraint directly into the objective with a penalty multiplier λ:

$$\min_{\mathbf{u}}(\mathbf{u}^T\mathbf{L}\mathbf{u} - \lambda(\mathbf{u}^T\mathbf{u} - 1))$$

The Lagrangian Relaxation provides a weak solution, meaning that any solution to the original problem implies a solution of the relaxation but not necessarily vice versa. A strong but more technical version can be found in the works of Fiedler (1973) or Shi (2000). To find the optimal solution, a minimum of this relaxed objective function is obtained by taking the partial derivative with respect to **u** and setting it to zero:

$$\frac{\delta}{\delta\mathbf{u}}(\mathbf{u}^T\mathbf{L}\mathbf{u} - \lambda(\mathbf{u}^T\mathbf{u} - 1)) = 0$$

Hence, with

$$\mathbf{L}\mathbf{u} = \lambda\mathbf{u} \tag{I.4}$$

the terms with **u** vanish and the derivative is zero. The formulation of I.4 is well known as an eigenproblem of which λ is an eigenvalue of **L** and **u** is the corresponding eigenvector. By the properties of the Laplacian L the smallest eigenvalue of a Laplacian is always 0 with a constant eigenvector, which does not help in the context of graph clustering. Therefore the second smallest eigenvalue (called algebraic connectivity) is used to approximate the min-cut problem and ultimately cluster the graph. The positive and negative entries of the associated eigenvector (called Fiedler vector) then give a bi-partitioning of the graph as shown in Fig. 9. Furthermore, the algebraic connectivity directly reflects the connectivity of the graph. Higher algebraic connectivity values indicate better graph connectivity, and in such cases, a cut may not be as helpful for clustering purposes. On the other hand, when the algebraic connectivity is low, it suggests the presence of well-separated clusters in the graph. For example, the graph

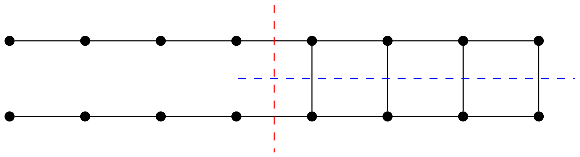


Figure 10: The cockroach graph. An optimal split is indicated in red which removes the legs from the body, while blue shows a suboptimal one. The spectral clustering approach can favor the blue split (Miettinen, 2017)

with all edges (Kuratowski graph) produces the highest possible algebraic connectivity of 1.

In summary, by finding the eigenvalues and eigenvectors of the Laplacian matrix **L**, the relaxed min-cut problem can be solved and the solution approximates the graph clustering task. It is important to note that the algebraic connectivity and Fiedler vector only provide approximations to the min-cut problem and may produce non-optimal splits in some instances as demonstrated with the cockroach graph Fig. 10 (Miettinen, 2017).

Eigenproblem

The eigenproblem, that is, the problem of finding eigenvalues and vectors of a given matrix, is a well-known and studied topic. There are plenty of different approaches to solving this:

The power iteration is a method for finding the largest eigenvalue of an eigenproblem. This method is based on the idea of iteratively multiplying a matrix with a vector and normalizing the resulting vector. By repeating this process, the method converges to the largest eigenvalue. It's simplicity and low memory requirements make it a popular choice for various applications in eigenvalue computations, one of which is `Proteinortho` (version v5), which utilizes the power iteration method to cluster the RBH graph. It employs a two-step approach, first transforming the Laplacian matrix to ensure that the algebraic connectivity becomes the largest eigenvalue and then applying the power iteration method. This approach offers the advantage of requiring only the edges present in the graph rather than the full adjacency matrix (Lechner, 2011). Consequently, this method significantly reduces memory footprint, enabling efficient processing of large matrices.

Another approach that is implemented in `Proteinortho` to solve the eigenproblem is utilizing the highly optimized Fortran library `Lapack` (Anderson, 1999). `Lapack` is a widely-used linear algebra library that provides efficient and reliable implementations of various numerical algorithms. Within `Lapack`, the `ssyevr` routine is designed specifically for solving the symmetric eigenvalue problem based on the *Relatively Robust Representation* (RRR) algorithm (Parlett, 2000). This algorithm is known for its ability to compute eigenpairs in linear time (Bientinesi, 2005), making it highly efficient for large-scale problems. On the flip side, RRR requires the full Laplacian matrix and thus has a higher memory consumption in total. A comparison between the two approaches is discussed in the third article.

II

Results and Discussion

This chapter provides the three research articles. Additional results that highlight the connection between the three topics, give further details on the discussed topic or provide additional information about the phylogenetic performance can be found in appendix A1, A2, and A3 as well as the supplementary repository[15].

[15] https://gitlab.uni-marburg.de/synmikro/ag-lechner/paul-klemm-dissertation-supplement

2.1 Kiwellins in Embryophyta 🥝

Original Research Article: **Klemm, P.**, Christ, M., Altegoer, F., Freitag, J., Bange, G., & Lechner, M. Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family. *Frontiers in plant science* 13 (2022): 4832. doi: 10.3389/fpls.2022.1034708

frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 22 December 2022
DOI 10.3389/fpls.2022.1034708

Check for updates

OPEN ACCESS

EDITED BY
Artemio Mendoza-Mendoza,
Lincoln University, New Zealand

REVIEWED BY
Wenwu Zhou,
Zhejiang University, China
Prithwi Ghosh,
Vidyasagar University, India

*CORRESPONDENCE
Marcus Lechner
lechner@staff.uni-marburg.de
Gert Bange
gert.bange@synmikro.uni-marburg.de

[†]These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION
This article was submitted to Plant Pathogen Interactions, a section of the journal Frontiers in Plant Science

RECEIVED 01 September 2022
ACCEPTED 09 November 2022
PUBLISHED 22 December 2022

CITATION
Klemm P, Christ M, Altegoer F, Freitag J, Bange G and Lechner M (2022) Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family. *Front. Plant Sci.* 13:1034708. doi: 10.3389/fpls.2022.1034708

Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family

Paul Klemm[1†], Marvin Christ[1†], Florian Altegoer[2], Johannes Freitag[3], Gert Bange[1,4*] and Marcus Lechner[1*]

[1]Center for Synthetic Microbiology (SYNMIKRO), Philipps-University Marburg, Marburg, Germany, [2]Institute of Microbiology, Heinrich Heine University Dusseldorf, Düsseldorf, Germany, [3]Department of Biology, Philipps-University Marburg, Marburg, Germany, [4]Molecular Physiology of Microbes, Max-Planck Institute for Terrestrial Microbiology, Marburg, Germany

Crop diseases caused by pathogens critically affect global food security and plant ecology. Pathogens are well adapted to their host plants and have developed sophisticated mechanisms allowing successful colonization. Plants in turn have taken measures to counteract pathogen attacks resulting in an evolutionary arms race. Recent studies provided mechanistic insights into how two plant Kiwellin proteins from *Zea mays* mitigate the activity of the chorismate mutase Cmu1, a virulence factor secreted by the fungal pathogen *Ustilago maydis* during maize infection. Formerly identified as human allergens in kiwifruit, the biological function of Kiwellins is apparently linked to plant defense. We combined the analysis of proteome data with structural predictions to obtain a holistic overview of the Kiwellin protein family, that is subdivided into proteins with and without a N-terminal kissper domain. We found that Kiwellins are evolutionarily conserved in various plant species. At median five Kiwellin paralogs are encoded in each plant genome. Structural predictions revealed that Barwin-like proteins and Kiwellins cannot be discriminated purely at the sequence level. Our data shows that Kiwellins emerged in land plants (embryophyta) and are not present in fungi as suggested earlier. They evolved *via* three major duplication events that lead to clearly distinguishable subfamilies. We introduce a systematic Kiwellin nomenclature based on a detailed evolutionary reconstruction of this protein family. A meta-analysis of publicly available transcriptome data demonstrated that Kiwellins can be differentially regulated upon the interaction of plants with pathogens but also with symbionts. Furthermore, significant differences in Kiwellin expression levels dependent on tissues and cultivars were observed. In summary, our study sheds light on the evolution and regulation of a large protein family and provides a framework for a more detailed understanding of the molecular functions of Kiwellins.

KEYWORDS

Kiwellins, plants, interaction, pathogen, symbionts, evolution, classification, nomenclature

Introduction

A better understanding of plant diseases caused by viruses, bacteria, and fungi as well as by oomycetes is critical to improve global food security. In many cases, pathogens employ secreted effector proteins to manipulate the host plant and promote infection (Stergiopoulos and de Wit, 2009; Rocafort et al., 2020). Effectors exhibit a wide range of functions, e.g. they can mask the pathogen, (down)-regulate host defense mechanisms, or target defense enzymes or toxins to render them harmless (Lanver et al., 2018). In turn, plants have evolved various defense mechanisms. Upon pathogen contact, pattern recognition receptors (PRR) in the plant plasma membrane recognize conserved molecules on the surface of the microorganisms, such as flagellin, chitin, or glucans from bacteria, fungi, and oomycetes respectively (Jones and Dangl, 2006; Cook et al., 2015). These microorganism-associated molecular patterns (MAMPs) lead to MAMP-triggered immunity (MTI) (Cook et al., 2015) triggering response mechanisms to restrict damage caused by the pathogen. Plants can furthermore recognize effectors that are secreted or translocated into the plant cytoplasm, resulting in the activation of a second layer of defense, the so-called effector-triggered immunity (ETI) (Du et al., 2016). Both types of responses are tightly interconnected and thus referred to as the plant immune system (Jones and Dangl, 2006; Nguyen et al., 2021). Recently, two studies suggested a crucial role for maize Kiwellins as proteins counteracting pathogen attack (Han et al., 2019; Altegoer et al., 2020).

Two *Z. mays* Kiwellins specifically bind to the secreted chorismate mutase Cmu1 of the smut fungus *U. maydis* and inhibit its enzymatic activity (Han et al., 2019; Altegoer et al., 2020). Cmu1 was shown previously to down-regulate salicylic acid synthesis in the host by diverting its substrate chorismate to the phenylpropanoid pathway, thereby decreasing maize resistance to *U. maydis* (Djamei et al., 2011). Kiwellins were originally identified in kiwifruit (*Actinidia spp.*) in which they account for about 30% of the total protein content (Tamburrini et al., 2005). Kiwifruit can cause allergies in humans (Fine, 1981; Wang et al., 2019). It was shown that Kiwellin proteins contribute to the allergic response and are recognized by immunoglobulin E (Tamburrini et al., 2005; Ciardiello et al., 2009). The crystal structure of a Kiwellin from *Actinidia chinensis* revealed that it is a modular protein formed by an N-terminal 4 kDa kissper domain and a C-terminal core domain (Hamiaux et al., 2014). Pore-forming activity was reported for the kissper domain in synthetic lipid-bilayers while the Kiwellin-core-domain contains a double-psi β-barrel fold and a β-hairpin (Tuppo et al., 2008). Kiwellins have a high structural similarity to another class of plant defense proteins termed Barwin. Barwin and Barwin-like proteins are pathogenesis-related (PR) proteins belonging to the PR4 family. This family is divided into two classes, Barwin-like proteins with a chitin-binding domain (class I) and without this domain (class II) (Sinha et al., 2014). These proteins are mainly found in plants but also occur in bacteria, algae, and fungi. The functions of the Barwin domain are manifold. They can bind sugars, cleave RNA and DNA depending on divalent cations, and show antifungal activity (Dabravolski and Frenkel, 2021).

Due to the identification of a biological role of Kiwellins as plant defense molecules and their widespread appearance in the kingdom of plants (Bange and Altegoer, 2019; Han et al., 2019) it was tempting to speculate about an evolutionarily conserved role of Kiwellins as regulators of biotic interactions. Therefore, we set out for a systematic phylogenetic and structural investigation of the Kiwellin protein family to provide a framework for further research on this large yet relatively uncharacterized group of proteins. We provide a detailed phylogenetic reconstruction of Kiwellin evolution based on published proteome data and introduce a nomenclature for these proteins. In addition, we show that many proteins annotated as Kiwellin-like are actually Barwin-like proteins and that Kiwellins are probably restricted to land plants. Finally, we reanalyzed 31 publicly available transcriptome data obtained from plants exposed to biotic and abiotic stresses. This uncovered remarkable transcriptional regulation patterns for Kiwellin encoding transcripts upon interaction of plants with microorganisms. These data hence suggest a crucial role of Kiwellin proteins as modulators of plant-microbe interactions.

Material and methods

Kiwellin annotation

Kiwellins were annotated in all complete reference proteomes provided by UniProt, v2022_01 (Consortium, 2019). We conducted an iterative procedure that combines an initial sequence-based model with a structure-based filtering strategy. Initially, published sequences from Han et al. (2019) were aligned using muscle v3.8.1551 (Edgar, 2004). A profile Hidden Markov Model was built from this alignment using HMMer v3.2.1 (Eddy, 1998) and then used to mine the proteomes of all kingdoms of life (adaptive e-value cutoff based on false positives, see Supplementary Data 1 chapter 4). The resulting proteins were trimmed at the signal peptide cleavage site which was predicted with SignalP v5.0b (Almagro Armenteros et al., 2019) if present within the first half. The region upstream of this site is cleaved *in vivo* and is thus not

Abbreviations: Kwl, Kiwellin; dpi, days post-infection; hpi, hours post-infection; LCA, lowest common ancestor; BL, Barwin-like; DPBB, double-psi β-barrel; FUN, Fungi; BRY, Bryophyta; LYC, Lycopodiopsida; AMB, Amborellales; LIL, Liliopsida; RAN, Ranunculales; MAG, Magnoliidae; SAX, Saxifragales; ROS, Rosids; CAR, Caryophyllales; AST, Asterids.

relevant for structural prediction. We employed AlphaFold2 v2.0.0 (Jumper et al., 2021) to accurately assess structural elements, which were then used to determine the fold type, see Figure 1. Predicted structures were superimposed with the Kiwellin crystal structure provided by Han et al. (2019) using PyMOL (DeLano and Bromberg, 2004) and compared based on the presence of structural elements, RMSD and overlap (see Supplemental Data 1 chapter 4 for details). In this way, false-positive hits, e.g. Barwin-like proteins, could be distinguished from Kiwellins and Kissper-Kiwellins. The process was iterated multiple times, further improving the profile Hidden Markov Model. We trained separate models for Kiwellins and Kissper-Kiwellins. In parallel, we trained a model for false positives e.g. Barwin-like proteins that allowed us to filter out members of this group already at the sequence-based stages.

In total, we identified 915 Kiwellins in 142 land plants (embryophyta) and one fungal species. No Kiwellins were detected in bacteria, archaea, or viruses/phages. A detailed description of the pipeline can be found in Supplementary Data 1 chapter 4. The implementation of the workflow along with the models generated at the last iteration is available *via* our GitLab repository, see *Data Availability Statement*.

Evolutionary reconstruction

The evolution of Kiwellins was reconstructed based on a phylogenetic gene tree that was modeled onto the associated species tree. Phylogenetic conflicts that arise in this process are resolved using a maximum likelihood-based approach concerning speciation, duplication, or loss events within the gene lineages.

First, all 915 Kiwellins with trimmed signal peptides were aligned using muscle v3.8.1551 (Edgar, 2004) to reconstruct the gene tree. The best-fit model for amino acid replacement with respect to the Bayesian information criterion (BIC) was determined using IQ-TREE v2.0.3 (Nguyen et al., 2015). The phylogenetic tree was then reconstructed with 100k bootstrap iterations based on the reported general amino-acid exchange rate matrix (WAG) (Whelan and Goldman, 2001) with a FreeRate model (Yang, 1995) of 8 categories of rate heterogeneity across sites, WAG+R8.

Second, the species tree was compiled based on the Open Tree of Life Synthetic Tree v13.4 (OpenTree, 2019a). The relevant subtree was re-rooted and pruned to the species with annotated Kiwellins within the reference proteomes provided by UniProt using ete3 v3.0.0b34 (Huerta-Cepas et al., 2016) based on the open tree taxonomy v3.3 (OpenTree, 2019b). To reduce the complexity, cultivars/subspecies were merged into a single species node representing all strains. As the Synthetic Tree does not encode distances for any species, these were estimated based on their core proteome. Orthologous proteins were determined using Proteinortho v6.1.1 (Lechner et al., 2011) with a fairly high e-value of 10^{-20}. The 204 orthologous groups that covered all species were aligned using muscle v3.8.1551 (Edgar, 2004). Alignment columns with more than 90% gap content were dismissed, to reduce complexity. The alignments of all orthologous groups were then concatenated to reconstruct a supertree covering the core proteome. The best-fit model for amino acid replacement with respect to BIC was determined using IQ-TREE v2.0.3 (Nguyen et al., 2015). The reported model was JTT (Jones et al., 1992) with a FreeRate model (Yang, 1995) of 9 categories of rate heterogeneity across the site and empirical base frequencies, JTT+F+R9. It was used to reconstruct the phylogenetic tree, constrained by the topology of the species tree.

Next, GeneRax v2.0.4 (Morel et al., 2020) was used for a species-tree-aware Maximum Likelihood-based gene family tree inference using the UndatedDL reconciliation model (including speciation, duplication and loss events). 20 iterations were performed to reconcile the gene and species trees. This evolutionary reconstruction was then visualized using iTOL v6 (Letunic and Bork, 2021). It forms the foundation of the Kiwellin nomenclature.

Nomenclature

Based on early duplication events, three major Kiwellin groups were identified Kwl1, Kwl2, and Kwl3. Kwl1 is closest to the lowest common ancestor node (LCA) that was predicted in the analysis and thus represents the primal Kiwellin subfamily. Kwl3 is the most recent subfamily. Within these major groups, the following duplication events with a representative number of species covered were used to further refine subgroups, e.g. Kwl1-1, Kwl1-2, and so on. Paralogs within species are then numbered (ascending by age/distance to LCA) using letters, e.g. Kwl1-1a, Kwl1-1b, and so on. A notable exception is Kwl3-1 which was grouped based on a speciation rather than a duplication event. However, this subfamily is specific to Liliopsida and was therefore threaded separately.

To ease reading and summarize our findings, species were matched to taxonomic groups according to the NCBI taxonomy (Sayers et al., 2021). These groups are referred to by a three-letter code. The abbreviations used here are as follows: Fungi (FUN), Bryophyta (BRY), Lycopodiopsida (LYC), Amborellales (AMB), Liliopsida (LIL), Ranunculales (RAN), Magnoliidae (MAG), Saxifragales (SAX), Rosids (ROS), Caryophyllales (CAR), and Asterids (AST).

Consensus

Kiwellin sequences with trimmed signal peptides were grouped according to their respective major subfamilies (Kwl1, Kwl2, Kwl3, see 'Nomenclature' above). As most but not all Kwl1

Kiwellins contain a kissper domain, this subfamily was further divided in Kissper-Kwl1 and Kwl1 (without kissper). To emphasize the major differences, Barwin-like proteins identified through the Barwin-Model (see 'Kiwellin annotation' above) were added as an additional group. Note that this set is biased as it was constructed from false positive Kiwellin annotations to discriminate Barwin-like proteins from Kiwellins already at the sequence level. The groups were aligned using muscle v3.8.1551 (Edgar, 2004). Alignment columns with more than 90% gap content were dismissed, to reduce complexity.

A consensus sequence was calculated for each group and visualized using jalview v2 (Waterhouse et al., 2009). The conservation score of the Kiwellin consensus sequences was calculated following Livingstone and Barton (1993). 1 to 9 indicates property conservation of the alignment column in ascending order. Full property-related conservation is indicated by +, perfect amino acid conservation by *. Physico-chemical properties are highlighted based on the color schema provided by Larkin et al. (2007) (known e.g. from clustalX). The full alignments are available *via* our GitLab repository, see *Data Availability Statement*.

Transcriptomics

Publicly available RNA-seq data sets in NCBI SRA (Sayers et al., 2021) were identified that were created to study either pathogenic or symbiotic interactions in at least two biological replicates. A complete list is compiled in the Supplemental Data. The amino acid sequences of the respective Kiwellins were mapped to the respective transcripts either based on the NCBI transcriptome (Sayers et al., 2021) or, if not available, based on the cDNA sequences provided by Ensembl Plants (Yates et al., 2022) using Proteinortho in autoblast mode (to match translated DNA/RNA with amino acid sequences) with a relaxed minimal sequence coverage of 20% and rather strict minimal percent identity of 90% and e-value threshold of 10^{-50}, see Supplementary Data 3.

RNA-seq libraries were quality trimmed using trim_glalore v0.4.4 Krueger et al. (2021) and sequencing adapters were removed using cutadapt v2.3 (Martin, 2011). For paired-end sequenced experiments, the mate reads are omitted. Reads were mapped to the transcriptome of the plant under study and its pathogenic or symbiotic partner organism if available using bwa v0.7.17 Li (2013) with default parameters. The average number of mapped transcripts per data set can be found in Supplementary Data 3. It was not always possible to assign a read to a single transcript due to close sequence similarity. In these cases reads with multiple hits were accounted proportionate to the targets. However, we conservatively neglected reads that were mapped to a Kiwellin and a non-Kiwellin transcript to avoid linking measured expression of two or more genes between both groups.

Differential gene expression analysis was performed using DESeq2 v1.22.2 (Love et al., 2014). To reduce background noise, transcripts with less than 20 reads combined for all replicates and conditions were neglected. Technical replicates were collapsed using the collapseReplicates routine. For each dataset, we picked relevant replicates and conditions to compare control or mock-treated versus infected or treated in pairwise analyses (Wald test). A detailed listing is provided in the Supplemental Datas 1, 3. Transcripts with a baseMean (a proxy for overall expression strength) above 80 were considered highly expressed. Only significantly regulated Kiwellin transcripts with a P-value below 5% and an absolute log2-fold-change of at least 1 were considered for further evaluation.

Results

Structural characteristics of Kiwellins, Kissper-Kiwellins and Barwin-like proteins

To get insights into the Kiwellin family of proteins we first worked out the structural characteristics. Approaches used to identify Kiwellins so far did not include this parameter to separate this protein family from Barwin-like proteins. Figure 1 highlights distinct structural features. Over 90% of the identified Kiwellins contain a signal peptide and thus can be secreted from the cell *via* the conventional pathway. The core of a Kiwellin protein is about 110 aa long. It consists of three α-helices and six parallel and antiparallel connected β-strands, that form a so-called double-psi β-barrel. This type of fold is also characteristic for the superfamily of Barwin-like proteins (Figure 1B). Two disulfide bonds between the two smallest β-strands $\beta5$ and $\beta6$ provide additional stability. The long flexible loop connecting $\alpha1$ and $\beta3$ is fixed to the barrel by another disulfide bridge, which is anchored between $\beta1$ and $\beta2$.

In contrast to Barwin-like proteins, Kiwellins have an additional N-terminal extension of about 25 to 45 aa (Figure 1B). This domain consists of two β-strands connected by a loop. The loop region between $\beta3'$ and $\beta4'$ is highly variable. It is stabilized by disulfide bridges at both ends of the sheets that allow linkage of the extension with external loops of the β-barrel. Another disulfide bridge connects the loop located between the two β-strands that form the β-hairpin to the loop between $\beta5$ and $\beta6$, likely to connect this flexible and long loop to the core of the protein. We will refer to this N-terminal region as the Kiwellin-extension (compared to Barwin-like proteins).

The second class of Kiwellins, the so-called Kissper-Kiwellins (Ciardiello et al., 2008) include one further N-terminal extension of about 40 aa (Figure 1C). This domain is enriched in disulfide bridges and short regions of secondary structure elements (Ciardiello et al., 2008; Hamiaux et al., 2014). Notably, the loop connecting $\beta3'$and $\beta4'$ is significantly shorter in Kissper-Kiwellins,

Enlarged image can be found in the appendix

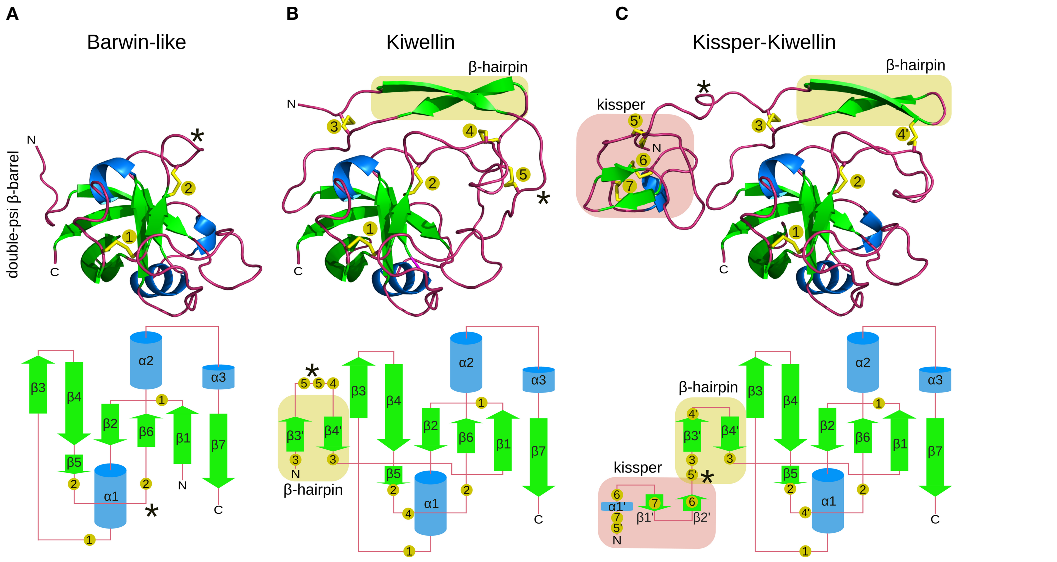


FIGURE 1
Structure-models of Barwin-like **(A)**, Kiwellin **(B)**, and Kissper-Kiwellin **(C)** proteins based on consensus sequences of all proteins identified for each group (signal peptide removed). Elements visible in the 3D structure on the top are indicated in a planar visualization on the bottom with identical coloring (green: β-sheets, blue: α-helices, red: loop regions). Highlighted in yellow is the β-hairpin and in red is the kissper domain. Numbered, yellow circles indicate the respective disulfide-boundforming cysteine residues. A loop region with variable length is indicated by *.

it decreases from 15 aa in Kiwellins to only 2 aa in Kissper-Kiwellins. This changes the arrangement of disulfide bridges in the Kiwellin-extension. The shorter loop provokes the absence of disulfide bridge 5 found in other Kiwellins and influences anchoring of disulfide bridge 4 in Kiwellins (compare the disulfide-bound forming cysteine residues 4, 4’, and 5 in Figures 1B, C). Three disulfide bridges are formed in the Kiwellin extension to stabilize the small fold (compare cysteine residues 5’, 6, and 7).

The evolution of Kiwellins

We reconstructed the phylogeny of Kiwellins and reconciled this data in the respective species tree to estimate duplication, speciation, and loss events along the evolution of this protein family. A summarized illustration is shown in Figure 2. The complete phylogenetic reconciliation including an annotation of duplication, speciation, and loss events can be found in Supplemental Data 2. The root of the tree was automatically estimated. It is located between the evolutionary oldest species in Bryophyta and Lycopodiopsida and the putative fungal prediction which we show here for the sake of completeness. We propose that Kiwellins close to the root represent the primal instances of this protein family.

Our analysis revealed three initial duplication events and we, thus, distinguish three major Kiwellin groups: Kwl1, Kwl2, and Kwl3 (see *Materials and methods* for details). Kwl1 is the most ancient group probably representing the original Kiwellin subfamily. It is present in most taxonomic groups and is enriched in evolutionary older groups. It is found in younger groups as well e.g. in rosids and asterids but was frequently lost. Out of 205 Kwl1 proteins, 137 contain a kissper domain. This domain is restricted to the Kwl1 subfamily. The enrichment of this additional domain in a specific group of plants was not observed. Notably, Kissper-containing Kiwellins are completely missing in LIL. Kiwellins with and without kissper domain are phylogenetically grouped next to each other. An alternative phylogenetic analysis based on Kiwellins with truncated kissper domains resulted in a comparable phylogeny (Figure S1), indicating that these longer sequences did not alter the phylogenetic reconstruction significantly.

The Kwl1 group spans 87 species including the oldest embryophyta *Physcomitrium patens* and *Selaginella moellendorffii* and five taxonomic groups BRY, LYC, LIL, ROS, and, AST. The Kwl2 subfamily contains 202 proteins and is restricted to a limited number of taxons. This subfamily is probably derived from a specific duplication found only in LIL, AMB, and MAG. Compared to the other groups LIL species predominantly contain Kwl2. Kwl3 is the largest group, comprising 508 Kiwellins in 115 species covering the six taxonomic groups ROS, AST, RAN, SAX, CAR, and LIL. For all taxonomic groups except LIL, Kwl3 is the overall youngest but also the most abundant subfamily.

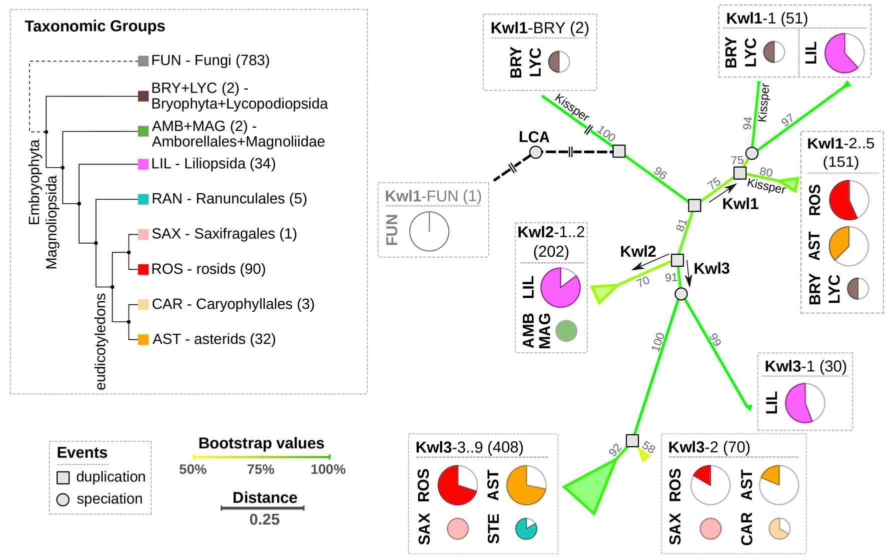


FIGURE 2
Left panel: Cladogram of the relevant taxonomic groups. The number of respective species in our data set is indicated in brackets. Right panel: Summary of the phylogenetically reconciled Kiwellin tree. The edge color and numbers refer to bootstrap percentages. Evolutionary events are indicated by a circle (speciation) or a square (duplication). Pie charts visualize the species coverage. Groups that mostly contain Kissper-Kiwellins are indicated (Kissper).

Sequence and structure conservation of Kiwellin subfamilies

The evolutionary reconciliation of Kiwellins identified three major subfamilies, Kwl1, Kwl2, and Kwl3. Kwl1 can be divided into two subgroups: one with and another without the kissper domain. Figure 3 shows a sequence-based alignment based on consensus sequences. While Kiwellin sequences are highly conserved, the subfamilies can be discriminated by the length of the variable loop region between *β3'* and 4' (Figure 4). In particular, this loop is short in Kissper-Kiwellins (median of only two amino acids). For Kwl1 this loop has a median length of 14,

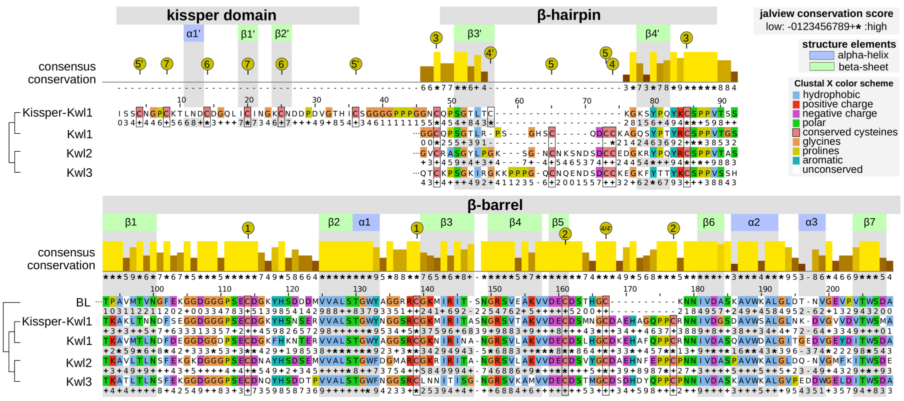


FIGURE 3
Aligned consensus sequences (without signal peptide) with secondary structure information of the Kiwellin subfamilies and a set of 391 BL proteins for reference. The conservation score of the consensus alignment for all Kiwellins is indicated above. The family-specific conservation is shown below the respective sequence: – (mostly gaps), 0...9, + (property conservation, ascending), ∗ (perfect conservation). Amino acids are colored according to their physicochemical properties. Positions and secondary structure elements were drawn corresponding to Kissper-Kwl1. Green represents β-sheets, blue α-helices. Numbered, yellow circles indicate the cysteine residues forming disulfide bounds.

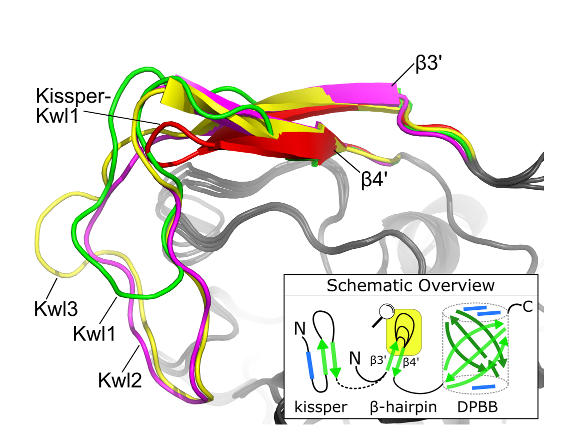


FIGURE 4
The consensus structure of the loop connecting β3′ and β4′ in the β-hairpin. The Kiwellin-groups are highlighted by different colors: red: Kissper-Kwl1, green: Kwl1, magenta: Kwl2, yellow: Kwl3. The bottom right: a schematic overview of (Kissper-) Kiwellins. Green arrow: β-sheet, blue rectangle: α-helix, yellow box: zoomed region.

Kwl2 17, and Kwl3 20 amino acids. While cysteines are highly conserved, the disulfide bridge pattern differs between Kiwellins with and without the kissper domain since the loop region in the Kissper-Kiwellins is shortened and thus contains only one cysteine. For Kiwellins without the kissper domain, the loop is extended. Hence, a total of six cysteines are found that form three additional disulfide bridges. Overall, the loop appears to be a modular region with the lowest overall sequence conservation, e.g. significantly lower than the barrel-giving *β*-sheets (one-sided Wilcoxon rank-sum test, p<0.02). Notably, this region is not present at all in BLs. Similarly, the loop between *β*5 and *β*6 is usually shorter in BLs. Both features make it possible to distinguish BL and Kiwellins. While this is sufficient in most cases, about one-third of the BLs contain a loop similar to Kiwellins (Figure S2). Notably, neither the length nor the sequence composition of this region did impact the structure predictions.

The amino acid sequences in the remaining secondary structure elements are strongly conserved. In particular, the *β*-sheets and the α-helices located in the barrel *β1*–*β7* and *α1*–*α3* are significantly conserved relative to the adjacent unstructured regions (one-sided Wilcoxon rank-sum test, p<0.05). Most of the amino acid positions in the secondary structure elements of the barrel are entirely conserved (*) especially *β*2, *α*1, *β*5 and *β*6. Less conservation was observed for other elements in the barrel. For example, *α*3, *β*3 and *β*7 have some positions with low conservation scores compared to the overall consensus, as well as in the individual subfamilies (e.g. position 1, 2 in *α*3 or 5 in *β*7).

Of interest, we identified 61 additional proteins containing a kiwellin domain as part of a larger protein distributed over 33 species with no clear taxonomic limitation. About half (30) of those proteins are considerably larger (three to four times), still, no further domain could be identified. 26 hits are duplications of the kiwellin domain (Figures S3A, B). In two cases, triplications were found (Figure S3C). Similar domain duplications were also found for Kissper-Kiwellins (Figure S3D). All identified fusion proteins are listed in Supplementary Data 3.

Dissemination

A species-wise view of Kiwellin subfamilies (Figure 5) shows that Kwl2 is exclusively found in LIL where it is typically present, apart from some *Oriza* cultivars. LIL species on the other hand do not encode any Kissper-Kiwellins. We observed two major Kiwellin loss events coinciding with a loss of BLs in the order of Brassicales (e.g. *A. thaliana*) as well as the division of Marchantiophyta. The latter is only represented by two species in our data set. Thus, a general conclusion cannot be drawn at this stage. Brassicales are represented by thirteen species. Kiwellins are present in species sharing a common ancestor with this group (e.g. *E. grandis*). Therefore, a loss event is likely and in line with a reported whole genome triplication in the group, followed by many loss events in members of Brassicaceae (Moghe et al., 2014).

The Kiwellin subfamily Kwl2 is lost in the younger taxonomic groups of ROS and AST, while the evolutionary older Kwl1 including the Kissper-Kwl1 is still present. Thus, Kwl2 was probably lost in the intermediate group of LIL. A loss of the otherwise predominant Kiwellin subfamily Kwl3 is found in the order of Cucurbitales (ROS).

We observed a median of five Kiwellins in the larger taxonomic groups LIL, ROS, and AST, however, with a significant difference in numbers at the genus level and, strikingly, already at the level of breeds and cultivars. This can be explained by the degree of genome expansion. One measure for this feature is the unreplicated haploid nuclear genome amount also known as 1C-value (Soltis et al., 2013). Minor genome expansion is reported in CAR, SAX, and AST while large expanded genomes are found in some clades, especially within LIL which is found to show an exceptionally large range of 1C-values compared to the other taxonomic groups (Leitch et al., 1998). The reported 1C ranges of AST and ROS are similar (AST: 0.3-24.8pg, ROS: 0.1-16.5pg). This coincides with the number of Kiwellins occurring in these groups (AST: 1-24 and ROS: 1-17). SAX and CAR are reported to exhibit lower 1C ranges and as well contain a below-median number of Kiwellins.

The allotetraploid pasta wheat *T. turgidum*, one of the oldest domesticated crops, is known for its potential to obtain resistance to biotic and abiotic stresses. It encodes the second-highest number of Kiwellins (28). The bread wheat *T. aestivum* (LIL) has the highest number of Kiwellins observed in our data set (52). Most belong to the Kwl2 subfamily. Notably, *T. aestivum* is an allohexaploid composed of the three species *T.*

Enlarged image can be found in the appendix

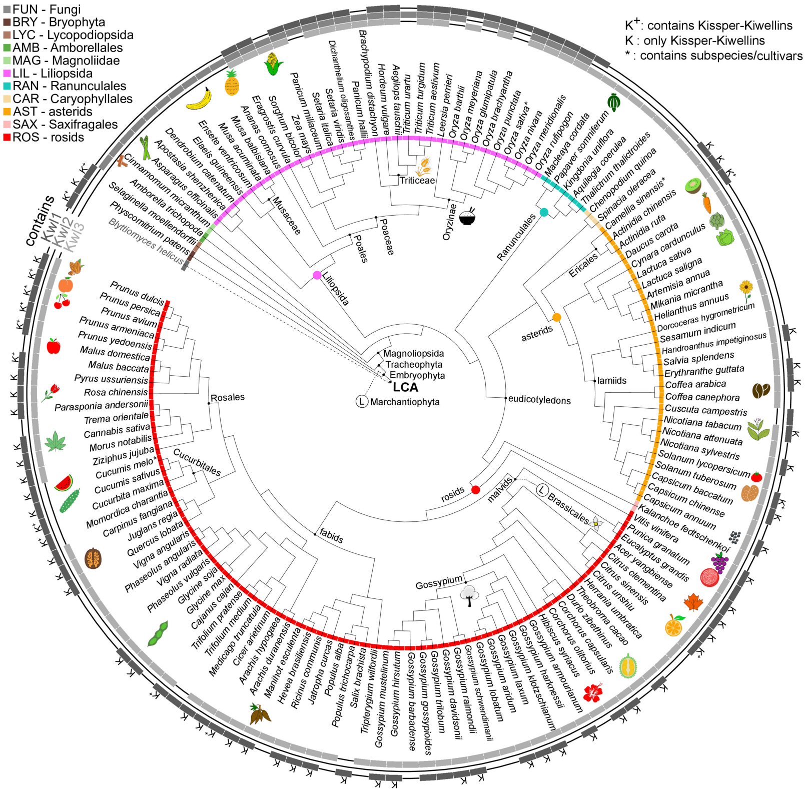


FIGURE 5
Species tree cladogram. The inner circle encodes taxonomic groups. The outer circle indicates if a species is found in the Kiwellin group Kwl1, Kwl2 or Kwl3. K$^+$: Kwl1 contains Kissper-Kiwellins, K: only Kissper-Kwl1, *: contains subspecies/cultivars, Ⓛ: putative loss event.

urartu, *A. tauschii* and an unknown close relative to *A. speltoides* (not in this analysis) (Feldman and Levy, 2012). *T. urartu* and *A. tauschii* contain Kiwellins above the median and predominantly of type Kwl2. It is thus reasonable to assume that *T. aestivum* kept most of the Kiwellins from the donor species. In contrast, *T. urartu* harbors a diploid genome (Liu et al., 2017) and codes for nine Kiwellins.

Screening and consolidation

A total of 20,630 full proteomes was screened for Kiwellins (see *Materials and methods* for details). None of the sequence-based predictions could be structurally verified in Bacteria, Archaea, or Viruses/Phages. Except for a single instance, no Kiwellins were found in the 783 fungal species in our data set. The single hit detected in Fungi is from *Blyttiomyces helicus* (A0A4P9WPM3), a saprophyte, which grows on pollen and cannot be cultured so far (Ahrendt et al., 2018). Given the phylogenetic position of this gene in our reconciliation, horizontal gene transfer from a plant or contamination is unlikely. The remaining 915 Kiwellins were identified in 142 land plants (embryophyta) (Figure 6).

The set of 'Kiwellin-like' proteins provided by InterPro (IPR039271) contains 2,362 entries that overlap to a large extent with our source data set of full proteomes (Blum et al., 2021). While it covers all Kiwellin entries identified here, it also contains many BL proteins and several unrelated proteins that we could not verify as Kiwellins. In total, we estimate only about half of the set to represent canonical Kiwellins. The published Kiwellin structure from *Actinidia chinensis* (PDB: 4PMK/Uniprot: P85261) corresponds to Kwl1-2a according to our nomenclature (Hamiaux et al., 2014). The crystal structures of *Zm*KWL1a (PDB: 6FPG/Uniprot: A0A1D6GNR3) and *Zm*KWL1b (PDB: 6TI2/Uniprot: K7U7F7) from *Zea mays* correspond to Kwl3-1b and Kwl2-2d (Han et al., 2019; Altegoer et al., 2020). *Zm*KWL4 (Uniprot: A0A1D6GNR6) corresponds to Kwl2-2e and *Zm*KWL6 as well as *Zm*KWL12 are identified as BL proteins. Similarly, a fungal rust effector protein that suppresses cell death in plants was found to be a BL protein as well (Jaswal et al., 2021). The structure from *Actinidia deliciosa* (PDB: 4X9U/Uniprot: P84527) was not recovered as the species is not part of the data set (Offermann et al., 2015). Nevertheless we identified

Enlarged image can be found in the appendix

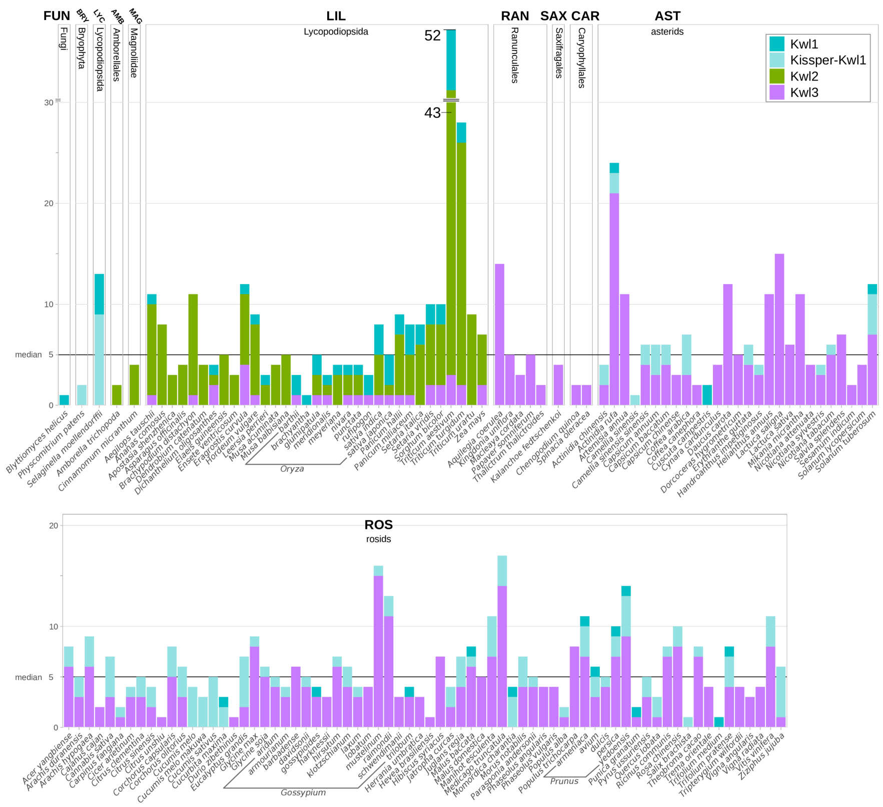


FIGURE 6
Kiwellins detected among specified species. The black horizontal line indicates the overall median of five Kiwellins per proteome. Colors indicate Kiwellin groups. Dark-turquoise: Kwl1, light-turquoise: Kissper-Kwl1, green: Kwl2, magenta: Kwl3.

this protein as a Kissper-Kwl1, orthologous to Kwl1-2a from *Actinidia chinensis*.

Meta-analysis of Kiwellin expression in biotic and abiotic interactions of plants

Kiwellins have been identified as relevant defense proteins against the pathogenic fungus *U. maydis* (Han et al., 2019). Moreover, it has also been suggested that Kiwellins show tissue-specific expression patterns (Altegoer et al., 2020). To gain a broader view of how Kiwellin proteins are expressed in various situations, we performed a meta-analysis of publicly available transcriptome datasets from NCBI SRA. We focused on experiments in which Kiwellin-containing land plants were exposed to either pathogens or symbionts. Moreover, we only selected datasets, which were documented by a publication, comprising at least two biological replicates, contained unambiguous sample and experimental descriptions, and produced reliable FastQC scores (further details in Supplementary Data 1 chapter 5). A total of 31 data sets (out of 70 initially selected) met these criteria (Table 1). 14 out of the 31 data sets showed strong expression levels, as well as significant changes in Kiwellin mRNA levels in response to interactions with symbiotic or pathogenic species. A detailed listing of individual findings and full experimental descriptions as well as the full workflow description is provided in Supplement Data 1 chapter 5. Overall, our meta-analysis revealed that strongly expressed Kiwellins are present in each of the three subfamilies Kwl1, Kwl2, and Kwl3. About half of these Kiwellins show a significant response upon the interaction of the analyzed plant species with the respective interaction partner, being either a pathogen or a symbiont (Table 1). Specifically, we detected strong and significant Kiwellin responses in 7 out of the 21 data sets analyzing a plant-pathogen interaction (see Figure S7). For the interaction of a plant with its symbiont, we have also detected strong expression levels and regulation of Kiwellin transcripts, however, only in 5 out of 15 experiments (see Figure S7). These findings might indicate a role of the Kiwellin in the interaction of plants with the cognate pathogens and symbionts as well. Taken together, our meta-analysis supports the idea that Kiwellins represent a

TABLE 1 Overview of strongly expressed and significantly regulated Kiwellin groups among all species with RNA-seq data sets.

Taxonomic Group		Species	Kwl1^{K}	Kwl1	Kwl2	Kwl3	#/total
Bryophyta (BRY)	🟫	*P. patens*	P↑*	–	–	–	1*/1
Liliopsida (LIL)	🟪	*M. acuminata*	–	–	S↑ T	–	2/3
	🟪	*Z. mays*	–	–	P↓*	P↓* P↑ S↑	2(+1*)/ 5
	🟪	*T. aestivum*	–	P↓ S↓ T	P↕ S↕ T	S↑ T	3/4
	🟪	*O. sativa*	–	P↕	∅	∅	1/2
asterids (AST)	🟧	*S. lycopersicum*	–	–	–	S↑	1/3
	🟧	*S. tuberosum*	∅	∅		∅	0/1
	🟧	*A. chinensis*	P↑*	–	–	∅	1*/1
Caryophyllales (CAR)	🟨	*C. quinoa*	–	–	–	∅	0/1
rosids (ROS)	🟥	*G. max*	P↑ S↕* T*	–	–	∅	1(+1*)/5
	🟥	*C. sativus*	S↑ P↑	–	–	–	2/2
	🟥	*M. truncatula*	∅	–	–	S↑ A↑ T	2/2
	🟥	*C. melo*	∅	–	–	–	0/1
							Σ 14 (+4*)/31

Kwl1^{K}: Kissper-Kwl1, P: pathogenic interaction, S: symbiotic interaction, T: tissue-specific, A: abiotic stress. –: Kiwellin group not present in this species, ∅: no Kiwellin found with significant regulation, ↑: up-regulated, ↓: down-regulated, ↕: up and down-regulated (compared to the respective control), *: weakly expressed but with significant differences, #: number of independently collected data sets with significantly regulated and strongly expressed Kiwellins.

regulatory layer in the plant-microbe interactions, although additional experiments are required to further consolidate this notion. We would also like to note that we observed a tissue-specific expression difference in three distinct species *M. truncatula*, *M. acuminata* and *T. aestivum* in roots compared to leaves and one in nodules compared to roots. Notably, in wheat Kwl2 and Kwl3 are enriched in roots, while Kwl1 seems to be more prevalent in leaves. These findings are in line with the previously recognized tissue-specific expression patterns of the maize Kiwellins (Altegoer et al., 2020). Altogether, our observations might suggest that the differential expression of Kiwellins in tissues is a general feature of land plants. Moreover, we also observed that Kiwellins can be upregulated during water limitation, which might suggest that also abiotic factors are able to induce a Kiwellin response (Riahi et al., 2019).

Discussion

Plants have developed numerous strategies to cope with abiotic and biotic stresses caused by e.g. drought or viral, bacterial, fungal, or herbivore pathogens (Draffehn et al., 2013; Quintana-Camargo et al., 2015; Mosquera et al., 2016). Kiwellin encoding genes have been shown to be regulated upon these challenges e.g. (Huang et al., 2017; Fiorilli et al., 2018; Lanver et al., 2018). Recent studies demonstrated that two Kiwellins from maize specifically target a secreted effector protein from *Ustilago maydis* (Han et al., 2019; Altegoer et al., 2020) making this protein family an prominent new candidate to better understand plant-microbe interactions.

Our study introduces a unified nomenclature and Kiwellin phylogeny to guide future research on Kiwellin proteins. Combining structural predictions and sequence-based analysis we can clearly distinguish between Kiwellins and BL proteins. Interestingly, Kiwellins appear to be a unique invention of land plants despite one singular hit found in the fungal kingdom. Further investigation will be required to understand the evolution of this putative Kiwellin. Our structural comparison suggests that Kiwellins are derivatives of BL proteins. Starting from the BL fold, Kiwellins may have evolved, for example, by extending the N-terminus, which may serve as a surface extension of the protein to perform specific functions.

Another feature present in many Kiwellin proteins is the kissper domain. Out of 915 Kiwellins identified in our study, 143 proteins harbor a kissper domain. It was shown in experiments that the *Actinida deliciosa* Kissper-Kiwellin can be cleaved into two domains kissper and kiwellin by actinidain, a cysteine protease highly abundant in kiwifruits *in vitro* (Tuppo et al., 2008). Structural comparisons have shown that the short kissper peptide has high similarities to cysteine-rich motifs such as the epidermal-growth-factor-like motif, or toxins from animals (Ciardiello et al., 2008). This region, which is only 40 amino acids long, contains 6 cysteines and can form 3 disulfide bridges. Remarkably, the kissper peptide exhibits pH-dependent and voltage-gated ion channel-forming activity in synthetic lipid bilayers (Ciardiello et al., 2008; Meleleo et al., 2012; Ciacci et al., 2014). The biological role of the kissper domain has not been elucidated. Suitable model systems to study the potential role of Kissper-Kiwellins for the interaction of plants with pathogenic or symbiotic microbes could be *Cucumis sativus* and *Glycine max*. Both are established model systems.

All three Kiwellin groups show significant responses as well as strong expression upon the interaction of plants with microbes. Therefore, we speculate that members of all Kiwellin classes may function as modulators of biotic interactions. Hence, manipulation of Kiwellins or Kiwellin expression might provide a novel means to develop new disease-resistant plants or plants with improved symbiotic capabilities e.g. for nitrogen-fixing bacteria. Suezawa et al. (2017) investigated the performance (a.o. photosynthetic rate, fruit quality, crop yield) of *A. chinensis* under poor drainage grafted on rootstocks of different *Actinidia* species. The best results were observed with *A. rufa* rootstocks, in which Kiwellins are highly abundant. Furthermore, Kisaki et al. (2019) showed increased tolerance to bacterial blossom blight in a hybrid breed of *A. chinensis* and *A. rufa*. This coincides with the difference in Kiwellins abundance between *A. rufa* (24 proteins) and *A. chinensis* (4 proteins).

Conclusion

Kiwellins have distinct structural characteristics that need to be addressed when annotating new proteins of this family. Otherwise especially BL proteins are likely to be missannotated as Kiwellins as shown in examples from literature and a protein database. In addition, we identified three evolutionary distinct subfamilies that can be distinguished a.o. based on the length of the Kiwellin loop at the β-hairpin. We hypothesize that Kiwellins are evolutionarily derived from BL proteins that belong to the pathogen-related family 4. They may have additional functions in plant immune response due to the N-terminal extensions. The provided nomenclature and grouping of Kiwellins along with evidence from transcriptomic data indicating Kiwellin proteins as mediators of plant-microbe interactions will aid to guide further research in the fields of plant-pathogen and -symbiont interactions.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/sra IDS: PRJEB4211, PRJNA10698, PRJNA116, PRJNA122, PRJNA14007, PRJNA17973, PRJNA182750, PRJNA190909, PRJNA20061, PRJNA20263, PRJNA207554, PRJNA225997, PRJNA225998, PRJNA232045, PRJNA232125, PRJNA238126, PRJNA240798, PRJNA241430, PRJNA243, PRJNA245122, PRJNA246165, PRJNA257217, PRJNA261643, PRJNA262552, PRJNA262907, PRJNA263939, PRJNA268357, PRJNA268358, PRJNA28131, PRJNA282644, PRJNA285087, PRJNA29019, PRJNA293435, PRJNA29797, PRJNA301363, PRJNA315994, PRJNA316327, PRJNA319578, PRJNA319678, PRJNA326436, PRJNA328963, PRJNA33471, PRJNA341501, PRJNA342685, PRJNA342701, PRJNA34677, PRJNA350852, PRJNA355166, PRJNA371634, PRJNA376605, PRJNA376608, PRJNA38691, PRJNA389730, PRJNA394209, PRJNA394242, PRJNA394253, PRJNA395588, PRJNA396054, PRJNA396063, PRJNA397875, PRJNA407962, PRJNA418295, PRJNA432228, PRJNA438537, PRJNA453230, PRJNA453787, PRJNA471752, PRJNA476953, PRJNA482138, PRJNA48389, PRJNA492326, PRJNA49677, PRJNA50439, PRJNA506972, PRJNA524157, PRJNA525136, PRJNA534520, PRJNA560384, PRJNA574457, PRJNA576248, PRJNA580467, PRJNA631757, PRJNA633601, PRJNA638679, PRJNA655717, PRJNA66163, PRJNA673911, PRJNA689611, PRJNA691360, PRJNA698663, PRJNA702515, PRJNA702529, PRJNA713846, PRJNA737421, PRJNA74771, PRJNA764258, PRJNA774802, PRJNA796348. The tools and data sets (tables, sequences, structure predictions) generated for this study can be found in the uni marburg gitlab repository (https://gitlab.uni-marburg.de/synmikro/ag-lechner/kiwellins).

Author contributions

GB conceived the study. ML supervised the project and drafted the manuscript. PK carried out the bioinformatic analyses and collected public RNA-seq data sets. PK, MC, and FA evaluated and verified structures. GB, ML, JF, and FA revised the manuscript. All authors wrote, read, and approved the final manuscript.

Funding

This research was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 460129525, NFDI4Mircobiota. GB thanks the European Research Council (ERC) for support through the project “KIWIsome” (Grant agreement number: 101019765).

Acknowledgments

We thank Clemens Thölken for fruitful discussions and Marlina Putri Indrayani for the graphic design of the plant pictograms. Computationally intensive calculations including structure prediction and evolutionary reconciliation were possible thanks to the HPC cluster MaRC3 located at the Philipps-University Marburg.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2022.1034708/full#supplementary-matHerial

SUPPLEMENTARY DATA SHEET 1

Additional figures and detailed algorithmic description (PDF). chapter 1, S1 Gene tree with and without truncated Kissper-Kiwellins. chapter 2, S2 Weblogo of consensus sequences. chapter 3, S3 Examples of kiwellin domain duplications and triplications. chapter 4 Detailed description of the kiwellin identification pipeline. S4 RMSDPMAS examples. S5 Range of descriptors used in find_kwl. S6 Kiwellin identification pipeline. chapter 5 Detailed RNA-seq results. S7 Workflow for re-evaluated RNA-seq data.

SUPPLEMENTARY DATA SHEET 2

Fully reconciled Kiwellin evolution (PDF).

SUPPLEMENTARY DATA SHEET 3

Listings, e.g. proteomes, RNA-seq sources, Kiwellin annotations (XLSX).

References

Ahrendt, S. R., Quandt, C. A., Ciobanu, D., Clum, A., Salamov, A., Andreopoulos, B., et al. (2018). Leveraging single-cell genomics to expand the fungal tree of life. *Nat. Microbiol.* 3, 1417–1428. doi: 10.1038/s41564-018-0261-0

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Altegoer, F., Weiland, P., Giammarinaro, P. I., Freibert, S.-A., Binnebesel, L., Han, X., et al. (2020). The two paralogous kiwellin proteins kwl1 and kwl1-b from maize are structurally related and have overlapping functions in plant defense. *J. Biol. Chem.* 295, 7816–7825. doi: 10.1074/jbc.RA119.012207

Bange, G., and Altegoer, F. (2019). Plants strike back: Kiwellin proteins as a modular toolbox for plant defense mechanisms. *Communicative Integr. Biol.* 12, 31–33. doi: 10.1080/19420889.2019.1586049

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2021). The interpro protein families and domains database: 20 years on. *Nucleic Acids Res.* 49, D344–D354. doi: 10.1093/nar/gkaa977

Ciacci, C., Russo, I., Bucci, C., Iovino, P., Pellegrini, L., Giangrieco, I., et al. (2014). The kiwi fruit peptide kissper displays anti-inflammatory and anti-oxidant effects in *in-vitro* and ex-vivo human intestinal models. *Clin. Exp. Immunol.* 175, 476–484. doi: 10.1111/cei.12229

Ciardiello, M. A., Giangrieco, I., Tuppo, L., Tamburrini, M., Buccheri, M., Palazzo, P., et al. (2009). Influence of the natural ripening stage, cold storage, and ethylene treatment on the protein and ige-binding profiles of green and gold kiwi fruit extracts. *J. Agric. Food Chem.* 57, 1565–1571. doi: 10.1021/jf802966n

Ciardiello, M. A., Meleleo, D., Saviano, G., Crescenzo, R., Carratore, V., Camardella, L., et al. (2008). Kissper, a kiwi fruit peptide with channel-like activity: Structural and functional features. *J. Pept. Science: Off. Publ. Eur. Pept. Soc.* 14, 742–754. doi: 10.1002/psc.992

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Cook, D. E., Mesarich, C. H., and Thomma, B. P. (2015). Understanding plant immunity as a surveillance system to detect invasion. *Annu. Rev. Phytopathol.* 53, 541–563. doi: 10.1146/annurev-phyto-080614-120114

Dabravolski, S. A., and Frenkel, Z. (2021). Diversity and evolution of pathogenesis-related proteins family 4 beyond plant kingdom. *Plant Gene* 26, 100279. doi: 10.1016/j.plgene.2021.100279

DeLano, W. L., and Bromberg, S. (2004). *Pymol user's guide* Vol. 629 (DeLano Scientific LLC).

Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., et al. (2011). Metabolic priming by a secreted fungal effector. *Nature* 478, 395–398. doi: 10.1038/nature10454

Draffehn, A. M., Li, L., Krezdorn, N., Ding, J., Lübeck, J., Strahwald, J., et al. (2013). Comparative transcript profiling by supersage identifies novel candidate genes for controlling potato quantitative resistance to late blight not compromised by late maturity. *Front. Plant Sci.* 4, 423. doi: 10.3389/fpls.2013.00423

Du, Y., Stegmann, M., and Villamil, J. C. M. (2016). The apoplast as battleground for plant–microbe interactions. *New Phytol.* 209, 34–38. doi: 10.1111/nph.13777

Eddy, S. R. (1998). Profile hidden markov models. *Bioinf. (Oxford England)* 14, 755–763. doi: 10.1093/bioinformatics/14.9.755

Edgar, R. C. (2004). Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Feldman, M., and Levy, A. A. (2012). Genome evolution due to allopolyploidization in wheat. *Genetics* 192, 763–774. doi: 10.1534/genetics.112.146316

Fine, A. J. (1981). Hypersensitivity reaction to kiwi fruit (chinese gooseberry, actinidia chinensis). *J. Allergy Clin. Immunol.* 68, 235–237. doi: 10.1016/0091-6749(81)90189-5

Fiorilli, V., Vannini, C., Ortolani, F., Garcia-Seco, D., Chiapello, M., Novero, M., et al. (2018). Omics approaches revealed how arbuscular mycorrhizal symbiosis enhances yield and resistance to leaf pathogen in wheat. *Sci. Rep.* 8, 1–18. doi: 10.1038/s41598-018-27622-8

Hamiaux, C., Maddumage, R., Middleditch, M. J., Prakash, R., Brummell, D. A., Baker, E. N., et al. (2014). Crystal structure of kiwellin, a major cell-wall protein from kiwifruit. *J. Struct. Biol.* 187, 276–281. doi: 10.1016/j.jsb.2014.07.005

Han, X., Altegoer, F., Steinchen, W., Binnebesel, L., Schuhmacher, J., Glatter, T., et al. (2019). A kiwellin disarms the metabolic activity of a secreted fungal virulence factor. *Nature* 565, 650–653. doi: 10.1038/s41586-018-0857-9

Huang, H., Nguyen Thi Thu, T., He, X., Gravot, A., Bernillon, S., Ballini, E., et al. (2017). Increase of fungal pathogenicity and role of plant glutamine in nitrogen-induced susceptibility (nis) to rice blast. *Front. Plant Sci.* 8, 265. doi: 10.3389/fpls.2017.00265

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Jaswal, R., Rajarammohan, S., Dubey, H., Kiran, K., Rawal, H., Sonah, H., et al. (2021). A kiwellin protein-like fold containing rust effector protein localizes to chloroplast and suppress cell death in plants. *bioRxiv*. doi: 10.1101/2021.08.20.456821

Jones, J. D., and Dangl, J. L. (2006). The plant immune system. *nature* 444, 323–329. doi: 10.1038/nature05286

Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* 8, 275–282. doi: 10.1093/bioinformatics/8.3.275

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., et al. (2021). "Highly accurate protein structure prediction with AlphaFold. *Nature* .596, 583–589. doi: 10.1038/s41586-021-03819-2

Kisaki, G., Shimagami, T., Matsudaira, K., Tsugi, Y., Moriguchi, K., Nakashima, K., et al. (2019). A kiwifruit cultivar crossbred with actinidia chinensis and actinidia rufa has practical tolerance to pseudomonas syringae pv. actinidiae biovar 3. *J. Plant Pathol.* 101, 1211–1214. doi: 10.1007/s42161-019-00349-9

Krueger, F., James, F., Ewels, P., Afyounian, E., and Schuster-Boeckler, B. (2021). *FelixKrueger/TrimGalore: v0.6.7.* doi: 10.5281/zenodo.5127899

Lanver, D., Müller, A. N., Happel, P., Schweizer, G., Haas, F. B., Franitza, M., et al. (2018). The biotrophic development of ustilago maydis studied by rna-seq analysis. *Plant Cell* 30, 300–323. doi: 10.1105/tpc.17.00764

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal w and clustal x version 2.0. *bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC Bioinf.* 12, 1–9. doi: 10.1186/1471-2105-12-124

Leitch, I. J., Chase, M. W., and Bennett, M. D. (1998). Phylogenetic analysis of dna c-values provides evidence for a small ancestral genome size in flowering plants. *Ann. Bot.* 82, 85–94. doi: 10.1006/anbo.1998.0783

Letunic, I., and Bork, P. (2021). Interactive tree of life (itol) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* 49, W293–W296. doi: 10.1093/nar/gkab301

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv*. doi: 10.48550/arxiv.1303.3997

Liu, W., Maccaferri, M., Chen, X., Laghetti, G., Pignone, D., Pumphrey, M., et al. (2017). Genome-wide association mapping reveals a rich genetic architecture of stripe rust resistance loci in emmer wheat (triticum turgidum ssp. dicoccum). *Theor. Appl. Genet.* 130, 2249–2270. doi: 10.1007/s00122-017-2957-6

Livingstone, C. D., and Barton, G. J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Bioinformatics* 9, 745–756. doi: 10.1093/bioinformatics/9.6.745

Love, M., Anders, S., and Huber, W. (2014). Differential analysis of count data–the deseq2 package. *Genome Biol.* 15, 10–1186. doi: 10.1186/s13059-014-0550-8

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Meleleo, D., Gallucci, E., Notarachille, G., Sblano, C., Schettino, A., and Micelli, S. (2012). Studies on the effect of salts on the channel activity of kissper, a kiwi fruit peptide. *Open Nutraceuticals J.* 5, 136–145. doi: 10.2174/1876396001205010136

Moghe, G. D., Hufnagel, D. E., Tang, H., Xiao, Y., Dworkin, I., Town, C. D., et al. (2014). Consequences of whole-genome triplication as revealed by comparative genomic analyses of the wild radish raphanus raphanistrum and three other brassicaceae species. *Plant Cell* 26, 1925–1937. doi: 10.1105/tpc.114.124297

Morel, B., Kozlov, A. M., Stamatakis, A., and Szöllősi, G. J. (2020). Generax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* 37, 2763–2774. doi: 10.1093/molbev/msaa141

Mosquera, T., Alvarez, M. F., Jiménez-Gómez, J. M., Muktar, M. S., Paulo, M. J., Steinemann, S., et al. (2016). Targeted and untargeted approaches unravel novel candidate genes and diagnostic snps for quantitative resistance of the potato (solanum tuberosum l.) to phytophthora infestans causing the late blight disease. *PloS One* 11, e0156254. doi: 10.1371/journal.pone.0156254

Nguyen, Q.-M., Iswanto, A. B. B., Son, G. H., and Kim, S. H. (2021). Recent advances in effector-triggered immunity in plants: new pieces in the puzzle create a different paradigm. *Int. J. Mol. Sci.* 22, 4709. doi: 10.3390/ijms22094709

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015). Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Offermann, L. R., Giangrieco, I., Perdue, M. L., Zuzzi, S., Santoro, M., Tamburrini, M., et al. (2015). Elusive structural, functional, and immunological features of act d 5, the green kiwifruit kiwellin. *J. Agric. Food Chem.* 63, 6567–6576. doi: 10.1021/acs.jafc.5b02159

OpenTree, Redelings, B., Sanchez Reyes, L. L., Cranston, K. A., Allman, J., Holder, M. T., McTavish, E. J., et al (2019a). Open Tree of Life Synthetic Tree (12.3). *Zenodo*. doi: 10.5281/zenodo.3937742

OpenTree, Cranston, K. A., Redelings, B., Sanchez Reyes, L. L., Allman, J., McTavish, E. J., Holder, M. T., et al (2019b). Open Tree of Life Taxonomy (3.2). *Zenodo*. doi: 10.5281/zenodo.3937751

Quintana-Camargo, M., Méndez-Morán, L., Ramirez-Romero, R., Gurrola-Díaz, C. M., Carapia-Ruiz, V., Ibarra-Laclette, E., et al. (2015). Identification of genes differentially expressed in husk tomato (physalis philadelphica) in response to whitefly (trialeurodes vaporariorum) infestation. *Acta physiologiae plantarum* 37, 1–19. doi: 10.1007/s11738-015-1777-z

Riahi, J., Amri, B., Chibani, F., Azri, W., Mejri, S., Bennani, L., et al. (2019). Comparative analyses of albumin/globulin grain proteome fraction in differentially salt-tolerant tunisian barley landraces reveals genotype-specific and defined abundant proteins. *Plant Biol.* 21, 652–661. doi: 10.1111/plb.12965

Rocafort, M., Fudal, I., and Mesarich, C. H. (2020). Apoplastic effector proteins of plant-associated fungi and oomycetes. *Curr. Opin. Plant Biol.* 56, 9–19. doi: 10.1016/j.pbi.2020.02.004

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., et al. (2021). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 49, D10. doi: 10.1093/nar/gkaa892

Sinha, M., Singh, R. P., Kushwaha, G. S., Iqbal, N., Singh, A., Kaushik, S., et al. (2014). Current overview of allergens of plant pathogenesis related protein families. *Sci. World J.* 2014, 543195. doi: 10.1155/2014/543195

Soltis, D. E., Soltis, P. S., Bennett, M. D., and Leitch, I. J. (2013). Evolution of genome size in the angiosperms. *Am. J. Bot.* 90, 1596–1603. doi: 10.3732/ajb.90.11.1596

Stergiopoulos, I., and de Wit, P. J. (2009). Fungal effector proteins. *Annu. Rev. Phytopathol.* 47, 233–263. doi: 10.1146/annurev.phyto.112408.132637

Suezawa, K., Fukuda, T., Mizutani, R., Yamashita, T., Otani, M., Abe, M., et al. (2017). Field performance of tetraploid Actinidia chinensis' Sanuki Gold' on Actinidia rufa rootstocks. *Acta Horticulturae*, Vol. 1218. 413–418. doi: 10.17660/ActaHortic.2018.1218.57

Tamburrini, M., Cerasuolo, I., Carratore, V., Stanziola, A. A., Zofra, S., Romano, L., et al. (2005). Kiwellin, a novel protein from kiwi fruit. purification, biochemical characterization and identification as an allergen. *Protein J.* 24, 423–429. doi: 10.1007/s10930-005-7638-7

Tuppo, L., Giangrieco, I., Palazzo, P., Bernardi, M. L., Scala, E., Carratore, V., et al. (2008). Kiwellin, a modular protein from green and gold kiwi fruits: evidence of *in vivo* and *in vitro* processing and ige binding. *J. Agric. Food Chem.* 56, 3812–3817. doi: 10.1021/jf703620m

Wang, J., Vanga, S. K., McCusker, C., and Raghavan, V. (2019). A comprehensive review on kiwifruit allergy: pathogenesis, diagnosis, management, and potential modification of allergens through processing. *Compr. Rev. Food Sci. Food Saf.* 18, 500–513. doi: 10.1111/1541-4337.12426

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J. (2009). Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Whelan, S., and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18, 691–699. doi: 10.1093/oxfordjournals.molbev.a003851

Yang, Z. (1995). A space-time process model for the evolution of dna sequences. *Genetics* 139, 993–1005. doi: 10.1093/genetics/139.2.993

Yates, A. D., Allen, J., Amode, R. M., Azov, A. G., Barba, M., Becerra, A., et al. (2022). Ensembl genomes 2022: an expanding genome resource for non-vertebrates. *Nucleic Acids Res.* 50, D996–D1003. doi: 10.1093/nar/gkab1007

frontiers

Supplementary Data

1 GENE TREE WITH AND WITHOUT TRUNCATED KISSPER-KIWELLINS

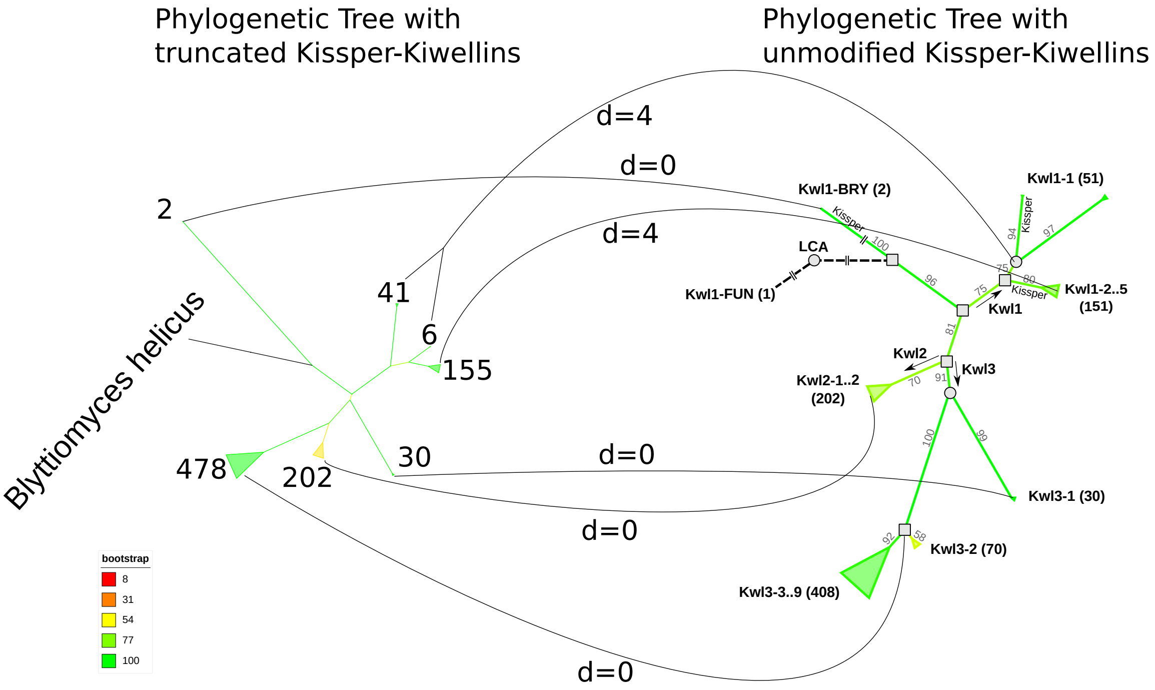


Figure S1: Left: phylogenetic tree with truncated Kissper-domains. The text on the collapsed branches indicates the number of proteins. Right: Kiwellin phylogenetic tree of Fig. 2. The text above the black lines indicates the number of differences between the connected sub-trees.

Supplementary Data

2 WEBLOGO OF CONSENSUS SEQUENCES

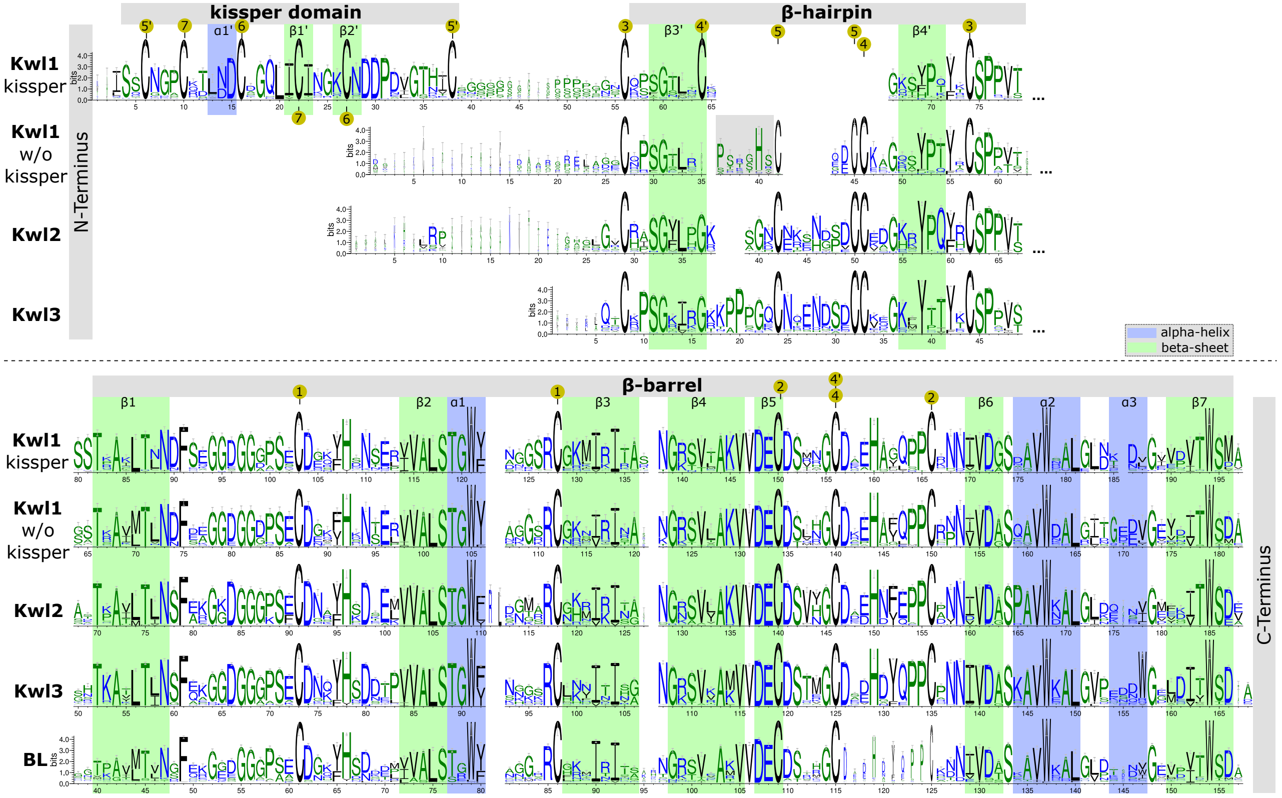
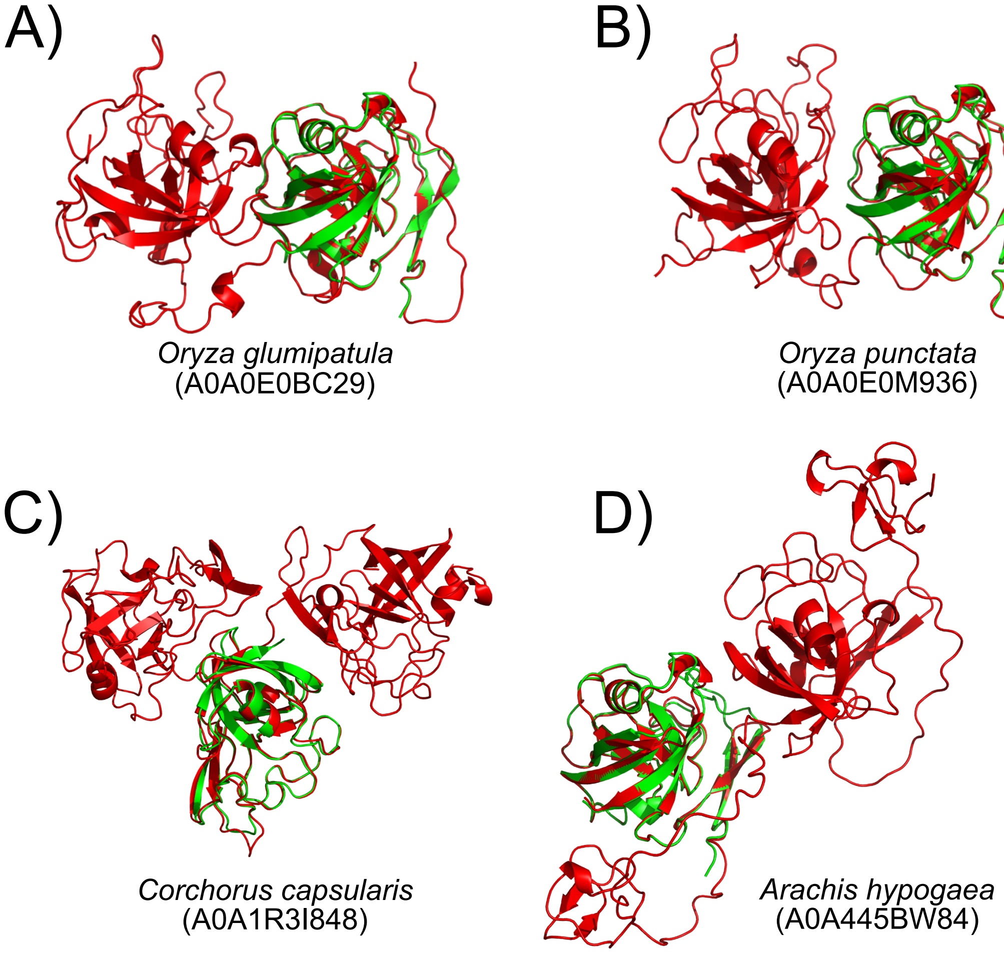


Figure S2: Weblogo of aligned consensus sequences (signal peptide trimmed) with secondary structure information of the Kiwellin groups Kissper-Kwl1, Kwl1 (Kiwellins without Kissper domain), Kwl2, and Kwl3 and a set of 391 BL proteins for reference. Green represents beta-sheets and blue alpha-helices. Numbered, yellow circles specify the cysteine residues forming disulfide bounds.

3 EXAMPLES OF KIWELLIN-DOMAIN-DUPLICATIONS AND -TRIPLICATIONS



Oryza glumipatula (A0A0E0BC29)

Oryza punctata (A0A0E0M936)

Corchorus capsularis (A0A1R3I848)

Arachis hypogaea (A0A445BW84)

Figure S3: Examples of proteins were found for Kiwellins using the algorithm with relaxed length parameters. A and B show Kiwellin-domain-duplications, C shows a triplication, and D shows a domain-duplication of Kissper-Kiwellins.

4 DETAILED DESCRIPTION OF THE KIWELLIN IDENTIFICATION PIPELINE

4.1 Re-identification ($\texttt{round}_1$)

Han et al. (2019) published 620 putative Kiwellin proteins in plants and fungi in a total of 61 species covering seven taxonomic groups, namely non-seed plants (NSP), gymnosperms (GYM), monocots (MON), stem eudicots (STE), asterids (AST), rosids (ROS) and fungi (FUN). These sequences were used as queries for the `UniProt` reference proteomes set release `Reference_Proteomes_2022_01` (Consortium, 2019). To cover all taxonomic groups, incomplete proteomes were added in accordance with the species covered in Han et al. (2019) (a total of 20.9k proteins were added).

All proteins of (Han et al., 2019) were initially mapped to the respective `UniProt` proteome using `Proteinortho` v6.1.1 (Lechner et al., 2011) with an E-Value threshold of 10^{-50}. Ultimately 411 of the initial 620 proteins could be re-identified in the current release of UniProt. Next SignalP v5.0b (Almagro Armenteros et al., 2019) was used to predict and trim a leading signal peptide. If no signal peptide was found the leading residues were removed one by one until either a signal peptide was found or less than 50% of the original protein is left. If no signal peptide was found the protein is left unmodified. Alphafold v2.0.0 (Jumper et al., 2020) was used to predict the 3D structure using the database bfd_database bfd_metaclust_clu_complete_id30_c90_final_seq, mgnify_database mgy_clusters_2018_12 and references pdb70, uniclust30_2018_08 and uniref90. Finally pyMOL v2.5.2 (DeLano and Bromberg, 2004) was used for manual inspection and visualization.

In the following, we refer to the manually verified set of the 235 Kiwellins, 117 BL, and 59 unrelated proteins as $\texttt{round}_1$ and e.g. with $\texttt{round}_1^{\text{kissper}}$ the set of Kissper-Kiwellins of round 1.

4.2 Advanced search ($\texttt{round}_2$)

In the following, we frequently investigated if a certain descriptor falls into a range defined by $\texttt{round}_1$. To allow more atypical values, we introduced a tolerance parameter of 25% to soften cutoffs, which is used throughout this pipeline.

Building on the knowledge from $\texttt{round}_1$ we scanned the UniProt database once more with a sophisticated pipeline named find_kwl. This tool can be subdivided into 3 main steps:

1. Pre-filtering and Pre-processing
2. Collect descriptors
3. hmmsearch and filtering

4.2.1 Pre-filtering and Pre-processing

To reduce the search space and save computation time we filtered proteins first by sequence length. The signal peptide trimmed Kiwellins identified in $\texttt{round}_1$ contain between 150 and 227 amino acids (Kiwellins with and without Kissper domain) and at least 3 cysteine residues (including the BL proteins). To account for a possible signal peptide the length limit is extended by 90. Therefore, only proteins of lengths $150 - 317 \pm 25\%$ and at least 3 cysteine residues were initially considered. Those sequences were then trimmed using SignalP as described for the sequences of $\texttt{round}_1$. The trimmed sequences were filtered again by length: $150 - 227 \pm 25\%$.

4.2.2 Collect descriptors

Besides the sequence length and the number of cysteine residues, we wanted to evaluate the 3D structure. Using AlphaFold 3D structure predictions were generated for all proteins passing the pre-filter. As AlphaFold predictions do not include secondary structure, the dssp routine of the R package bio3d v2.4-1.9 (Grant et al., 2006) was used to define the number of continuous region of β-sheets (b_regions). E.g. the Kiwellin 3-1b (A0A1D6GNR3):

primary: FPYRSLLQTCQPSGSIQGRSGNCNTENGSECCKNGRRYTTYGCSPPVTGSTRAVLTL
secondary: -------------BBBB--------AAA--------BBBB----------BBBBBBB
primary: NSFAEGGDGGAAACTGKFYDDSKKVVALSTGWYNGGSRCRKHIMIHAGNGNSVSAL
secondary: ------------------------BBBBBAAAA--------BBBBBB----BBBBB
primary: VVDECDSTVGCDKDHNFEPPCRNNIVDGSPAVWDALGLNKDDGQAQITWSDE
secondary: BBBBB-------------------BBBB-AAAAAAA---AAA-BBBBBBBB-
removed signal peptide precursor: MATVGGNRALYAVVALPLLATLLHGPMRLSHA
B:β-sheet, A:α-helix, -:unstructured, b_regions : 8

Furthermore, we rated the 3D structure with a special focus on the different domains (barrel, kissper, clamp) of the putative Kiwellins. Thus, a set of reference structures from different species of $\texttt{round}_1$ was hand curated, i.e. multiple reference structures were used to combat a possible underfitting:

- 4 kissper domains: Kwl1-2b (A0A2R6PEY1), Kwl0-1a (A0A2K1KL29), Kwl1-5b (A0A251NR03), Kwl1-2a (A0A067F280)
- 5 clamp domains Kwl3-8b (A0A2R6RCR6), Kwl1-3c (M1AEA5), Kwl2-2i (T1MCJ8), Kwl2-2a (A0A1Z5RFC7), Kwl1-1a (D8S9G1) and
- 4 barrel domains: Kwl3-4d (M0ZG50), Kwl3-4e (M0ZG49), Kwl3-1b (A0A1D6GNR3), Kwl3-6a (A0A0R0KPS2)
- *Zm*KWL1a (A0A1D6GNR3) crystal structure from Han et al. (2019)

The extracted structures for the kissper and clamp domain were reduced to the first 60 residues and the one for the barrel domain to the 6 β-sheets $\beta1, ...\beta6$.

The structure prediction was superimposed using PyMOL with the set of references to calculate the RMSD (the lower the better) and the number of matching atoms (MA, the higher the better). The RMSD can be arbitrarily small even for unrelated proteins with short overlaps (low MA) and a low RMSD does not necessarily follow from a high number of matching atoms. Therefore we combined both values and define

$$\texttt{RMSDPMAS} := \frac{\texttt{RMSD}}{\texttt{MA}^2}$$

The lower RMSDPMAS the better the superimposition as shown in Fig. S4. With that, the smallest RMSDPMAS was determined for each set of reference domains (later referred to as kissper, clamp, and barrel RMSDPMAS). To ensure comparability with the kissper and clamp domain only the leading 60 residues of a protein were compared to the reference structures. As shown in Fig. S5 this leads to a reliable identification of the Kiwellins with and without a kissper domain using the kissper RMSDPMAS and the clamp RMSDPMAS respectively. Additionally, the barrel and *Zm*KWL1a RMSDPMAS were used to exclude the unrelated proteins.

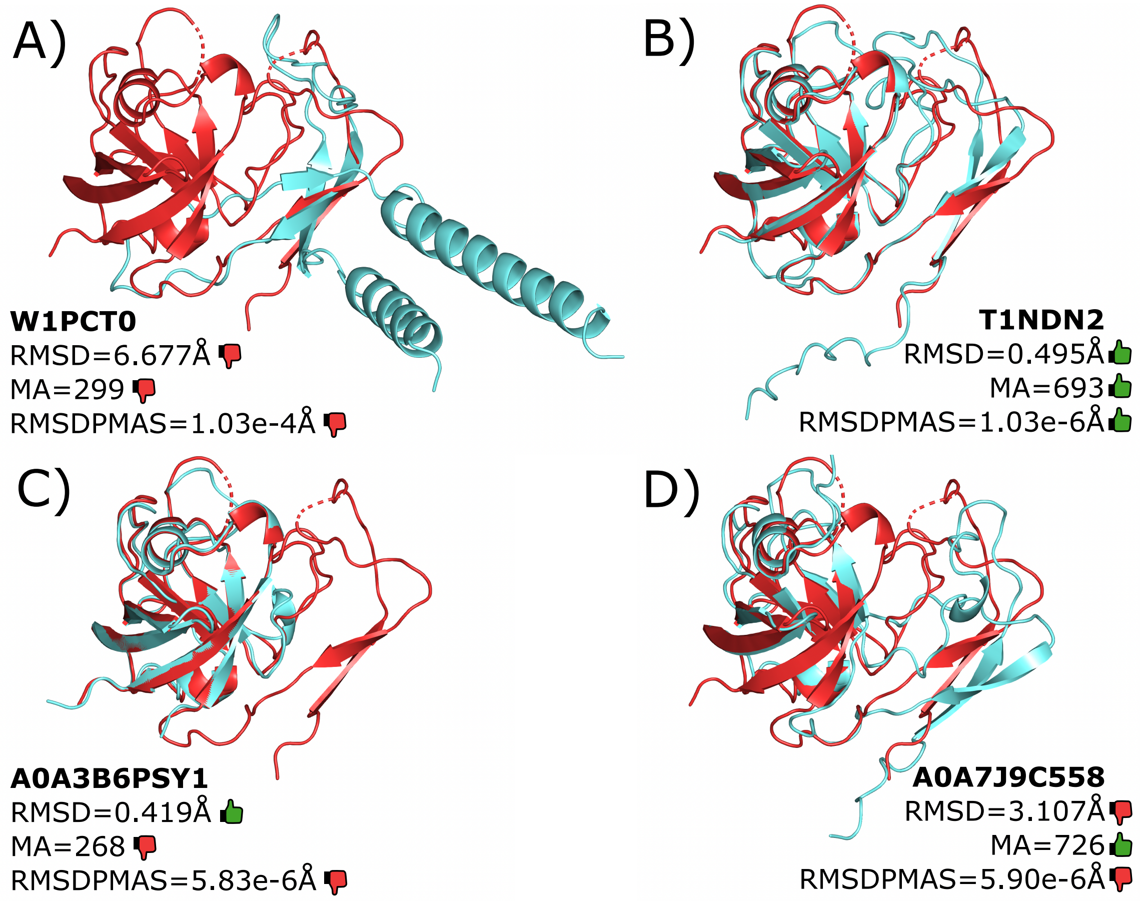


Figure S4: Superimposition of Kwl3-1b (A0A1D6GNR3, crystal structure of Han et al. (2019)) in red and blue the proteins A) W1PCT0, B) Kwl2-2c (T1NDN2), C) A0A3B6PSY1 and D) A0A7J9C558. A) shows an alignment with a high `RMSD` and low `MA` values resulting in a high `RMSDPMAS`. In contrast that the very similar structures of B) result in a low `RMSDPMAS`. In comparison the non-optimal alignments of C (short overlap with high similarity) and D (long overlap with low similarity) with similar `RMSD` or `MA` values, respectively both result in a higher `RMSDPMAS`.

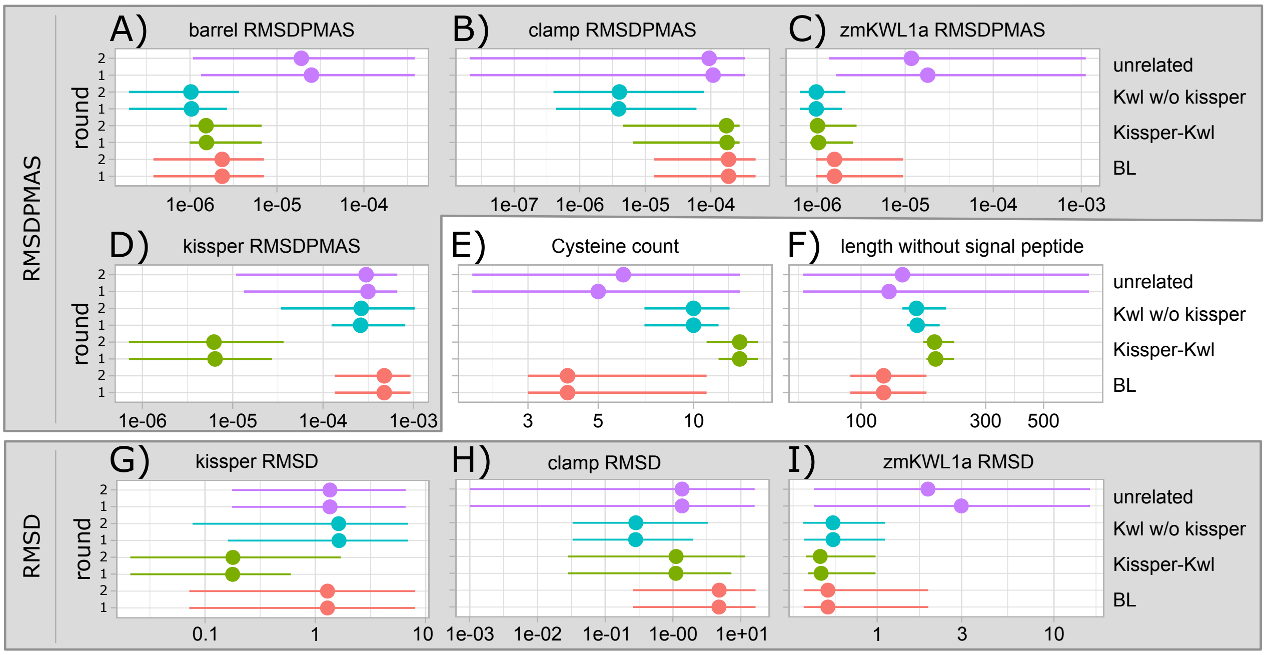


Figure S5: Min-Max range of the different descriptors used in `find_kwl` (A-F) as described in 4.2.2. Dots indicate the median value. G-I show some raw `RMSD` used to compute the `RMSDPMAS` of B-D. *Zm*KWL1a refers to the published crystal structure of A0A1D6GNR3 (Han et al., 2019).

Although the Kissper-Kiwellins contain a clamp domain the clamp `RMSDPMAS` is usually orders of magnitudes worse than the Kiwellins without the kissper domain since the first 60 residues do usually not contain the clamp.

To summarize, we collect the following descriptors for any protein that passes the pre-filter checks of 4.2.1:

- length of sequence after signalp trimming
- number of cysteine residues
- the number of continuous regions of β-sheets `b_regions`
- the structure scores `RMSDPMAS` against the set of reference structures

4.2.3 `hmmsearch` and filter

For each of the curated sets of $\texttt{round}_1$ (Kiwellins, Kissper-Kiwellins, and BLs) first an alignment was generated using a `muscle`. Next, those columns were removed that almost only consisted of gaps ($> 90\%$), and a hidden Markov model (HMM) was assembled using HMMer v3.2.1 (Eddy, 1998). `hmmsearch` was used to query the 3 models against all reference proteomes of UniProt. The resulting E-value of a match (m, p) between a model m and a protein p is denoted by $E_{(m,p)}$. To determine a suitable E-value cutoff for each model the 59 as unrelated identified proteins of $\texttt{round}_1$ were used as a negative control set:

$$c_m := \min_{up:\texttt{unrelated protein}} E_{(m,up)} = \begin{cases} 5.2 \cdot 10^{-42}, m = \texttt{BL} \\ 3.8 \cdot 10^{-54}, m = \texttt{kissper} \\ 3.2 \cdot 10^{-56}, m = \texttt{kiwellin} \end{cases}$$

Next, we wanted to assess if a descriptor (e.g. number of cysteine residues) is untypical compared to the values of $\texttt{round}_1$ with a predefined tolerance parameter of $t = 25\%$. For that we will say that a descriptor *is part of* $\texttt{round}_1$ if the value v lies in the range $[\text{mi}, \text{ma}]$ of values of $\texttt{round}_1$ extended by the tolerance parameter t:

$$\overbrace{\text{mi} - (\text{ma} - \text{mi}) \cdot t \leq v}^{\textit{at least}} \underbrace{\leq \text{ma} + (\text{ma} - \text{mi}) \cdot t}_{\textit{below}}$$

On the same note we defined that a descriptor is *at least* or *below* $\texttt{round}_1$ for only the left or right inequality respectively.

Furthermore, we defined a set of filtered matches as the subset of all reported matches (`hmmsearch`) that fulfill the following set of rules:

- the E-value below the defined model specific cutoff: $E_{(m,p)} \leq c_m$
- the number of cysteine residues *at least* $\texttt{round}_1^m$
- the sequence length is *part of* $\texttt{round}_1^m$
- for $m =$ Kiwellin (without kissper domain):
  - the clamp `RMSDPMAS` is smaller than the kissper `RMSDPMAS` up to the tolerance $t = 25\%$
  - the clamp, barrel and *Zm*KWL1a `RMSDPMAS` is *below* $\texttt{round}_1^m$
  - if the protein length is below the midpoint of the lengths of $\texttt{round}_1^m$ then $\texttt{b\_regions} \geq 2$
- for $m =$ kissper (Kiwellin with kissper domain):
  - the kissper `RMSDPMAS` is smaller than the clamp `RMSDPMAS` up to the tolerance t
  - the kissper, barrel and *Zm*KWL1a `RMSDPMAS` is *below* $\texttt{round}_1^m$
- for $m =$ BL (Barwin-like):
  - the barrel and *Zm*KWL1a `RMSDPMAS` is *below* $\texttt{round}_1^m$

For each protein among the set of filtered hits, we report the model with minimal E-value as the best match for that protein.

In total 683 new Kiwellins were found, i.e. 59 with and 589 without a kissper domain. The steps $2 - 3$ of the pipeline were repeated with the new extended set of Kiwellins as the input (alignment, HMM model, descriptor ranges), and 15 further Kiwellins were identified. In a final step, this set of Kiwellins was checked by hand again and almost all entries (98.2%) could be verified as correctly classified. We removed only 17 entries. Most of which were faulty Kiwellins (missing $\beta 7$). A final set of 915 Kiwellins (62 with and 772 without a kissper domain) were reported and are denoted as $\texttt{round}_2$ in Fig. S5.

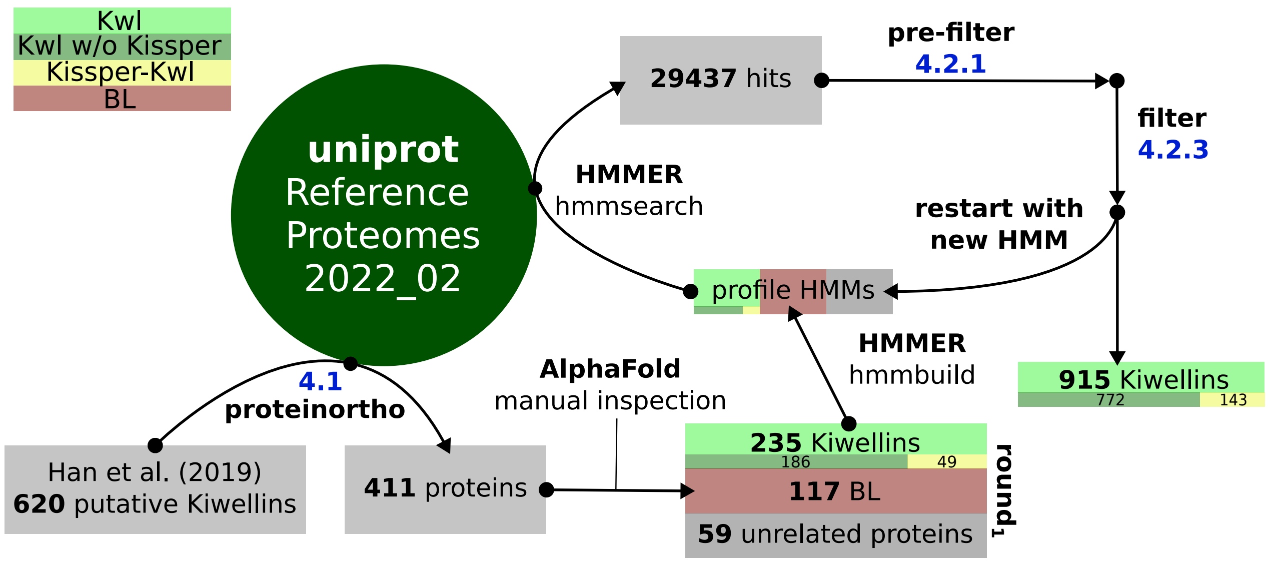


Figure S6: Flow chart of the Kiwellin identification pipeline. More details on the blue-marked processing steps are described in the respective chapters.

5 DETAILED RNA-SEQ RESULTS

To get an idea of the functions of Kiwellins we re-analyzed publicly available RNA-seq data sets from the NCBI SRA. 70 experiments were obtained using the following filter characteristics: RNAseq, RNA, stress keywords (pathogenic, symbiotic, water, ...), and the scientific name of the plants. The 70 data sets were checked on quality parameters of the raw data using fastQC (Andrews et al., 2010) and on data integrity (at least 2 replicates, associated publication, unambiguous sample naming, and experimental descriptions) resulting in 31 data sets. To determine if a Kiwellin was significantly regulated an FDR threshold of 5% was used. We consider an entry to be differentially expressed if the absolute log2 transformed fold changes (L2FC) is above 1 and the P-value is below the above FDR threshold. Furthermore, we define a Kiwellin group as strongly expressed if the baseMean (baseM) is at least 80 (a proxy for the overall expression strength; $\log_{10}(80) \approx 1.9$). Finally, we grouped the experiments by experimental parameters (pathogenic, symbiotic, abiotic, tissue-specific responses).

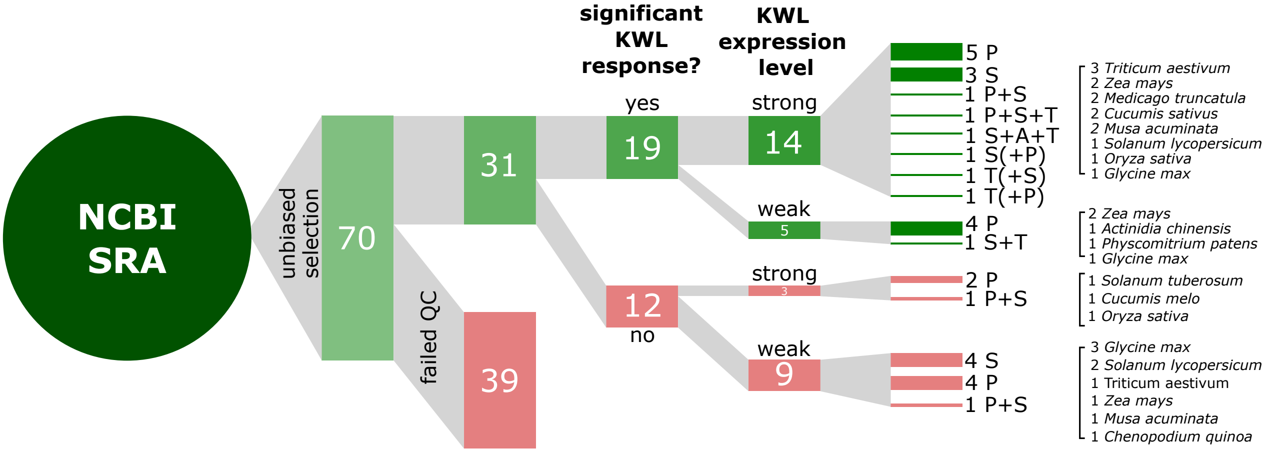


Figure S7: Sankey diagramm illustrates from left to right the workflow for re-evaluated RNA-seq data. P: pathogenic, S: symbiotic, A: abiotic, T: tissue specific responses. QC: Quality check. More information can be found in the chapter 5. The number of analysed experiments per species are given on the right side. Brackets indicate non signifiant interactions not significant (e.g. S(+P): significant response only to the symbiotic and not to pathogenic partner).

In the following chapters, we shortly describe the 31 analyzed case studies and examine the regulation of the Kiwellins groups. Groups were formed from indistinguishable Kiwellins concerning the associated transcript(s) as described in the Material and Methods. A Kiwellin group can include multiple proteins as well as transcripts. For example lets consider the kiwellin group 'Kwl2-2t,2v,2s' of *T. aestivum* shown in PRJNA743515. This group includes the three almost identical Kiwellin proteins Kwl2-2t, Kwl2-2v, Kwl2-2, that share 188/197 identical residues. With the help of proteinortho two similar transcripts (XM_044541154.1, XM_044500936.1) were identified from the respective transcriptome. Since the proteins as well as the transcripts are almost indistinguishable, we combine the results of this group into one entry. All identified groups of transcripts are listed in the table "Nomenclature-KWL" and the used transcriptomes in "transcriptome sources" of Supplementary Data 3 (proteinortho transcripts).

For each experiment, a heatmap is shown to visualize the L2FC on the left panel from blue (down-regulated) to red (up-regulated). Gray is shown if the comparison does not exhibit significant changes. The middle panel specifies the log10 transformed baseM. A * symbol in the name indicates that the Kiwellin group surpasses the baseMean threshold of 80 and thus is considered to be strongly expressed. More details can be found in the Material and Methods. The right panel shows average normalized counts as well as standard deviations between the replicates of all conditions.

The results of the 31 analyzed experiments were divided into sections according to Fig. S7:

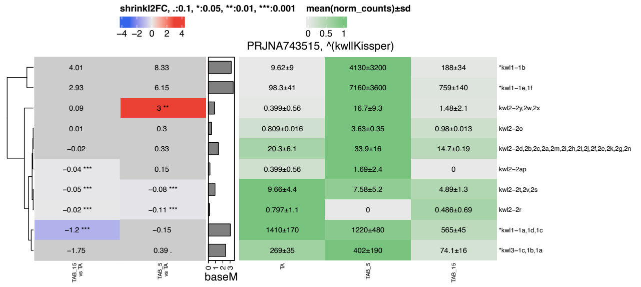
5.1 Significant response and strong expression

5.1.1 Pathogenic response: P

Triticum aestivum (PRJNA743515)

Bipolaris sorokiniana is a hemibiotrophic fungus responsible for several plant diseases. The study Zhang et al. (2022) aimed to investigate how genes are regulated when *Triticum aestivum* is infected by pathogenic fungus (TAB). Uninfected plants (TA) served as control groups. Plants were soil-inoculated and samples of root and basal stems were harvested 5 and 15 days after infection. RNA was isolated from the samples and sequenced.

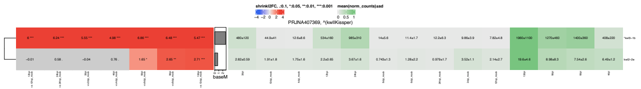
We found 4 strongly expressed groups (*-prefix), three from Kwl1 and one from Kwl3, and 6 further weak expressed groups of Kwl2. We found that one group of the weakly expressed Kwl2 to be differentially regulated 5 dpi (3 L2FC). One group of Kwl1 showed a slight down-regulation late in the infection stage (-1.2 L2FC) and remained unchanged at 5 dpi.



Zea mays (PRJNA407369)

Zea mays was syringe-infected with *Ustilago maydis*. In Lanver et al. (2018) infected plant material was harvested at different time points ($\frac{1}{2}$, 1, 2, 4, 6, 8, 12 dpi) and mRNA was analyzed. Axenic *U. maydis* culture and water-inoculated plants (mock) served as controls or comparison groups.

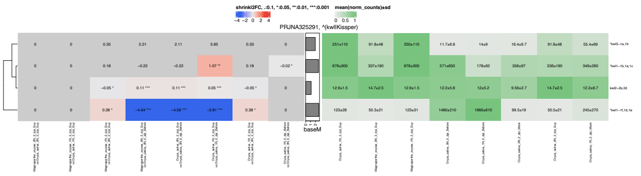
We found 2 Kiwellins of Kwl3 and Kwl2, of which Kwl3-1b is strongly expressed (*-prefix). Furthermore, Kwl3-1b showed a strong up-regulation ($\approx 5-7$ L2FC) among all time points compared to the mock-inoculated control. The weakly expressed Kwl2 protein showed a late response with a slight up-regulation ($\approx 1-3$ L2FC) starting at 6 dpi.



Oryza sativa (PRJNA325291)

In Huang et al. (2017), rice was inoculated with the fungus *Magnaporthe oryzae* (with=Guy, without=Before). It is known that nitrogen fertilization increases the effects of many diseases. The authors studies whether the external addition or omission of nitrogen led to differentially expressed genes during infection in both species to explain Nitrogen-Induced Susceptibility (NIS). For this purpose, rice plants were infected with water or the fungus and 0 dpi or 2 dpi shoot tissue of the plants were harvested. Nitrogen was omitted from the fertilizer in one series of experiments (0N) and added in the form of ammonium nitrate in another (1N). Subsequently, mRNA was isolated from the obtained tissue and analyzed.

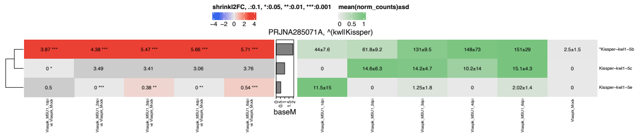
We found 2 groups of two Kwl1 and one Kwl3 to be highly expressed. The kwl3 group shows no regulation in response to the infection. One Kwl1 group was down-regulated (4 L2FC), and in the second group, a slight up-regulation upon infection under nitrate treatment was found. No ammonium nitrate-specific response was observed among all groups.

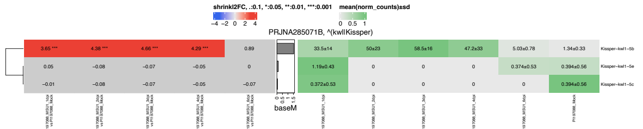


Cucumis sativus (PRJNA285071)

In Burkhardt and Day (2016), a resistant strain (PI197088) and an susceptible strain (Vlaspik) of *Cucumis sativus* were infected with the fungus *Pseudoperonospora cubensis* and water (mock), respectively. Leaves of the plant were harvested 1, 2, 3, 4, and 6 dpi and mRNA levels were determined in each case.

We found a group of Kissper-Kiwellins (Kwl1) to be strongly expressed in the susceptible strain and moderately in the resistant strain. Furthermore, we found a strong up-regulation upon infection ($\approx 4-6$ L2FC) in the susceptible strain throughout infection, while the resistant strain showed a slight dampening of the differential response ($\approx 3-4$ L2FC). Furthermore, the response vanishes at 6 hpi for the resistant strain.

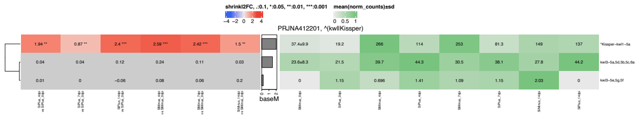




Glycine max (PRJNA412201)

It is known that silicon can protect plants from biotrophic and hemibiotrophic pathogens. To better understand this mechanism, Glycine max was infected with *Phytophthora sojae* in Rasoolizadeh et al. (2018). Silicon was added in one case (SiPlus) and omitted (SiMinus) in the other plants. After 21 days of infection, root samples were collected and mRNA was isolated and sequenced.

We found one group of Kissper-Kiwellins (Kwl1) to be highly expressed with a differential response to the infection ($\approx 1-3$ L2FC). The silicon treatment slightly reduced effect (≈ 1 L2FC less).

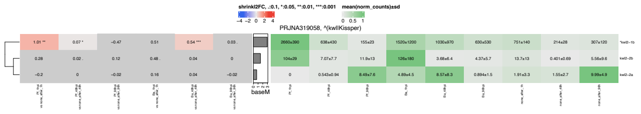


5.1.2 Symbiotic response: S

Musa acuminata (PRJNA319058)

In Gamez et al. (2019) seedlings of Musa acuminata were inoculated with two species of growth-promoting rhizobacteria: *Bacillus amyloliquefaciens* (Ba) and *Pseudomonas fluorescens* (Pf). 1 hpi, 2 dpi and 4 dpi whole seedlings were collected, and the mRNAs were isolated. These data sets were compared with water-inoculated seedlings.

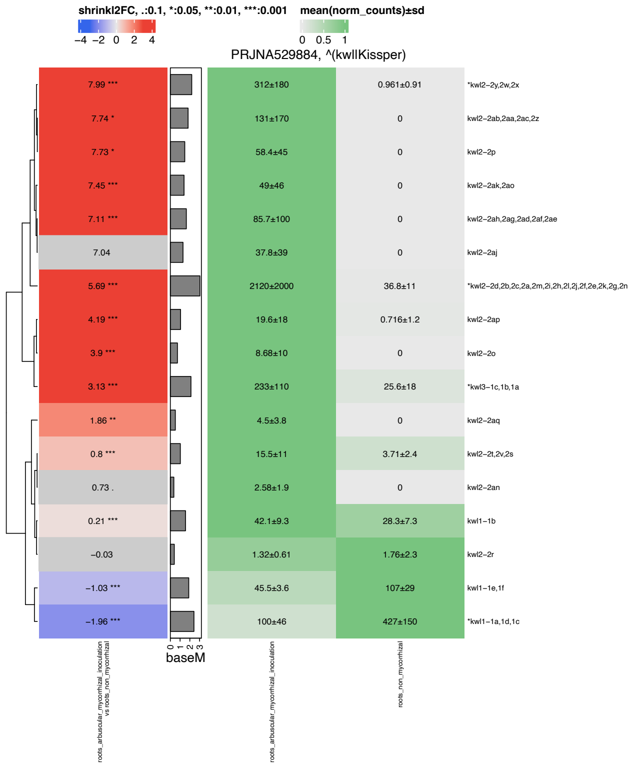
We detected three Kiwellins groups of Kwl2. Kwl2-1b was the only strong expressed group and showed a weak up-regulation upon infection with Pf after 1 hpi and remains inconspicuous otherwise. The other two Kiwellins showed no differential response to the infection.



Triticum aestivum (PRJNA529884)

In Li et al. (2018a) *Triticum aestivum* was infected with the arbuscular mycorrhizal fungus *Rhizophagus irregularis*. After 42 days of infection, shoot tissues of the plants were harvested and mRNA was extracted and analyzed. This data set was compared with non-infected plants.

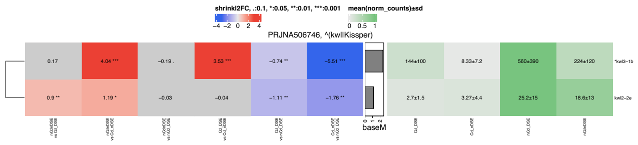
Our analysis revealed 4 groups (Kwl1, Kwl3, and 2 Kwl2 variants) of highly expressed Kiwellins. Members of Kwl2 and Kwl3 showed a strong up-regulation upon infection ($\approx 3-8$ L2FC) and Kwl1 a down-regulation (≈ 2 L2FC).



Zea mays (PRJNA506746)

In Shen et al. (2020), the cadmium tolerance of *Zea mays* roots was investigated, which previously treated with the endophyte *Exophiala pisciphila*. Roots of three-day-old maize seedlings were first inoculated with the fungus (with=DSE, without=nDSE). 10 days later, plants were fertilized with cadmium for 31 days (with=Cd, without=nCd). Plants not treated with cadmium and/or the fungus served as the control. Finally, roots were harvested and mRNA was extracted and analyzed.

We found one group of Kwl3 to be strongly expressed. In the case where the fungi were absent, this group show a strong down-regulation with cadmium (4 L2FC) but no change upon fungal treatment was detectable. If cadmium is absent, the infection does not significantly impact expression but if cadmium is introduced into the system we see an up-regulation (3.5 L2FC) in infected plants.

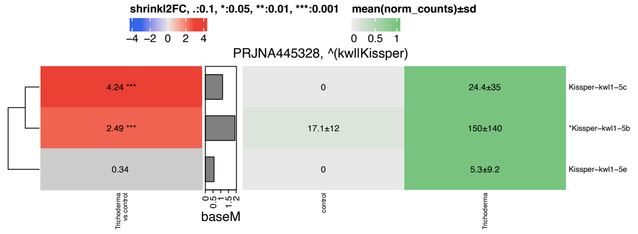


5.1.3 Pathogenic but no symbiotic response: P+S

Cucumis sativus (PRJNA445328)

To better understand *Trichoderma*-induced plant resistance to many plant pathogens, cucumber plants were infected with *Botrytis cinerea* in the presence or absence of *Trichoderma* in Yuan et al. (2019). At the three-leaf stage, plants were inoculated with *Trichoderma* and 24 hours later *B. cinerea* was injected into the leaves. Samples of the leaves were harvested 96 hours later and examined for differential gene expression.

We detected 3 groups of kwl1 Kissper-Kiwellins. One group was highly expressed and showed a significant up-regulation in response to the symbiont and pathogen (≈ 2.5 L2FC).

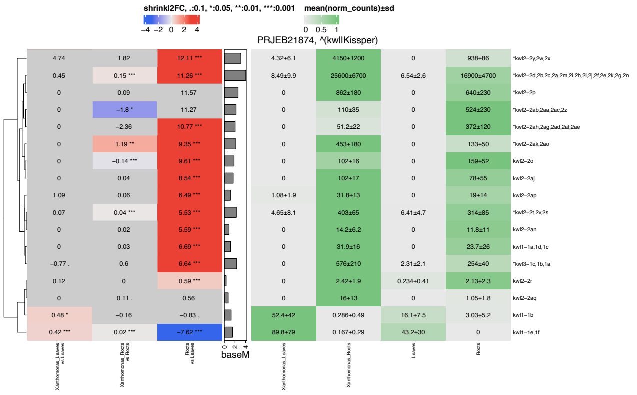


5.1.4 Pathogenic, symbiotic and tissue specific effect: P+S+T

Triticum aestivum (PRJEB21874)

Triticum aestivum was infected with the bacterial pathogenic *Xanthomonas translucens* in Fiorilli et al. (2018). It was tested whether the mycorhizal fungus *Funneliformis mossae* influenced the infection. After plants were colonized by mycorrhiza for 49 days, plants were inoculated with the pathogenic bacterium. One day after infection, samples of roots and leaves were isolated and the mRNA levels of the three species were examined.

We found 8 highly expressed Kiwellin groups most of which belong to Kwl2 but and to Kwl3. The Kwl3 group showed no significant response and the results for Kwl2 were mixed. In roots, we observed one group of Kwl2 to be down-regulated (≈ 2 L2FC) and one group to be up-regulated (1 L2FC) in response to the pathogen and symbiont. Differences between roots and leaves can be observed for Kwl1, Kwl2, and Kwl3. Overall the expression strength in roots was observed to be higher compared to leaves.

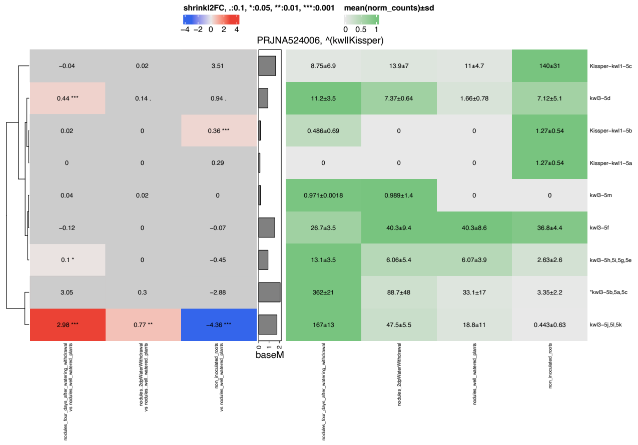


5.1.5 Symbiotic and abiotic response and tissue-specific effects: S+A+T

Medicago truncatula (PRJNA524006)

In Sańko-Sawczenko et al. (2019), the Fabaceae *Medicago truncatula* was evaluated for their response to water stress when the roots were inoculated by nitrogen-fixing bacteria *Sinorhizobium meliloti*. After the roots were successfully colonized by the bacteria, the plants were subjected to water stress. For this, the plants were not watered for up to 4 days after colonization. At the end of the four days, root nodules were harvested from watered and non-watered plants. Uninfected plants served as control, here the roots were harvested. The mRNA was isolated from the collected samples and analyzed.

We found one highly expressed group of Kwl3 that shows a strong up-regulation after 4 days of water withdrawal (2 L2FC).

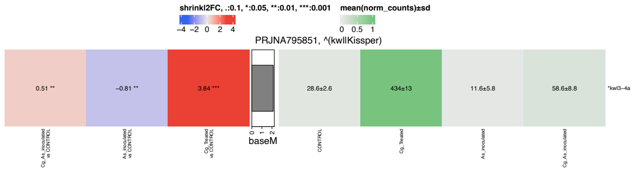


5.1.6 Symbiotic but no pathogenic response: S(+P)

Solanum lycopersicum (PRJNA795851)

Biological control agents (BCA) play a major role to combat plant pathogens. Singh et al. (2021) aimed to investigate the transcriptonal response to treat with the BCA fungus *Chaetomium globosum* (Cg) on plants infected with the pathogenic fungi *Alternaria solani* (As). First, 21-day-old tomato plants were inoculated with the BCA. Another 24 hours later, the plants were spray-inoculated with the pathogen. After five days of infection, infected leaves were harvested, and RNA was isolated and sequenced. In total, four data sets resulted from this experiment: plants not infected (CONTROL), plants infected with both fungi (Cg_As_inoculated), and plants infected with only one fungus each (Cg_inoculated, As_inoculated).

We found one group of Kwl3 to be highly expressed. An up-regulation could be observed in case of infection with the BCA fungus *C. globosum* (≈ 4 L2FC). Furthermore, a slight down-regulation (below 1 L2FC) was observed upon infection with the pathogen *A. solani*. In response to combinatorical treatment with the pathogen, a slight up-regulation (below 1 L2FC) was observed.

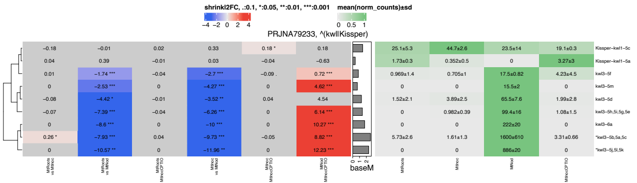


5.1.7 Tissue specific but no symbiotic response: T(+S)

Medicago truncatula (PRJNA79233)

In Boscari et al. (2013), the transcription of developing nodules on *Medicago truncatula* was investigated. For this purpose, plants were infected with their symbiont *Sinorhizobium meliloti* and samples were taken from different stages of the nodules/roots and the mRNA was isolated and sequenced. Roots of the plant that were not infected were collected 4 days after infection (developing nodules), and 12 days after infection (matured nodules) were examined. Furthermore, a nitric oxide scavenger (cPTIO) was added to an infected plant and the effect on nodules was studied. Thus, a total of four data sets were obtained and compared: unified roots (MtRoots), infected roots (MtInoc), nodules (MtNod), and infected roots treated with cPTIO (MtInocCPTIO). All samples except the nodule sample were collected and analyzed 4 days after infection or mock infection.

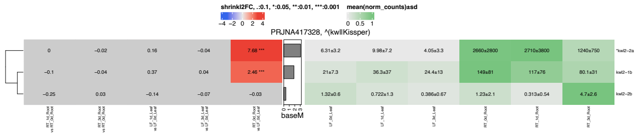
We found 2 groups of Kwl3 to be strongly expressed but no difference between infected and non-infected roots (cPTIO independent) was observed. Remarkably, Kiwellins were almost exclusively found in nodules.



5.1.8 Tissue specific but no pathogenic response: T(+P)

Musa acuminata (PRJNA417328)

Benzothiadiazole (BTH) is an inducer of plant resitance that stimulates the defense response in bananas and protects against pathogen infection. In Cheng et al. (2018), via RNA-seq, the effect of BTH was investigated at the gene expression level by spraying young plants with a BTH solution. For this purpose, plant samples from roots (RT) and leaves (LF) 1 and 3 days post-infection with the fungal pathogen *Fusarium oxysporum* were compared with their respective controls (0 dpi). We found one group of highly expressed Kwl2 members (in roots but almost absent in leaves). Furthermore, no significant changes were observed in response to BTH. Kiwellins were almost exclusively found in roots (compared to leaves).



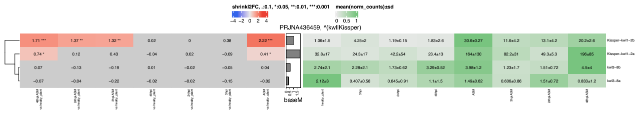
5.2 Significant response and weak expression

5.2.1 Pathogenic response: P

Actinidia chinensis (PRJNA436459)

Michelotti et al. (2018) investigated the effect of acibenzolar-A-methyl (ASM, a bactericidal component) on the course of infection of *Pseudomonas syringae pv. actinidiae* on its host the kiwifruit plant (*Actinidia chinensis*). For this purpose, plants were treated with or without ASM and inoculated with the bacterium or buffer. 3, 24, and 48 hpi samples were obtained, and the mRNA was isolated from ground tissue and analyzed.

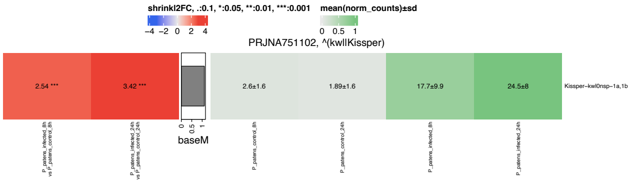
We found 4 groups of Kiwellins of which 2 are Kissper-Kiwellins (Kwl1) and two belong to Kwl3. The groups of kwl1 are moderately expressed in one group and we see an ASM-specific up-regulation of $\approx 1 - 2$ L2FC. In all comparisons without ASM, no effects are were observed throughout the infection.



Physcomitrium patens (PRJNA751102)

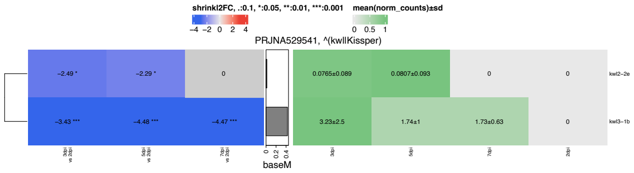
Otero-Blanca et al. (2021) investigated the defense mechanisms of *Physcomitrium patens* against *Colletotrichum gloeosporioides*. For this purpose, plants were spray-inoculated with the pathogen. Samples were harvested and analyzed 8 and 24 hours after infection and uninfected plants served as controls.

We found a group of two Kissper-Kiwellins from Kwl1 (here called kissper-kwl0nsp). Although this group was not strongly expressed we found a response to the infection at both time points ($2 - 3$ L2FC).



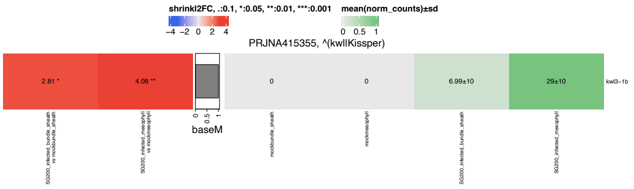
Zea mays (PRJNA529541)

Garcia-Ceron et al. (2021) infected *Zea mays* with *Fusarium graminearum* and examined the change in gene expression. For this purpose, the leaves of the plants were injured and disk-infected with the fungus. Leaf tissue was collected after 3, 5, and, 7 days respectively, and the mRNA was examined. The Comparison was made with uninfected plants and fungi grown in axenic culture. No Kiwellin was found to be highly expressed but a down-regulation for a group of Kwl2 and a group of Kwl3 was observed upon infection ($\approx 2 - 4$ L2FC).



Zea mays (PRJNA415355)

In Li et al. (2018b), mRNA levels were examined organ-specifically during tumor development of *Ustilago maydis* on *Zea mays*. For this purpose, data sets of bundle sheaths and the mesophyll of fungus-infected plants were compared with water-inoculated plants (mock). We found no strongly expressed Kiwellin groups but an up-regulation ($\approx 2 - 4$ L2FC) of Kwl3 in infected bundle sheath and mesophyll tissue.

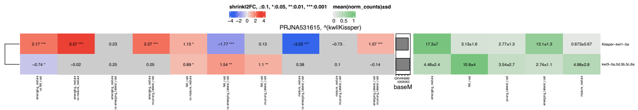


5.2.2 Pathogenic and tissue specific response: S+T

Glycine max (PRJNA531615)

Adhikari et al. (2019) investigated the influence of nodulation on the roots of the soybean plant. For this purpose, *Glycine max* was infected with the bacterium *Bradyrhizobium diazoefficiens*. The infected root tissue on which nodules were formed was harvested at $5 - 7$ days after infection (emerging nodules) or $14 - 16$ days after infection (mature nodules). Root tissue was collected above and/or below the nodules as control groups.

We found no strongly expressed Kiwellin group but differences were found e.g. between emerging and mature nodules as well as between mature nodules and uninfected roots (NA_root).



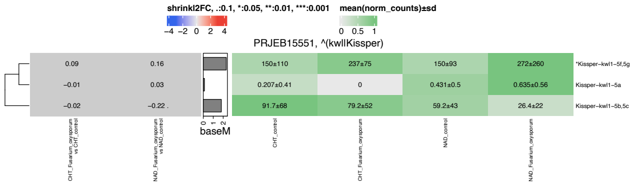
5.3 Strong expression but no significant response

5.3.1 Pathogenic response: P

Cucumis melo (PRJEB15551)

Two strains of *Cucumis melo* were infected with *Fusarium oxysporum f. sp. melonis* Snyd. & Hans race 1.2 (FOM1.2) in Silvia Sebastiani et al. (2017). One of the melon lines is the NAD strain, which is capable of early recognition of pathogens and developing resistances. The second melon genotype Charentais (CHT) is susceptible to the fungus. Plantlets of the two strains were infected with the fungus and 1 and 2 days. Stems of the small plants were harvested and mRNA levels were determined and compared.

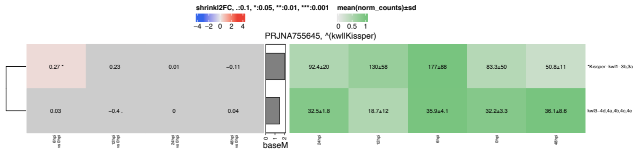
In our reanalysis, we found one highly expressed group of Kwl1 (Kissper-Kiwellin). None of the groups are significantly differentially expressed.



Solanum tuberosum (PRJNA755645)

Alternaria solani is a necrotrophic fungus that infects potatoes and other crops. The aim of Brouwer et al. (2021) was to investigate how the transcriptome of *Solanum tuberosum* changes during infection with this fungus. For this purpose, the leaves of six-week-old potato plants were infected with the pathogen. Samples of infected leaves were collected 1, 6, 12, 24, and 48 hours after infection, and the mRNA was analyzed and sequenced. Uninfected plants served as the control group (0 hpi).

Our reanalysis was able to identify Kiwellins 3 groups of Kissper-Kwl1. One group can be considered as strongly expressed but no response to the infection was observed.

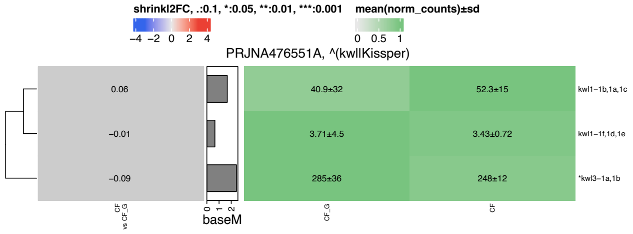


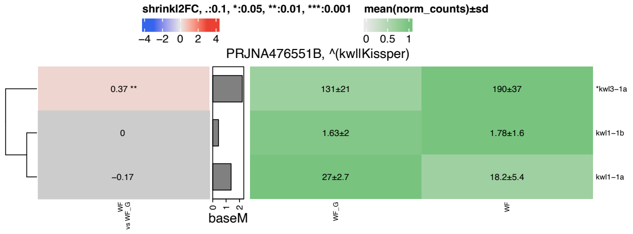
5.3.2 Pathogenic and symbiotic response: P+S

Oryza rufipogon and Oryza sativa (PRJNA476551)

Tian et al. (2019) investigated the differences between wild rice (*Oryza rufipogon*) and cultivated rice (*Oryza sativa*) inoculated with the arbuscular mycorrhizal fungus *Rhizoglomus intraradices* upon infection with the pathogen *Magnaporthe oryzae*. For this purpose, ten-day-old rice plants were first inoculated with the mycorrhizal fungus, and after another 45 days, the leaves of the plants were spray-inoculated with the pathogen. After another seven days, the roots of the plants were harvested and the RNA was isolated and analyzed.

For both *Oryza* species we found a group of Kwl3 to be highly expressed but no significant differential changes were observed.





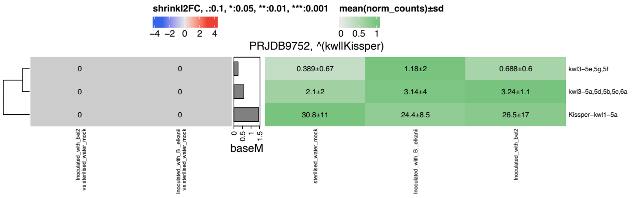
5.4 Weak expression and no significant response

5.4.1 Symbiotic response: S

Glycine max (PRJDB9752)

Roots of *Glycine max* strain EN1282 (nfr1-mutant - a strain lacking in a Nod factor receptor) was infected with the symbiotic bacterium *Bradyrhizobium elkanii* USDA61 in Ratu et al. (2021). The wild-type strain of the bacterium was compared with a T3SS (Type 3 Secretion system) deletion strain. The roots of the seedlings were harvested 30 days after infection and mRNA levels of the bacterium and the plant were measured.

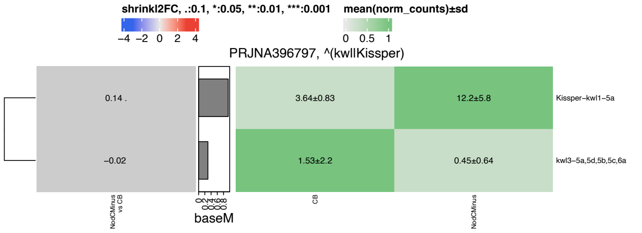
We found 3 groups of Kiwellins of which 2 belong to Kwl3 and one Kissper-Kiwellins to Kwl1. The Kwl3 groups were weakly expressed and Kissper-Kiwellin showed a moderate expression. Furthermore, all groups in this experiment showed no significant response to the infection.



Glycine max (PRJNA396797)

To investigate transcriptional changes associated with nodule formation genes in soybean, *Glycine max* roots were infected with *Bradyrhizobium japonicum* in Hayashi et al. (2012). Two strains of the bacterium were used and compared: a wild-type strain and a NodC strain that cannot synthesize Nod factors. The infected roots were harvested at 2 dpi, the mRNA was analyzed and compared.

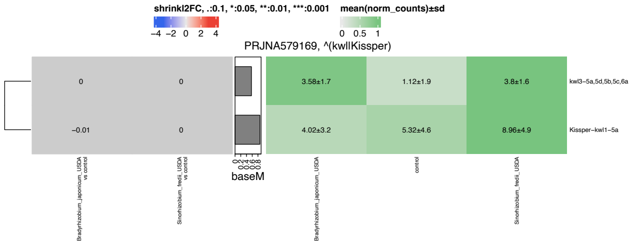
We found a Kwl1 (Kissper-Kiwellins) and one Kwl3 group. No group was strongly expressed or differentially regulated in this experimental setup.



Glycine max (PRJNA579169)

Using the *Rj2* allele, soybean plants can exclude poorly nitrogen-fixing or less useful rhizobia such as *B. japonicum* USDA122 or *Rhizobium fredii* USDA257 from a symbiotic relationship. Host immunity is mediated by the secretory rhizobium type-III-protein NopP and the previously described host resistance protein Rj2. In Shine et al. (2019) transcriptional changes in leaves of *Rj2* virus-silenced plants, each infected with buffer, or one of the two rhizobacteria, will be used to better understand the mechanism of systemic resistance induced by incompatible rhizobia. For this purpose, infected roots were harvested and the mRNA was isolated and analyzed.

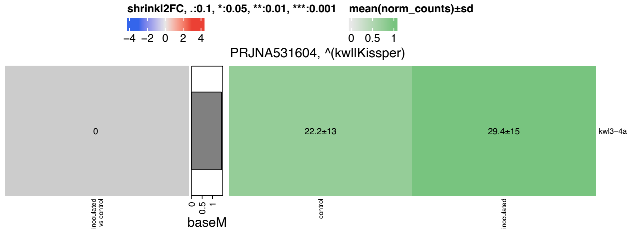
We found 2 groups of weakly expressed Kiwellins (Kissper-Kiwellins of Kwl1 and Kwl3). But in all comparisons, no differential regulation was detected.



Solanum lycopersicum (PRJNA531604)

In Li et al. (2018a) the influence of the endophyte *Pochonia chlamydosporia* on the response of *Solanum lycopersicum* was investigated. For this purpose, plants were infected with the fungus, and samples of the roots were harvested 4, 7, and 21 days after infection. From these tissue samples, mRNA was isolated and analyzed.

No Kiwellin was found to be highly expressed and no differential regulation was observed in response to the fungi.

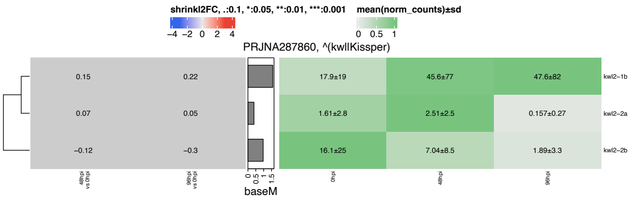


5.4.2 Pathogenic response: P

Musa acuminata (PRJNA287860)

Roots of two-month-old banana seedlings were infected with *Fusarium oxysporum* Race 4 (FocR4)-C1 HIR in Munusamy and Zaidi (2021). Infected root samples were harvested at 2, 48, and 96 hours. The 2 hpi root sample represents the control with which the other two samples were compared.

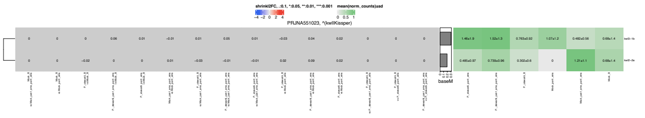
Our reanalysis found 3 weakly to moderately expressed Kiwellins belonging to Kwl2. Neither of these groups shows differential regulation in this experiment.



Zea mays (PRJNA551023)

Pantoea stewartii is the causal agent of Stewart's bacterial wilt of corn and is investigated in Doblas-Ibáñez et al. (2019). With the help of a mutation in the *pan1* gene, it is possible to create resistant corn plants to bacterial disease. Consequently, heterozygous and homozygous (related to the *pan1* gene) maize lines were created by crosses and infected with the bacterium. Subsequently, infected material was harvested one day post-infection, mRNA was isolated, and differences in transcription levels between the different maize lines infected or mock-infected were analyzed.

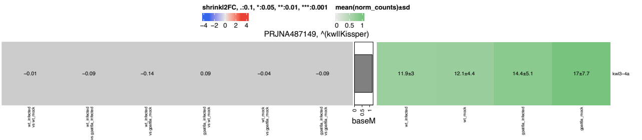
We found Kwl3-1a and Kwl2-2e but both were neither strongly expressed nor showed any differential response.

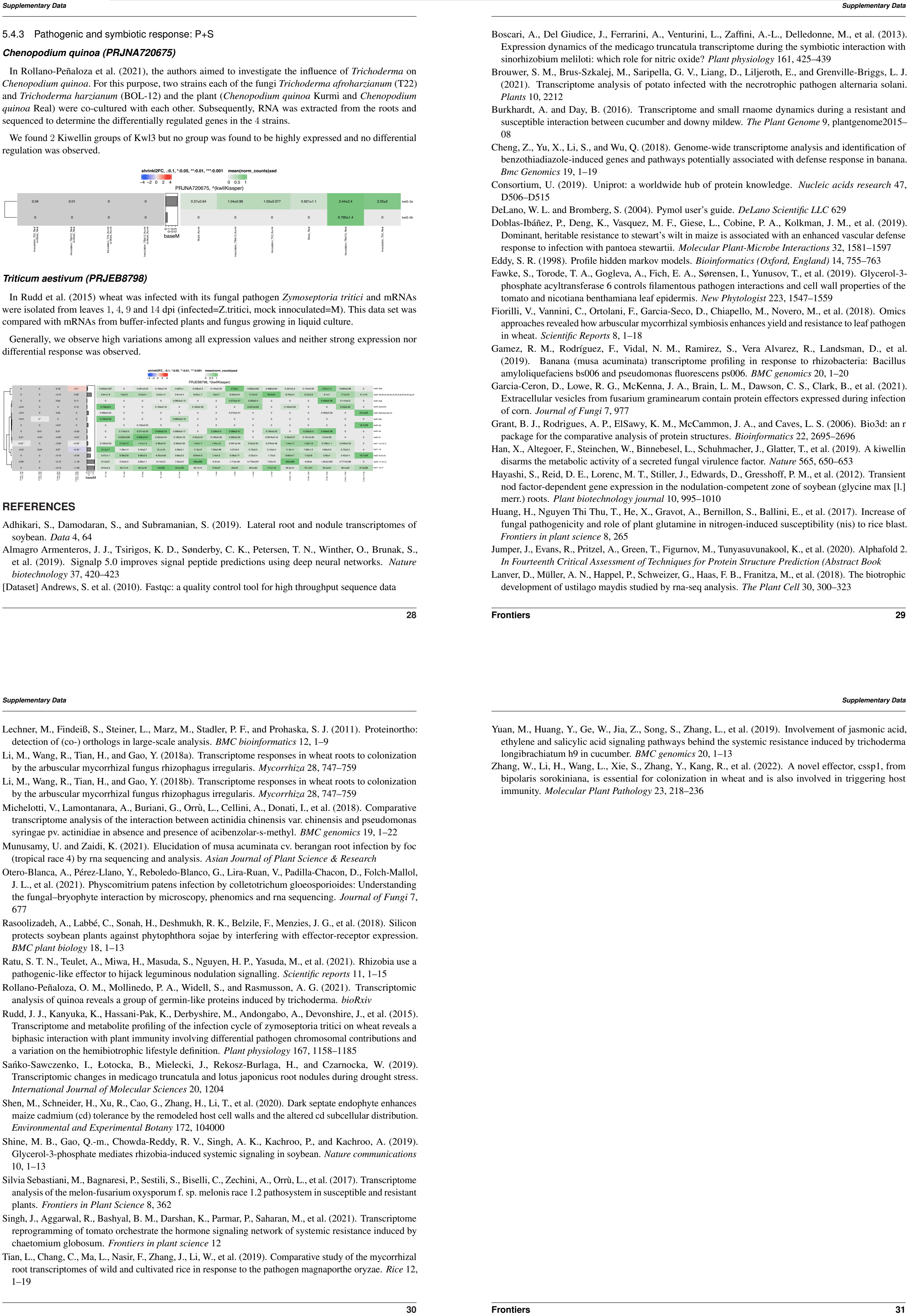


Solanum lycopersicum (PRJNA487149)

In Fawke et al. (2019), the influence of glycerol-3-phosphate acyltransferases on the resistance of *Solanum lycopersicum* to its host *Phytophthora infestans* was investigated. For this purpose, tomato wild-type plants and plants with a loss-of-function mutation in the *gpat6* gene were infected with the fungus. Three days after infection, the leaves were harvested, the RNA isolated, reverse transcribed and the data analyzed.

We found one member of the Kwl3 group that was neither strongly expressed nor showed a differential response.





5.4.3 Pathogenic and symbiotic response: P+S

Chenopodium quinoa (PRJNA720675)

In Rollano-Peñaloza et al. (2021), the authors aimed to investigate the influence of *Trichoderma* on *Chenopodium quinoa*. For this purpose, two strains each of the fungi *Trichoderma afroharzianum* (T22) and *Trichoderma harzianum* (BOL-12) and the plant (*Chenopodium quinoa* Kurmi and *Chenopodium quinoa* Real) were co-cultured with each other. Subsequently, RNA was extracted from the roots and sequenced to determine the differentially regulated genes in the 4 strains.

We found 2 Kiwellin groups of Kwl3 but no group was found to be highly expressed and no differential regulation was observed.

Triticum aestivum (PRJEB8798)

In Rudd et al. (2015) wheat was infected with its fungal pathogen *Zymoseptoria tritici* and mRNAs were isolated from leaves 1, 4, 9 and 14 dpi (infected=Z.tritici, mock innoculated=M). This data set was compared with mRNAs from buffer-infected plants and fungus growing in liquid culture.

Generally, we observe high variations among all expression values and neither strong expression nor differential response was observed.

shrinkl2FC, .:0.1, *:0.05, **:0.01, ***:0.001
mean(norm_counts)±sd
PRJEB8798, ^(kwllKissper)
baseM

REFERENCES

Adhikari, S., Damodaran, S., and Subramanian, S. (2019). Lateral root and nodule transcriptomes of soybean. *Data* 4, 64

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). Signalp 5.0 improves signal peptide predictions using deep neural networks. *Nature biotechnology* 37, 420–423

[Dataset] Andrews, S. et al. (2010). Fastqc: a quality control tool for high throughput sequence data

Boscari, A., Del Giudice, J., Ferrarini, A., Venturini, L., Zaffini, A.-L., Delledonne, M., et al. (2013). Expression dynamics of the medicago truncatula transcriptome during the symbiotic interaction with sinorhizobium meliloti: which role for nitric oxide? *Plant physiology* 161, 425–439

Brouwer, S. M., Brus-Szkalej, M., Saripella, G. V., Liang, D., Liljeroth, E., and Grenville-Briggs, L. J. (2021). Transcriptome analysis of potato infected with the necrotrophic pathogen alternaria solani. *Plants* 10, 2212

Burkhardt, A. and Day, B. (2016). Transcriptome and small rnaome dynamics during a resistant and susceptible interaction between cucumber and downy mildew. *The Plant Genome* 9, plantgenome2015–08

Cheng, Z., Yu, X., Li, S., and Wu, Q. (2018). Genome-wide transcriptome analysis and identification of benzothiadiazole-induced genes and pathways potentially associated with defense response in banana. *Bmc Genomics* 19, 1–19

Consortium, U. (2019). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research* 47, D506–D515

DeLano, W. L. and Bromberg, S. (2004). Pymol user's guide. *DeLano Scientific LLC* 629

Doblas-Ibáñez, P., Deng, K., Vasquez, M. F., Giese, L., Cobine, P. A., Kolkman, J. M., et al. (2019). Dominant, heritable resistance to stewart's wilt in maize is associated with an enhanced vascular defense response to infection with pantoea stewartii. *Molecular Plant-Microbe Interactions* 32, 1581–1597

Eddy, S. R. (1998). Profile hidden markov models. *Bioinformatics (Oxford, England)* 14, 755–763

Fawke, S., Torode, T. A., Gogleva, A., Fich, E. A., Sørensen, I., Yunusov, T., et al. (2019). Glycerol-3-phosphate acyltransferase 6 controls filamentous pathogen interactions and cell wall properties of the tomato and nicotiana benthamiana leaf epidermis. *New Phytologist* 223, 1547–1559

Fiorilli, V., Vannini, C., Ortolani, F., Garcia-Seco, D., Chiapello, M., Novero, M., et al. (2018). Omics approaches revealed how arbuscular mycorrhizal symbiosis enhances yield and resistance to leaf pathogen in wheat. *Scientific Reports* 8, 1–18

Gamez, R. M., Rodríguez, F., Vidal, N. M., Ramirez, S., Vera Alvarez, R., Landsman, D., et al. (2019). Banana (musa acuminata) transcriptome profiling in response to rhizobacteria: Bacillus amyloliquefaciens bs006 and pseudomonas fluorescens ps006. *BMC genomics* 20, 1–20

Garcia-Ceron, D., Lowe, R. G., McKenna, J. A., Brain, L. M., Dawson, C. S., Clark, B., et al. (2021). Extracellular vesicles from fusarium graminearum contain protein effectors expressed during infection of corn. *Journal of Fungi* 7, 977

Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A., and Caves, L. S. (2006). Bio3d: an r package for the comparative analysis of protein structures. *Bioinformatics* 22, 2695–2696

Han, X., Altegoer, F., Steinchen, W., Binnebesel, L., Schuhmacher, J., Glatter, T., et al. (2019). A kiwellin disarms the metabolic activity of a secreted fungal virulence factor. *Nature* 565, 650–653

Hayashi, S., Reid, D. E., Lorenc, M. T., Stiller, J., Edwards, D., Gresshoff, P. M., et al. (2012). Transient nod factor-dependent gene expression in the nodulation-competent zone of soybean (glycine max [l.] merr.) roots. *Plant biotechnology journal* 10, 995–1010

Huang, H., Nguyen Thi Thu, T., He, X., Gravot, A., Bernillon, S., Ballini, E., et al. (2017). Increase of fungal pathogenicity and role of plant glutamine in nitrogen-induced susceptibility (nis) to rice blast. *Frontiers in plant science* 8, 265

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., et al. (2020). Alphafold 2. *In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book*

Lanver, D., Müller, A. N., Happel, P., Schweizer, G., Haas, F. B., Franitza, M., et al. (2018). The biotrophic development of ustilago maydis studied by rna-seq analysis. *The Plant Cell* 30, 300–323

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC bioinformatics* 12, 1–9

Li, M., Wang, R., Tian, H., and Gao, Y. (2018a). Transcriptome responses in wheat roots to colonization by the arbuscular mycorrhizal fungus rhizophagus irregularis. *Mycorrhiza* 28, 747–759

Li, M., Wang, R., Tian, H., and Gao, Y. (2018b). Transcriptome responses in wheat roots to colonization by the arbuscular mycorrhizal fungus rhizophagus irregularis. *Mycorrhiza* 28, 747–759

Michelotti, V., Lamontanara, A., Buriani, G., Orrù, L., Cellini, A., Donati, I., et al. (2018). Comparative transcriptome analysis of the interaction between actinidia chinensis var. chinensis and pseudomonas syringae pv. actinidiae in absence and presence of acibenzolar-s-methyl. *BMC genomics* 19, 1–22

Munusamy, U. and Zaidi, K. (2021). Elucidation of musa acuminata cv. berangan root infection by foc (tropical race 4) by rna sequencing and analysis. *Asian Journal of Plant Science & Research*

Otero-Blanca, A., Pérez-Llano, Y., Reboledo-Blanco, G., Lira-Ruan, V., Padilla-Chacon, D., Folch-Mallol, J. L., et al. (2021). Physcomitrium patens infection by colletotrichum gloeosporioides: Understanding the fungal–bryophyte interaction by microscopy, phenomics and rna sequencing. *Journal of Fungi* 7, 677

Rasoolizadeh, A., Labbé, C., Sonah, H., Deshmukh, R. K., Belzile, F., Menzies, J. G., et al. (2018). Silicon protects soybean plants against phytophthora sojae by interfering with effector-receptor expression. *BMC plant biology* 18, 1–13

Ratu, S. T. N., Teulet, A., Miwa, H., Masuda, S., Nguyen, H. P., Yasuda, M., et al. (2021). Rhizobia use a pathogenic-like effector to hijack leguminous nodulation signalling. *Scientific reports* 11, 1–15

Rollano-Peñaloza, O. M., Mollinedo, P. A., Widell, S., and Rasmusson, A. G. (2021). Transcriptomic analysis of quinoa reveals a group of germin-like proteins induced by trichoderma. *bioRxiv*

Rudd, J. J., Kanyuka, K., Hassani-Pak, K., Derbyshire, M., Andongabo, A., Devonshire, J., et al. (2015). Transcriptome and metabolite profiling of the infection cycle of zymoseptoria tritici on wheat reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions and a variation on the hemibiotrophic lifestyle definition. *Plant physiology* 167, 1158–1185

Sańko-Sawczenko, I., Łotocka, B., Mielecki, J., Rekosz-Burlaga, H., and Czarnocka, W. (2019). Transcriptomic changes in medicago truncatula and lotus japonicus root nodules during drought stress. *International Journal of Molecular Sciences* 20, 1204

Shen, M., Schneider, H., Xu, R., Cao, G., Zhang, H., Li, T., et al. (2020). Dark septate endophyte enhances maize cadmium (cd) tolerance by the remodeled host cell walls and the altered cd subcellular distribution. *Environmental and Experimental Botany* 172, 104000

Shine, M. B., Gao, Q.-m., Chowda-Reddy, R. V., Singh, A. K., Kachroo, P., and Kachroo, A. (2019). Glycerol-3-phosphate mediates rhizobia-induced systemic signaling in soybean. *Nature communications* 10, 1–13

Silvia Sebastiani, M., Bagnaresi, P., Sestili, S., Biselli, C., Zechini, A., Orrù, L., et al. (2017). Transcriptome analysis of the melon-fusarium oxysporum f. sp. melonis race 1.2 pathosystem in susceptible and resistant plants. *Frontiers in Plant Science* 8, 362

Singh, J., Aggarwal, R., Bashyal, B. M., Darshan, K., Parmar, P., Saharan, M., et al. (2021). Transcriptome reprogramming of tomato orchestrate the hormone signaling network of systemic resistance induced by chaetomium globosum. *Frontiers in plant science* 12

Tian, L., Chang, C., Ma, L., Nasir, F., Zhang, J., Li, W., et al. (2019). Comparative study of the mycorrhizal root transcriptomes of wild and cultivated rice in response to the pathogen magnaporthe oryzae. *Rice* 12, 1–19

Yuan, M., Huang, Y., Ge, W., Jia, Z., Song, S., Zhang, L., et al. (2019). Involvement of jasmonic acid, ethylene and salicylic acid signaling pathways behind the systemic resistance induced by trichoderma longibrachiatum h9 in cucumber. *BMC genomics* 20, 1–13

Zhang, W., Li, H., Wang, L., Xie, S., Zhang, Y., Kang, R., et al. (2022). A novel effector, cssp1, from bipolaris sorokiniana, is essential for colonization in wheat and is also involved in triggering host immunity. *Molecular Plant Pathology* 23, 218–236

2.2 6S RNA in LAB

Original Research Article: Cataldo, P. G., **Klemm, P.**, Thüring, M., Saavedra, L., Hebert, E. M., Hartmann, R. K., & Lechner, M. (2021). Insights into 6S RNA in lactic acid bacteria (LAB). *BMC Genomic Data*, 22, 1-15. doi: 10.1186/s12863-021-00983-2

Cataldo *et al. BMC Genomic Data* (2021) 22:29
https://doi.org/10.1186/s12863-021-00983-2

BMC Genomic Data

RESEARCH ARTICLE Open Access

Insights into 6S RNA in lactic acid bacteria (LAB)

Pablo Gabriel Cataldo[1], Paul Klemm[2], Marietta Thüring[2], Lucila Saavedra[1], Elvira Maria Hebert[1], Roland K. Hartmann[2] and Marcus Lechner[2,3]*

Abstract

Background: 6S RNA is a regulator of cellular transcription that tunes the metabolism of cells. This small non-coding RNA is found in nearly all bacteria and among the most abundant transcripts. Lactic acid bacteria (LAB) constitute a group of microorganisms with strong biotechnological relevance, often exploited as starter cultures for industrial products through fermentation. Some strains are used as probiotics while others represent potential pathogens. Occasional reports of 6S RNA within this group already indicate striking metabolic implications. A conceivable idea is that LAB with 6S RNA defects may metabolize nutrients faster, as inferred from studies of *Echerichia coli*. This may accelerate fermentation processes with the potential to reduce production costs. Similarly, elevated levels of secondary metabolites might be produced. Evidence for this possibility comes from preliminary findings regarding the production of surfactin in *Bacillus subtilis*, which has functions similar to those of bacteriocins. The prerequisite for its potential biotechnological utility is a general characterization of 6S RNA in LAB.

Results: We provide a genomic annotation of 6S RNA throughout the *Lactobacillales* order. It laid the foundation for a bioinformatic characterization of common 6S RNA features. This covers secondary structures, synteny, phylogeny, and product RNA start sites. The canonical 6S RNA structure is formed by a central bulge flanked by helical arms and a template site for product RNA synthesis. 6S RNA exhibits strong syntenic conservation. It is usually flanked by the replication-associated recombination protein A and the universal stress protein A. A catabolite responsive element was identified in over a third of all 6S RNA genes. It is known to modulate gene expression based on the available carbon sources. The presence of antisense transcripts could not be verified as a general trait of LAB 6S RNAs.

Conclusions: Despite a large number of species and the heterogeneity of LAB, the stress regulator 6S RNA is well-conserved both from a structural as well as a syntenic perspective. This is the first approach to describe 6S RNAs and short 6S RNA-derived transcripts beyond a single species, spanning a large taxonomic group covering multiple families. It yields universal insights into this regulator and complements the findings derived from other bacterial model organisms.

Keywords: 6S RNA, SsrS, ncRNA, CcpA, cre site, Lactic acid bacteria, LAB

*Correspondence: lechner@staff.uni-marburg.de
[2]Philipps-Universität Marburg, Institut für Pharmazeutische Chemie, Marbacher Weg 6, 35032 Marburg, Germany
[3]Philipps-Universität Marburg, Center for Synthetic Microbiology (Synmikro), Hans-Meerwein-Straße 6, 35043 Marburg, Germany
Full list of author information is available at the end of the article

Background

Lactic acid bacteria

Lactic acid bacteria (LAB) constitute a genotypically, phenotypically, and phylogenetically diverse group of Gram-positive bacteria that belongs to the taxonomic order of the *Lactobacillales*. Shared metabolic characteristics and evolutionary relationships have been used as common markers for the identification, classification, typing, and phylogenetic analysis of LAB species [1]. During the last few decades, the analysis of 16S rRNA gene similarity was combined with the study of the carbohydrate fermentation profile to classify new bacterial isolates. The ongoing exploration of the *Lactobacillus* genus has led to frequent taxonomic rearrangements [2]. One reason is the presence of odd similarities and ambiguities in 16S rRNA gene sequence comparisons, resulting in a biased annotation of strains, species, and even LAB genera at short and long phylogenetic distances [3]. Currently, LAB are grouped into six families: *Aerococcaceae, Carnobacteriaceae, Enterococcaceae, Lactobacillaceae, Leuconostocaceae*, and *Streptococcaceae*. These groups share the ability to catabolize sugars for the efficient production of lactic acid [4]. LAB constitute the most competitive and technologically relevant group of microorganisms *G*enerally *R*ecognized *a*s *S*afe (GRAS). Their biotechnological relevance is a result of the many beneficial features that can be exploited, for instance, as starter cultures in the food industry, mediating the rapid acidification of raw material [4], or as probiotics, preventing the adherence, establishment, and replication of several enteric mucosal pathogens via exerting multiple antimicrobial activities [5]. Nevertheless, some LAB are opportunistic pathogens and can cause infections in individuals presenting some underlying disease or predisposing condition. The most prominent opportunistic pathogens are members of the genera *Streptococcus (S.)* and *Enterococcus* [6].

LAB are usually exposed to a wide range of harsh stresses, both in industrial environments and throughout the gastrointestinal tract. This includes acid, cold, drying, osmotic, and oxidative stresses [7]. Surviving these unfavorable conditions is a prerequisite to exert their expected activities [8]. While main stress-resistance systems have been documented in some LAB species, their regulation at the molecular level, including the role of non-coding RNAs (ncRNAs), is still far from being understood [9].

6S RNA

Over the last decades many small non-coding RNAs have been identified as key regulators in a variety of bacterial stress response pathways and in bacterial virulence [10–12]. A prominent example among these is 6S RNA encoded by a gene frequently termed *ssrS* according to the original gene designation in *Escherichia coli* [13, 14]. A 6S gene is found in nearly all bacterial genomes sequenced so far [15, 16]. This includes species with highly condensed genomes such as the hyperthermophile *Aquifex aeolicus*, species that obtain energy through photosynthesis like *Rhodobacter sphaeroides*, as well as pathogens such as *Helicobacter pylori* [16–19]. The dissemination of 6S RNA and its usually growth phase-dependent and condition-specific expression profile are indicators of the RNA's regulatory impact. Its mechanistic features have been more intensely studied for the two model organisms *E. coli* and *Bacillus subtilis* [20, 21]. The latter belongs to the *Bacillales*, a sister-order of *Lactobacillales*. 6S RNA is about 160-200 nucleotides in length and adopts a rod-shaped structure with an enlarged internal loop or bulge flanked by large helical arms on both sides [22, 23]. 6S RNA can bind the DNA-dependent RNA polymerase (RNAP) in complex with the housekeeping sigma factor (σ^{70} in *E. coli* and σ^{A} in *B. subtilis*) in competition with regular DNA promoters. This sequestration of RNAP alters the housekeeping transcription at a global level that is seemingly advantageous when facing numerous types of stress [22, 24, 25]. When RNAP is bound, it can utilize 6S RNA as a template for the transcription of short product RNAs (pRNAs). Upon relief of stress, the transcribed pRNAs become increasingly long. When reaching a certain length ($\sim$14 nt in *B. subtilis*), pRNAs can persistently rearrange the structure of 6S RNA to induce RNAP release, thus restoring regular transcription [21, 26–30]. Studies in *E. coli* have provided evidence that nutrients are metabolized faster in 6S RNA knockout strains than in the parental wild type strain [29, 31]. Furthermore, knockout strains might have the so far unexplored potential to produce elevated levels of secondary metabolites such as surfactants.

6S RNA in lactic acid bacteria

The importance of 6S RNA in LAB is indicated by studies that report its abundant expression as well as metabolic changes upon its knockout. However, specific 6S RNA analyses in this important group of bacteria are scarce or the studied ncRNA was not recognized as 6S RNA. It is annotated only in about half of all LAB species analyzed in this study (539/1,092 genomes). Here, we identified it in about 91% of all known LAB species. An example is *L. delbrueckii*, an industrial starter for dairy products, where a highly abundant ncRNA was reported [32]. Though its function could not be specified further, the authors suspected it to act as an antisense RNA. In our study, we identified this 210 nt long ncRNA as 6S RNA. In another study, 6S RNA was identified along with two types of pRNAs via RNA sequencing of *S. pyogenes* [33].

For *Lactococcus lactis*, the expression of 6S RNA has been linked to the carbon catabolite repression protein CcpA that binds to DNA at *cis*-acting sequences. These sites are called catabolite responsive elements (*cre*) [34];

cre sites are degenerate pseudo-palindromes. In Bacilli a CcpA dimer was shown to bind to dsDNA upon association with the Ser46-phosphorylated form of histidine-containing phosphocarrier protein (HPr-Ser46-P) [35]. In *L. lactis*, 6S RNA levels were found to be increased during stationary and exponential phase in the presence of galactose or cellobiose, but not fructose, as the sole carbon source. CcpA repression is known to be relieved by galactose and cellobiose, but not by fructose. Moreover, 6S RNA was found to be about 3-fold upregulated in a CcpA-deficient mutant [34] and a *cre* element was identified upstream of the -35 region of its promoter. This indicates a potential interaction between CcpA and the 6S RNA gene that might be relevant for LAB in general. Notably, *B. subtilis* 6S-1 and 6S-2 RNA were not identified as a target for CcpA [36].

For *E. faecalis*, a major opportunistic human pathogen, an additional transcript antisense to 6S RNA was detected [37]. The authors proposed its participation in degradation or maturation of 6S RNA as both ncRNA products were present in a processed form. To our knowledge, an equivalent antisense product is not described for *E. coli* [37], *B. subtilis* or any other species to date (own observation). However, interdependent expression of genes around the 6S RNA locus was noticed for other bacteria, e.g. *R. sphaeroides* (Proteobacteria), where a salt stress-induced membrane protein gene on the opposite strand immediately downstream of the 6S RNA locus is expressed at elevated levels in a 6S RNA knockout strain [18].

Apart from these isolated findings, little is known about the sequence, structure, and physiological role of this regulatory ncRNA in the large and widely heterogeneous group of LAB. In this study, we have annotated and analyzed 6S RNAs systematically to lay a foundation for further investigations regarding its role in stress responses, metabolic processes and interactions with eukaryotic cells. Moreover, we investigated how wide-spread and universally relevant the species-specific observations stated above are for LAB (link to CcpA and the presence of an antisense transcript). This is also the first comparative study covering 6S RNAs in a set of taxonomic families, thus making it possible to draw more representative conclusions than in species-wise studies.

Results

Dissemination & phylogeny

We searched 6S RNA sequences in 1,092 genomes covering strains from all 371 sequenced LAB species publicly available in the NCBI database at the time of this study [38]. While two 6S RNA copies were reported for some *Firmicutes* including *Bacillus subtilis*, *Bacillus halodurans*, *Clostridium acetobutylicum*, *Oceanobacillus iheyensis*, and *Thermoanaerobacter tengcongensis* [15], only one copy is present in LAB species. It shows more similarity to the major and well described *Bacillus subtilis* 6S-1 RNA than to its paralog 6S-2 RNA [39].

6S RNA was located in 1001 genomes (> 91%). Additional File 1 lists all loci. Genomes in which a 6S RNA gene could not be identified are predominantly partial genomes with a large number of contigs or scaffolds. When a 6S RNA gene was found in genomes of closely related species/strains, we assumed that the ncRNA is present but not part of the assembly yet. A peculiarity is the genus *Weissella* of the *Leuconostocaceae* family, represented with 13 species in our dataset. While only a weak 6S RNA locus was predicted in no more than four species of this genus, a significant amount of transcription could be shown for the syntenically conserved intergenic region downstream of *rarA* in publicly available RNA-Seq data for *W. confusa* and *W. koreensis* [40, 41]. Moreover, this locus is confined by a transcription terminator in most *Weissella* species. See Additional File 8 for details. This indicates that 6S RNAs in *Weissella* have a distinct singularity that was hardly picked up by our covariance-based search strategy. The typical rod-shaped structure with a central loop or bulge could not be confirmed for these non-canonical candidates.

Figure 1 shows the phylogeny of canonical 6S RNAs identified here based on their sequences and structural properties reconstructed using `RNAclust` [42] and `mlocarna` [43]. An alternative version with a resolution that reaches the species level is provided in Additional File 2. The phylogeny well resembles the taxonomic units at the level of genera. A minor exception is the *Carnobacteriaceae* group (blue) that includes *Abiotrophia defectiva* (*Aerococcaceae*) and *Bavariicoccus seileri* (*Enterococcaceae*). At the level of taxonomic families, the genus *Vagococcus* is significantly different from other *Enterococcaceae* (green). Similarly, *Aerococcus* is different from other *Aerococcaceae*. *Lactobacillus* is known to be the most heterogeneous genus within LAB [1]. This is also reflected phylogenetically since the 6S RNAs of this genus are divided into eight well distinguishable groups (*Lactobacillus* 1-7, *Pediococcus*, brown).

Relation to 16S rRNA phylogeny

The phylogenetic reconstruction of LAB species based on a sequence alignment of selected 16S rRNA sequences is shown analogous to the 6S RNA-based reconstruction in Additional File 3. As expected, the 16S rRNA-based approach better resembles the current taxonomic annotation [2, 44]. The majority of *Lactobacillaceae* species share a common subtree. Notably, a number of species from the Lactobacillus 6 group (6S RNA-based, see Fig. 1) is also located in a separate subtree in the 16S rRNA phylogeny. Similarly, the *Vagococcus* group is isolated from the remaining *Enterococcaceae* in both phylogenies and

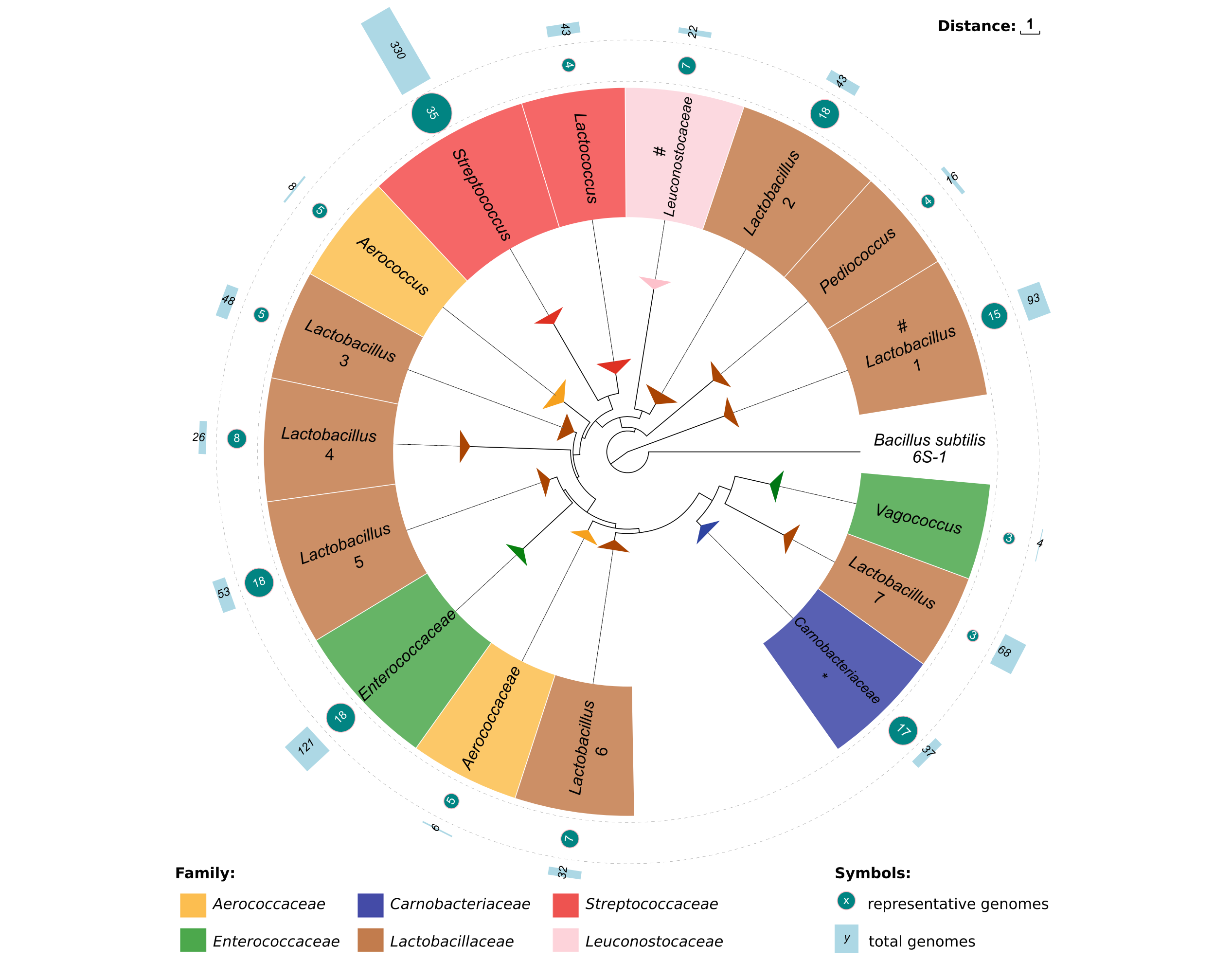


Fig. 1 Phylogenetic reconstruction of LAB based on sequence and structure of 6S RNA. 6S-1 RNA from *B. subtilis* is used as an outgroup. The number of different LAB strains is indicated on the outer ring. Turquoise circles show the number of unique 6S RNA sequences within each group. The asterisk at *Carnobacteriaceae* indicates that two species in the group belong to another family. The number sign at *Leuconostocaceae* and *Lactobacillus* 1 remarks non-canonical secondary consensus structures

the same two family-foreign species are found within the *Carnobacteriaceae* subtree, namely *A. defectiva* (*Aerococcaceae*) and *B. seileri* (*Enterococcaceae*). In the 16S rRNA tree, the grouped *Aerococcaceae* are closely related to *Carnobacteriaceae*. The 6S RNA tree, in contrast, splits this group into two subgroups that are not closely related to *Carnobacteriaceae*.

Synteny

To characterize the genomic locus of 6S RNA in LAB, a synteny analysis was performed. Proteinortho [45] was used to group the protein-coding genes in the vicinity of the 6S RNA locus. An overview of the genomic context of 6S RNA in LAB is shown in Fig. 2 and in more detail in Additional File 4. The genomic neighborhood of 6S RNA is conserved at the family level. Typically, the same genes are encoded up- and downstream of 6S RNA in the majority of genera from the same taxonomic family but not across LAB in general. Exceptions are the replication-associated recombination protein A gene (*rarA*), that is found upstream of the 6S RNA locus in nearly all species, and the universal stress protein A gene (*uspA*), that is found downstream across almost all species except for *Streptococcaceae* and a few *Aerococcaceae* members.

The upstream *rarA* gene is part of a highly conserved family of ATPases found in prokaryotes as well as eukaryotes. Homologs are known as *mgsA* in *E. coli*, *mgs1* in yeast (maintenance of genome stability A/1), and *WRNIP1* (Werner interacting protein 1) in mammals. The encoded protein is involved in cellular responses to stalled or collapsed replication forks, likely by modulating replication restart [46–48].

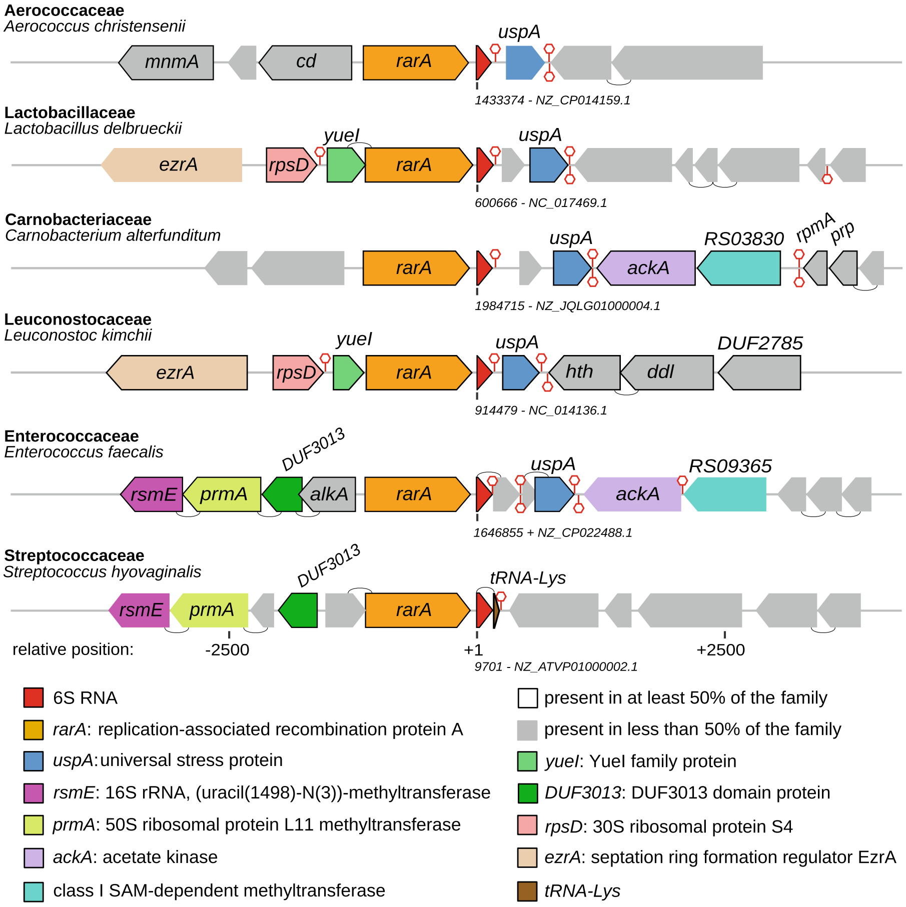

Aerococcaceae
Aerococcus christensenii
mnmA
cd
rarA
uspA
1433374 - NZ_CP014159.1
Lactobacillaceae
Lactobacillus delbrueckii
ezrA
rpsD
yueI
rarA
uspA
600666 - NC_017469.1
Carnobacteriaceae
Carnobacterium alterfunditum
rarA
uspA
ackA
RS03830
rpmA
prp
1984715 - NZ_JQLG01000004.1
Leuconostocaceae
Leuconostoc kimchii
ezrA
rpsD
yueI
rarA
uspA
hth
ddl
DUF2785
914479 - NC_014136.1
Enterococcaceae
Enterococcus faecalis
DUF3013
rsmE
prmA
alkA
rarA
uspA
ackA
RS09365
1646855 + NZ_CP022488.1
Streptococcaceae
Streptococcus hyovaginalis
DUF3013
rsmE
prmA
rarA
tRNA-Lys
relative position:
-2500
+1
+2500
9701 - NZ_ATVP01000002.1
🟥 6S RNA
🟧 rarA: replication-associated recombination protein A
🟦 uspA: universal stress protein
🟪 rsmE: 16S rRNA, (uracil(1498)-N(3))-methyltransferase
🟨 prmA: 50S ribosomal protein L11 methyltransferase
🟪 ackA: acetate kinase
🟦 class I SAM-dependent methyltransferase
□ present in at least 50% of the family
▨ present in less than 50% of the family
🟩 yueI: YueI family protein
🟩 DUF3013: DUF3013 domain protein
🟥 rpsD: 30S ribosomal protein S4
🟧 ezrA: septation ring formation regulator EzrA
🟫 tRNA-Lys

Fig. 2 Genomic context of 6S RNA in LAB (4 kb upstream and downstream of the 6S RNA gene). For each LAB family, the genomic locus of one representative species is shown. Genes present in ≥ 50% of the respective family are indicated with a solid border. Genes found in multiple families are colored. Hypothetical and less conserved proteins are unmarked. Putative Rho-independent terminators are indicated by red hexagons. Genes in close proximity (<20 nt) are indicated by a semicircle connecting them. These could be part of a polycistronic transcript. The complete list of genomic contexts including the NCBI reference codes is provided in Additional File 4. Further gene locus abbreviations: *mnmA*, tRNA 2-thiouridine(34) synthase MnmA; *cd*, cystein desulfurase; *rpmA*, 50S ribosomal protein L27; *prp*, ribosomal-processing cysteine protease Prp; *hth*, helix-turn-helix domain-containing protein; *ddl*, D-alanine-D-alanine ligase; *alkA*, DNA-3-methyladenine glycosylase (adaptive response to alkylative DNA damage)

The downstream *uspA* gene belongs to a superfamily that encompasses an ancient and highly conserved group of proteins that are widely distributed among bacteria, archaea, fungi, flies, and plants. It was found to be induced during metabolic, oxidative, and temperature stress in *Salmonella typhimurium* [49] and linked to cell sensitivity to ultraviolet light in *E. coli* [50]. *uspA* is known to be differentially expressed in response to a large number of different environmental stresses such as acid and salt stresses, starvation, exposure to heat, oxidants, metals, ethanol, antibiotics, and other stimulants - particularly within the genera *Lactobacillus*, *Streptococcus*, *Enterococcus* and *Lactococcus* [51–53].

Structure and sequence conservation

The consensus structure and sequence conservation of 6S RNA in LAB based on a mLocARNA [43] alignment combined with RNAalifold [54] is illustrated in Fig. 3. Additional File 5 shows the consensus structures at the family level. The consensus of 6S RNA in LAB follows the well-known secondary structure of the canonical 6S RNA [15, 23], featuring an outer closing stem with smaller bulges and loops, a large 5'-central bulge and an apical stem with smaller internal loops capped by the terminal loop L1. Opposite to the 5'-central bulge a hairpin is predicted that was also shown to form in *B. subtilis* 6S-1 RNA [26]. The central bulge harbors the initiation site for

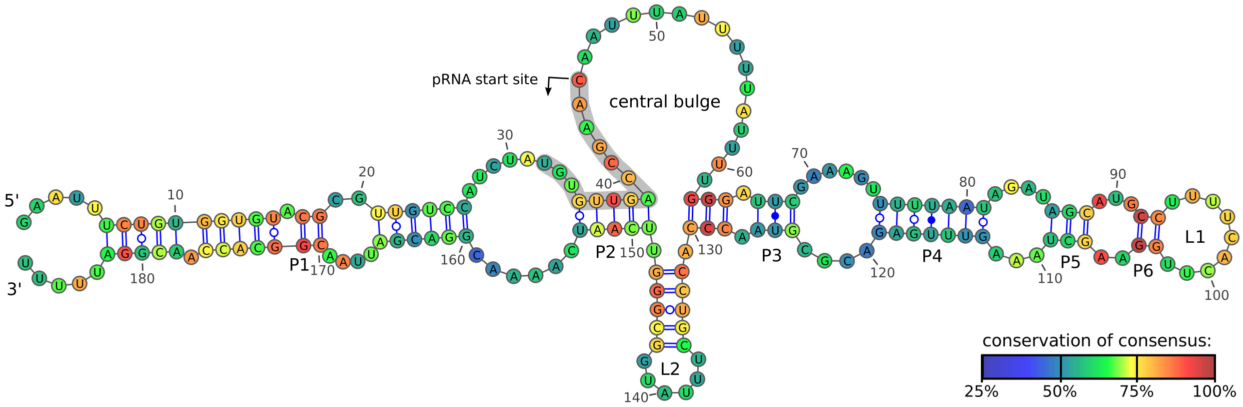


Fig. 3 Consensus secondary structure of 6S RNA in LAB. The structure is derived from a sequence-structure-based alignment of 172 unique representative sequences (see Materials and Methods for further details). Colors indicate sequence conservation within LAB. Paired regions P1-P6, the 5'-central bulge, terminal loops L1/L2, and the putative transcription start site of pRNAs are indicated

product RNA (pRNA) transcription. This consensus and canonical 6S secondary structure is evident in most of the 6S RNA groups: *Aerococcaceae*, *Aerococcus*, *Carnobacteriaceae*, *Vagococcus*, *Enterococcaceae*, *Pediococcus*, *Lactobacillus* 2, 3, 4, 6, 7, *Streptococcus*, and *Lactococcus*, see Additional File 5.

Product RNAs

Putative pRNA transcription start sites were inferred from a structural alignment (see Materials and Methods) of 172 representative 6S RNA sequences from LAB species and in relation to those of *E. coli*, *R. spheroides* and *B. subtilis* for which the start sites are experimentally proven. Fig. 4 shows the overall sequence motif. The first eleven nucleotides of the pRNAs are well conserved. This conservation diminishes starting at position 12. GG at position 5/6 as well as AA at position 9/10 are the most conserved in this group. Two G residues are also conserved in experimentally verified pRNAs from more distantly related bacteria such as the Gram-negatives *E. coli*, *A. aeolicus* and *R. spheroides*, but in these cases at positions 4/5 (Fig. 4). Notably, a highly conserved adenine immediately upstream of the pRNA start sites was identified in the 6S RNAs of LAB species as well as in the reference 6S RNAs included in Fig. 4.

Based on the pRNA sequence (positions 1-15), LAB pRNAs are closely related to pRNAs synthesized from

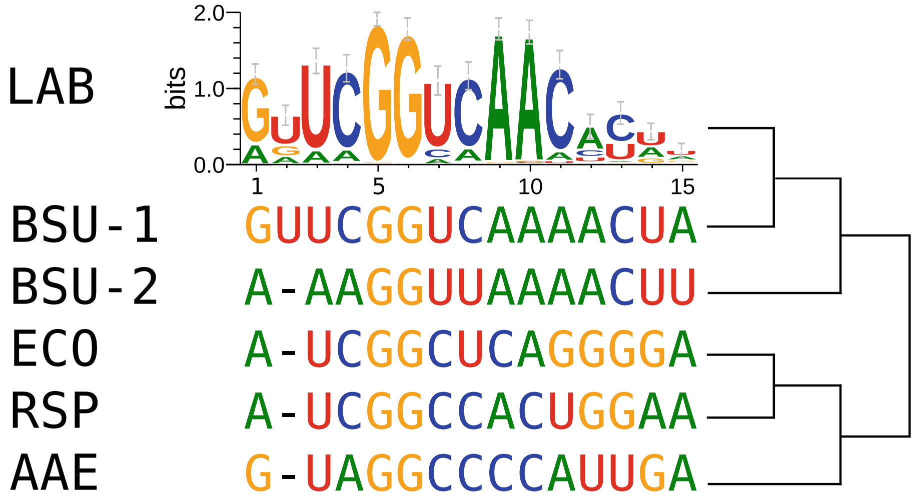


Fig. 4 Consensus sequence motif of 6S RNA-derived pRNAs in LAB. The motif found in LAB is indicated at the top. Positions are numbered from the pRNA 5'-end. Known pRNA sequences of other organisms are shown below the motif (BSU-1/2: *B. subtilis* 6S-1 and 6S-2 RNA, ECO: *E. coli*, RSP: *R. spheroides*, AAE: *A. aeolicus*). The conserved GG at position 4/5 or 5/6 is also encoded in 6S RNAs of bacteria outside the LAB group. A neighbor-joining tree based on the LAB consensus and the pRNA sequences (positions 1-15) is indicated on the right

B. subtilis 6S-1 RNA as template (Fig. 4). Although the 6S-1 pRNA sequence shows differences to the LAB pRNA consensus, major hallmarks (upstream adenine, GG dinucleotide, AA at position 9/10) are still present. Hence, despite the considerable phylogenetic distance, similarities to the pRNA sequence found in LAB are clearly recognizable.

We screened 115 publicly available RNA-Seq datasets for expression of 6S RNA and the presence of pRNAs. These small transcripts are usually depleted in sample preparation for RNA-Seq or neglected in data processing that typically focuses on longer RNAs such as tRNAs or mRNAs. Moreover, we found that pRNAs are underrepresented in adapter ligation libraries compared to poly(A)-tailing libraries [55]. It is thus not surprising that only small numbers of pRNA reads were identified in most RNA-Seq libraries. We yet found robust evidence for pRNAs in *Streptococcus pneumoniae* and *Streptococcus pyogenes* RNA-Seq data (Fig. 5), which also supports the predicted pRNA start site (Figs. 3 and 4) [56, 57]. Two pRNA transcripts were previously reported for *S. pyogenes*, but their sequences were not provided [33]. Here we confirm these findings. We find one alternative transcription start site (pRNA*) located around position 136 that starts at the beginning of the L2 loop (see Fig. 3). The alternative pRNA transcript likely results from 6S RNA binding RNAP in inverse orientation. Similar observations have been made for *Helicobacter pylori* [19]. Notably, neither the pRNA nor the pRNA* sequences have alternative matches in the respective genomes. It is thus unlikely that these transcripts derive from another locus. Additional File 6 illustrates further RNA-Seq results. While pRNAs were also found in libraries from *E. faecalis*, the number of reads is too low to draw safe conclusions.

CcpA-binding catabolite responsive elements

A functional *cre* site upstream of the 6S RNA promoter was reported in *L. lactis*, suggesting that 6S RNA expression is regulated depending on the available carbon source [34]. An equivalent *cre* site could be found in about one-third of all LAB species. Fig. 6 illustrates the location and sequence conservation of the two *cre* sites at the 6S RNA locus. Additional File 2 shows a detailed overview of all species with *cre* sites in the 6S RNA region. Additional File 7 lists the respective motif sequences along with their positions and p-values. *cre* sites are most frequently found in *Enterococcaceae* but also in several *Streptococcaceae* and the *Lactobacillus* groups 6 and 7 (see Fig. 1). Mainly in *Streptococcaceae* and *Lactobacillus* group 6, potential *cre* sites were also identified within the 6S RNA coding sequence. Notably, *L. coryniformis*, *L. rennini*, *L. vaginalis*, *S. canis*, *S. didelphis*, *S. equi*, *S. pantholopis* and *S. phocae* do not have a strong, detectable *cre* site at the 6S RNA promoter but only within the 6S RNA coding region; both sites were detected in *L. backii*, *L. bifermentans*, *S. castoreus*, *S. gallolyticus*, *S. halotolerans*, *S. ictaluri*, *S. iniae*, *S. parauberis* and *S. uberis*.

Expression and antisense transcripts

A total of 115 publicly available RNA-Seq libraries representing 24 different LAB genera were screened for the expression of 6S RNA, pRNAs and long antisense transcripts as described for the *Enterococcus faecalis* V583 strain [37]. Detailed results for each library are shown in Additional File 6.

6S RNA transcripts were highly abundant in general (usually 1-2% of all reads in the RNA-Seq libraries), indicating active transcription in LAB grown under a wide variety of culture conditions and stresses. In line with previous findings [37], however, we did not find evidence for long antisense transcripts of 6S RNA in any RNA-Seq library including those from other *Enterococcus faecalis* strains (OG1RF, 12030, and ATCC 29212), indicating that such transcripts are not a common trait among LAB.

Discussion

Here we identified the 6S RNA gene at a well-conserved genomic locus in LAB species that distinguishes this bacterial group from related bacterial clades. While the

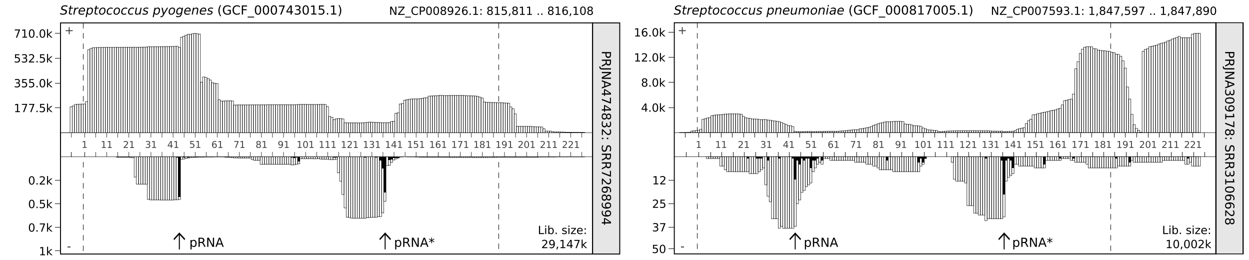


Fig. 5 Publicly available RNA-Seq datasets of *Streptococcus pyogenes* (left) and *Streptococcus pneumoniae* (right) mapped to the 6S RNA locus. 6S RNA transcripts are shown in the upper part. pRNA sequences are shown in the lower part in antisense direction. In each case, two short antisense transcripts can be found (pRNA, pRNA*, arrows indicate start sites)

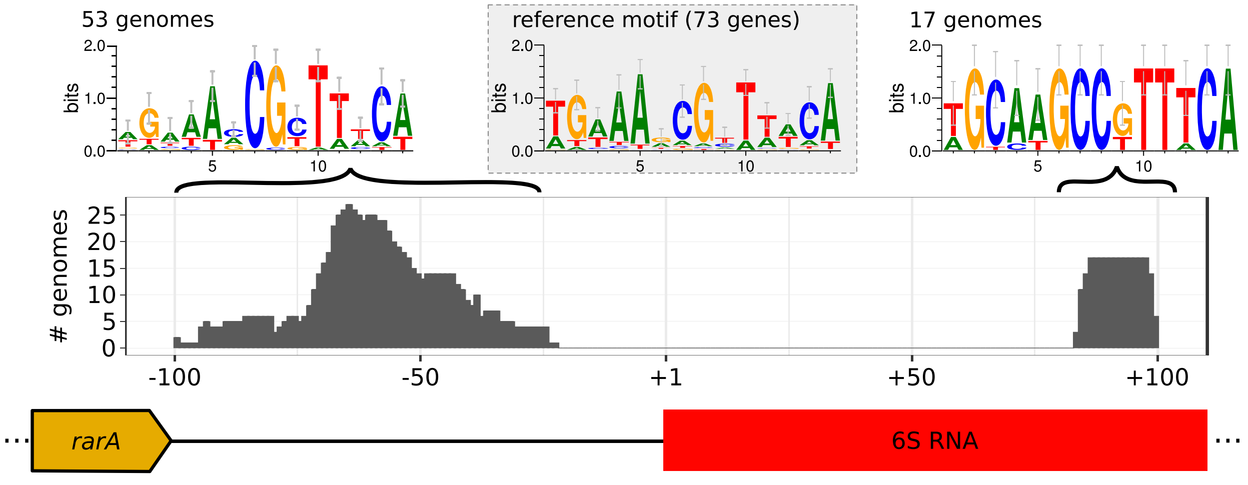


Fig. 6 Position and motif of located *cre* sites. Motifs indicated at the top represent the *cre* sites upstream of the 6S RNA promoter (left) and within the 6S RNA gene (right). Both show high conservation. The experimentally verified *cre* motif of 73 genes of *L. lactis* [79] is shown in the center for comparison

consensus secondary structure is typically canonical as described for *B. subtilis* 6S-1 RNA, we could not verify this for candidates of the genus *Weissella*. Nevertheless, we identified evidence for significant transcription of the respective loci in publicly available RNA-Seq libraries for two strains, see Additional File 8. This confirms a weak 6S RNA candidate in *W. koreensis*. Although no relevant match was found for *W. confusa*, the intergenic region downstream of the syntenically conserved *rarA* showed transcription that matched a 6S RNA transcript even though its putative secondary structure did not match a canonical 6S RNA. A TATAAT sequence is present at the -10 region of all candidates reported for *Weissella*, indicating the presence of a promoter. Similarly, a rho-independent terminator was predicted at the RNA's proposed 3'-end. Thus, the presence of an actively transcribed 6S RNA-like transcript can be assumed. It will be interesting to investigate the functional consequences of this structural alteration.

Carbon catabolite control is a major regulatory mechanism for the modulation of metabolic activity of microorganisms to optimize carbon metabolism and energy use. It involves both carbon catabolite repression and activation. In most low-GC-content Gram-positive bacteria this regulation is mediated by the catabolite control protein A (CcpA) that binds to DNA at *cis*-acting sequences. These are called catabolite responsive elements (*cre*) and are located either in the promoter region or within the coding sequence of the regulated gene [36]. CcpA can function as an activator or may repress transcription depending on its location within a regulated gene or operon [58]. We found strong evidence for *cre* sites upstream of the 6S RNA promoter in about a third of all LAB species, mainly in *Enterococcaceae* but also in *Streptococcaceae* and some *Lactobacillus* subgroups. For *Streptococcaceae* and *Enterococcaceae*, the presence and regulatory importance of these *cre* sites has been reported and studied previously [59, 60]. On the basis of previous reports, our findings suggest that 6S RNA expression is under the negative control of CcpA in many LAB species. This was shown e.g. for *L. lactis* where 6S RNA is 3-fold upregulated upon deletion of the *ccpA* gene [34].

For several 6S RNA genes, *cre* sites were also identified internally - in some cases in addition to the site at the 6S RNA promoter (see Additional File 2). The presence of two *cre* sites regulating the expression of *cid* and *lrg* genes in *Streptococcus mutans* has already been described, but in this case both sequences were upstream of the transcription start site of the above-mentioned genes [61]. In *B. subtilis*, *cre* sites upstream of promoters were found to be primarily activated by CcpA, while *cre* sites overlapping promoters had repressing effects [35]. As the *cre* sites in LAB overlap the -35 region of 6S RNA gene promoters (Fig. 6), CcpA-binding is likely inhibitory; *cre* sites located further downstream of the transcription start site may act as roadblocks or repress initiation of transcription through interaction with RNAP [62]. Future studies may address the interplay of the two *cre* sites at/within the 6S RNA gene. Although speculative at present, it is also a possibility that CcpA binds to 6S RNA at the internal *cre* site, taking into account that 6S RNAs mimic an open DNA promoter [22].

The identified *cre* sequences share a high degree of similarity to the consensus sequences previously described for other LAB such as *L. lactis*, (see Fig. 6) as well as to other Gram-positive bacteria such as *B. subtilis* [36, 63].

Recent studies on the promoter region of the *PTS-IIC* gene cluster of *L. lactis* demonstrated the importance of nucleotide identity at positions 7 and 12 of the 14-nt long *cre* site. Specific mutations within the -35 promoter element resulted in constitutive expression of the downstream gene in the presence of glucose, while other mutations enhanced promoter activity in the presence of cellobiose [63].

The prediction of transcription start sites for pRNAs was based on the structural alignment to other 6S RNAs and could be verified by RNA-Seq data in two cases. This study is the first that deduces pRNAs for a large taxonomic group covering multiple families. We found a highly conserved sequence up to around position 11. This may point to similar kinetics of pRNA synthesis and pRNA-induced 6S RNA refolding [26]. Strikingly, GG at positions 5/6 or 4/5 of the pRNAs appears to be a key feature conserved beyond LAB.

A general property of the 6S RNA locus in LAB is its location between the *rarA* and *uspA* genes. Gene order conservation can be used not only to evaluate the orthology of genomic regions but might also hint at functional relationships between genes [64]. RarA is proposed to act at stalled DNA replication forks upon DNA damage and UspA alters the expression of a variety of genes that help to cope with stresses. As 6S RNA was shown to have a role in cellular stress responses to ensure long-time cell survival, all three gene products might be part of an overachrching stress response network. The *rarA* gene is in close vicinity to the 6S RNA locus across all families including the 6S-1 RNA locus of the non-LAB firmicute *B. subtilis* (see Additional File 4). In the latter, however, *rarA* is encoded in the opposite direction and known to be monocistronic [65]. The RNA-Seq data presented in Additional Files 6 and 8 and the presence of a downstream terminator in most species indicates that the 6S RNA gene is monocistronic as well. However, several *Streptococcaceae* members encode a tRNA-Lys immediately downstream of 6S RNA, suggesting that both genes are part of the same operon. This assumption is supported by RNA-Seq data for *S. pneumoniae* (Additional File 6, p. 43) showing that both ncRNAs have the same transcript level [56]. Thus, both RNAs are likely processing products of the same primary transcript. Other notable syntenic bonds are not universally preserved for LAB but within and also across particular LAB families. Examples are the acetate kinase, class I SAM-dependent methyltransferase, 16S rRNA methyltransferase, and the 50S ribosomal protein L11 methyltransferase. While the function of the other frequently linked genes is unknown so far, this data suggests a cluster of growth-relevant and stress-related genes that 6S RNA is part of. Typically, these genes appear to be transcribed independently (with the exception of 6S RNA and tRNA-Lys in a number of *Streptococcaceae*). Therefore, the possibility of a common functional context remains vague at present.

Conclusions

Lactic acid bacteria include highly heterogenous species and the study of the role of non-coding RNA molecules, particularly 6S RNA, in the regulation of the response of these bacteria to different stress conditions has many potential applications, both within industrial and health contexts. The global transcription regulator 6S RNA is present in nearly all species and well-conserved throughout this group. It generally resembles the canonical form that is well described for *B. subtilis* 6S-1 RNA. LAB 6S RNAs also share the syntenic proximity to *rarA*, located upstream of 6S RNA in nearly all LAB genomes. Many species additionally encode the UspA protein downstream of 6S RNA, which makes its identification comparably easy. The experimental evidence that was processed and analyzed in this study also demonstrated that 6S RNA is expressed in a multitude of LAB species across all taxonomic families and under varying culture conditions. This also highlights the important regulatory role of this ncRNA in bacterial metabolism, further supported by the frequent presence of *cre* sites in its promoter and coding region. The conservation of 6S RNAs makes it plausible to generally apply our findings to any LAB species in order to explore its biotechnological potential.

Methods

Genomes

Several thousand genomes representing 576 species that cover 48 genera were listed as part of the *Lactobacillales* order according to the NCBI taxonomy classification (date of retrieval 10/09/2018) [38]. In order to work with a reasonably representative set, we focused on the genomes with the best respective assembly status for each species. The species *Enterococcus faecium* for example comprises 1109 genomes/subspecies. Fifty-one out of these are marked as “Complete Genome” and were thus considered in the present work. *Lactobacillus fuchuensis* is represented with three genomes out of which the most complete assembly is marked as “Chromosome” that was thus considered, and so on. Additionally, we added 13 strains that were characterized by our institute (CERELA-CONICET) even though they did not meet this criterion. Species with yet unclear specific names (sp.) were neglected. A total of 1,092 genomes were considered in this study. An overview of the genera analyzed here can be found in Table 1. A detailed list of the species and genome assembly levels is provided in Additional File 1. The respective genomes and genomic annotations were downloaded via ftp.ncbi.nlm.nih.gov from the NCBI database [38].

6S RNA prediction

Putative 6S RNAs encoded in LAB genomes were identified in multiple steps. A BLAST-based approach was performed using available 6S RNA annotations given in the NCBI RefSeq annotation, from Wehner *et al.*, and from the Rfam seed sequences for the 6S/SsrS RNA family (RF00013, Version 14) to cover the currently known 6S RNAs [16, 66, 67]. An e-value threshold of 10^{-30} was applied. Previously not annotated 6S RNAs were identified with a covariance-based search performed with INFERNAL (v1.1.1) [68] using the "6S/SsrS RNA" family model as query (see above). Initially, no thresholds were set. Based on the assumption that each genome should encode at least one 6S RNA gene, the highest-scoring hit for each genome was assumed as a true hit. Compared to this, the e-values of the second-best hits

Table 1 Genomes overview

Family	Genus	Genomes used / Genomes available
Aerococcaceae	*Abiotrophia*	1 / 2
	Aerococcus	8 / 61
	Dolosicoccus	2 / 3
	Eremococcus	1 / 2
	Facklamia	3 / 9
	Globicatella	1 / 4
Carnobacteriaceae	*Agitococcus*	1 / 1
	Alkalibacterium	1 / 8
	Allofustis	1 / 1
	Atopobacter	1 / 1
	Atopococcus	1 / 1
	Carnobacterium	9 / 41
	Dolosigranulum	10 / 12
	Granulicatella	1 / 7
	Jeotgalibaca	1 / 4
	Lacticigenium	1 / 1
	Marinilactibacillus	1 / 5
	Trichococcus	7 / 15
Enterococcaceae	*Bavariicoccus*	1 / 1
	Enterococcus	114 / 2105
	Melissococcus	2 / 14
	Tetragenococcus	5 / 19
	Vagococcus	4 / 6
Lactobacillaceae	*Lactobacillus*	460 / 1680
	Pediococcus	25 / 61
	Sharpea	1 / 4
Leuconostocaceae	*Convivina*	1 / 1
	Fructobacillus	5 / 9
	Leuconostoc	23 / 118
	Oenococcus	3 / 208
	Weissella	23 / 43
Streptococcaceae	*Floricoccus*	2 / 2
	Lactococcus	44 / 168
	Streptococcus	328 / 12076

Distribution and number of genomes that were retrieved and downloaded from the NCBI database according to the "most complete genome" criterion were worse by orders of magnitude. A manual inspection on a sample basis confirmed that those were not likely to be valid 6S RNA candidates. Hence, an e-value threshold of 10^{-8} was applied. In this case, a primary hit was found in most species while unexpected secondary hits were rare and could be judged manually in later stages. Overlapping hits were joined. Hits were found in 973 out of 1092 genomes. Redundant sequences were merged to a single representative sequence resulting in 330 unique sequences that were aligned using `Clustal Omega` (v1.2.1) [69]. Sequences with an edit distance of ten or less were merged to their consensus sequence to further reduce the amount of redundancy. 188 representative 6S RNA sequences remained. We checked for isolated sequences in the secondary structure clustering analysis (see below) and non-canonical secondary structures using RNAfold (v2.1.9) [54]) as well as suspicious alignments to further remove non-canonical and doubtful hits. The following sixteen 6S RNA candidates were discarded manually in the first round: *Agitococcus lubricus*, *Lactococcus fujiensis*, *Facklamia hominis*, *Pediococcus damnosus*, *Lactobacillus babusae*, *Pediococcus cellicola*, *Lactobacillus cacaonum*, *Lactobacillus mucosae*, *Lactobacillus coleohominis*, *Lactobacillus gastricus*, *Lactobacillus equigenerosi*, *Lactobacillus malefermentans*, *Lactobacillus oryzae*, *Oenococcus oeni*, *Weissella kandleri*, and *Weissella koreensis*. In total 172 representative 6S RNA sequences covering 947 genomes remained. This set was used for further analyses.

For each genome without an annotated canonical 6S RNA (including those discarded manually in the first round), a second search iteration was performed with a LAB-specialized covariance model that was build based on all canonical 6S RNAs identified before. The e-value threshold was reduced to 0.1 and all search heuristics were turned off (`cmsearch -max`). In addition, the correct genomic locus was ensured by only allowing hits within 2000 nt from *upsA* and/or *rarA* homologs. Both are typically encoded in close vicinity to 6S RNA gene (see Results section “Synteny”). The homologs were annotated using `BLAST` (v2.8.1+) [66] with an e-value of 10^{-40} based on the sequences found in the synteny analysis. In this way, additional syntenically supported 6S RNA candidate genes were identified in 54 genomes. These are marked as “2nd-iteration” in Additional File 1 that lists all 6S RNAs annotated for LAB.

Prediction of rho-independent terminators

Terminators were predicted using `TransTermHP` (v2.09) [70]. An adaptive threshold was used to ascertain significant predictions. Each genome was shuffled ten times while preserving its mono- and di-symbol composition. We then compared the number of hits above any given threshold between the shuffled genomes and original genome. The threshold was chosen such that the average number of hits in the shuffled genomes was no more than 5% compared to the hits in the original genome. E.g. if we find 100 hits above a score of 90 in the genome, the average number of hits in the shuffled genomes above the same score cannot exceed 5, otherwise a higher threshold is chosen. In the absence of significance values provided by the prediction tool, this method roughly estimates a p-value threshold of 0.05 for terminator hits. Overlapping hits were merged. In additon, `RNIE` (v0.01) was used with default parameters for a genome-wide prediction [71]. For the relevant regions, the results were a subset of the former predictions.

Consensus secondary structure

All representative 6S RNA candidates were aligned using `mLocARNA` (v2.0.0RC8), a local structural alignment algorithm for RNA secondary structures [43]. To locate the putative start sites for pRNAs in LAB, three well-studied 6S RNA instances were added as references from which the start sites were then projected to the LAB 6S RNAs. Namely *Escherichia coli* K12 (GCF_000005845.2) and *Bacillus subtilis* 168 (GCF_000009045.1), which codes for two paralogs, 6S-1 and 6S-2 RNA (also known as BsrA and BsrB) [39, 72]. The consensus secondary structure was then calculated with `RNAalifold` (v2.4.13) [54] and visualized using `VARNA` (v.3.93) [73], excluding the folding references.

Prediction of pRNAs

The transcription start of 6S RNA-derived pRNAs was determined based on the structural alignment mentioned above. Based on previously characterized transcription start sites in other bacteria [26, 55, 74], we assumed the equivalent positions within LAB 6S RNAs. The putative pRNA sequences of 16 nt length were aligned with `Clustal Omega` (v1.2.1) [69]. We found a strong consensus sequence motif (see Results) that we used to further adjust the pRNA start site by shifting it for up to three nucleotides in case of suboptimal matches. The motif composition was calculated using `WebLogo` (v2.8.2) [75].

Phylogeny with secondary structure clusters

The sequences of the 6S RNA candidates identified in the first round were clustered hierarchically based on their structured RNA motifs using `RNAclust` [42]. This approach combines the base pair probability matrix of the secondary structure distributions (via `RNAfold` (v2.1.9) [54]) and a sequence-structure alignment based on `LocARNA` [43]. *Bacillus subtilis* 168 (GCF_000009045.1) 6S-1 RNA (BsrB) was added as an outgroup [39]. The resulting tree can be found in Additional File 2, while a condensed version is shown in Fig. 1, visualized using `Evolview` (v3) [76].

16S rRNA phylogeny

16S rRNA sequences were identified using BLAST (v2.8.1+) [66] with an e-value of 10^{-20} based on the 16S rRNA reference sequences provided by the NCBI database [38]. Redundant sequences were merged. Sequences were aligned using muscle (v3.8.1551) [77]. The 5'- and 3'-end of the 16S rRNA alignment were trimmed such that $<$ 25% of all sequences had remaining gaps in these regions. The phylogenetic reconstruction was performed with RAxML (v8.1.20) [78] using the General Time Reversible model (GTR) with optimization of substitution rates and the GAMMA model of rate heterogeneity and 1000 bootstrap iterations. The phylogenetic reconstruction was visualized using Evolview (v3) [76].

Synteny

The amino acid sequences of ten protein-coding genes 5000 nt up- and downstream of the predicted 6S RNA locus were fetched from the NCBI database. Orthologous groups were predicted with Proteinortho (v6.13) [45]. To avoid an overrepresentation bias, equivalent and similar 6S RNA sequences were represented by a single reference strain rather than all strains of the respective species (see "Detection of 6S RNAs"). Genes found in fewer than 50% of each family were omitted from the analysis. For each LAB family, one species that best represented the genomic context of all family members was chosen.

CcpA-binding catabolite responsive elements

The sequence motif for *cre* sites was derived from experimental *B. subtilis* data [36] that also fits previously derived *L. lactis* data [79] as shown in Fig. 6. However, we preferred the former as it yields a higher number of underlying sequences, which strengthens the derived p-values for motif matches and thus avoids false positive predictions. The 6S RNA sequences along with their 100 nt upstream regions were used to find sequences matching the *cre* motif using MAST [80]. Typically, this position overlapped with the 3'-end of the *rarA* gene. Hence, we did not expect binding sites further upstream to be relevant to 6S RNA. We used the dinucleotide distribution of the respective genomes as background for each e-value calculation. The default e-value threshold of 10 and p-value threshold of 10^{-5} was applied. The resulting motifs were separated in two groups: Upstream of the 6S RNA promoter and within the 6S RNA coding region as shown in Fig. 6.

Expression

Available RNA-Seq datasets for LAB were located in the NCBI SRA archive and downloaded on 12-11-2018 [38]. In total 115 RNA-Seq libraries were analyzed covering 24 different LAB species. Read sequences were extracted using the NCBI-provided fastq-dump (v2.8.2). Adapter removal and read trimming was performed using cutadapt (v1.12) [81] followed by a quality control with fastqc (v0.11.5) [82]. Processed reads were mapped to the respective genomes with segemehl (v0.2.0) [83]. An e-value threshold of 0.0001 was applied. The mapped data was visualized for each 6S RNA locus using custom scripts. Additional File 6 shows all results and data sources in detail.

Abbreviations

GRAS: Generally Recognized as Safe; LAB: Lactic acid bacteria; RNA: ribonucleic acid; RNAP: DNA-depended RNA polymerase complex; cre site: ccpA-binding catabolite responsive element

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12863-021-00983-2.

Additional file 1: List of genomes and 6S RNAs (xls). List of LAB genomes used in this study including tax annotation, assembly status, location of the predicted 6S RNA.

Additional file 2: Full 6S RNA phylogeny (pdf). Sequence- and structure-based reconstruction of 6S RNA phylogeny in LAB including the annotation of species with located *cre* sites. Full taxonomic resolution of Fig. 1.

Additional file 3: 16S rRNA phylogeny (pdf). Phylogenetic reconstruction of LAB 16S rRNA.

Additional file 4: Full genomic context of 6S RNA in LAB (pdf). Full genomic context of 6S RNA in LAB. Full taxonomic resolution of Fig. 2.

Additional file 5: 6S RNA grouped consensus alignment (pdf). Folded consensus structure of the 6S RNA groups analogous to Fig. 3.

Additional file 6: RNA-Seq results (pdf). Visualization of RNA-Seq libraries mapped to the respective 6S RNA loci.

Additional file 7: Predicted *cre* site motifs (xls). Predicted *cre* sites sequences and positions relative to the 6S RNA start site.

Additional file 8: 6S RNA evidence in *Weissella* (pdf). RNA-Seq data, genomic context and sequences of putative 6S RNA loci in *Weissella*.

Acknowledgments

We thank Florian Taube for implementing the RNA-Seq visualizations.

Author's contributions

ML conceived the study. PGC, PK, MT and ML carried out the bioinformatic analyses. PGC, PK, RKH, LS, EMH and ML wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by Deutscher Akademischer Austauschdienst (Short-Term Research Grant) and Deutsche Forschungsgemeinschaft (RTG 2355). Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

Genomes of lactic acid bacteria were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/ (2019-08-07). Species names, chromosome and tax ids, fasta paths and annotated 6S RNAs are provided in Additional File 1. RNA-Seq data was retrieved from NCBI SRA at https://www.ncbi.nlm.nih.gov/sra. Bioproject and SRA ids are listed in Additional Files 6 and 8. The 6S/SsrS RNA family seed sequences RF00013 provided by RFAM at http://rfam.xfam.org/family/RF00013 (Version 14) and the sequences by Wehner et al. provied at https://www.rna.uni-jena.de/supplements/6SRNA/ (2019-08-07) were used for initial 6S RNA prediction.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Centro de Referencia para Lactobacilos (CERELA-CONICET), Chacabuco 145, 4000 San Miguel de Tucumán, Argentina. [2]Philipps-Universität Marburg, Institut für Pharmazeutische Chemie, Marbacher Weg 6, 35032 Marburg, Germany. [3]Philipps-Universität Marburg, Center for Synthetic Microbiology (Synmikro), Hans-Meerwein-Straße 6, 35043 Marburg, Germany.

Received: 21 June 2021 Accepted: 12 August 2021
Published online: 03 September 2021

References

1. Duar RM, Lin XB, Zheng J, Martino ME, Grenier T, Pérez-Muñoz ME, Leulier F, Gänzle M, Walter J. Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. FEMS Microbiol Rev. 2017;41(Supp_1): 27–48. https://doi.org/10.1093/femsre/fux030.
2. Zheng J, Wittouck S, Salvetti E, Franz CMAP, Harris HMB, Mattarelli P, O'Toole PW, Pot B, Vandamme P, Walter J, Watanabe K, Wuyts S, Felis GE, Gänzle MG, Lebeer S. A taxonomic note on the genus *Lactobacillus*: Description of 23 novel genera, emended description of the genus *Lactobacillus* Beijerinck 1901, and union of *Lactobacillaceae* and *Leuconostocaceae*. Int J Syst Evol Microbiol. 2020;70(4):2782–858. https://doi.org/10.1099/ijsem.0.004107.
3. Salvetti E, Harris HMB, Felis GE, O extquoterightToole PW. Comparative genomics of the genus *Lactobacillus* reveals robust phylogroups that provide the basis for reclassification. Appl Environ Microbiol. 2018;84(17):. https://doi.org/10.1128/AEM.00993-18. https://aem.asm.org/content/84/17/e00993-18.full.pdf.
4. Leroy F, De Vuyst L. Lactic acid bacteria as functional starter cultures for the food fermentation industry. Trends Food Sci Technol. 2004;15(2): 67–78. https://doi.org/10.1016/j.tifs.2003.09.004.
5. Hill C, Guarner F, Reid G, Gibson GR, Merenstein DJ, Pot B, Morelli L, Canani RB, Flint HJ, Salminen S, Calder PC, Sanders ME. The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. Nat Rev Gastroenterol Hepatol. 2014;11(8):506–514. https://doi.org/10.1038/nrgastro.2014.66.
6. Mattila-Sandholm T, Mättö J, Saarela M. Lactic acid bacteria with health claims—interactions and interference with gastrointestinal flora. Int Dairy J. 1999;9(1):25–35. https://doi.org/10.1016/S0958-6946(99)00041-2.
7. Smid EJ, Hugenholtz J. Functional genomics for food fermentation processes. Ann Rev Food Sci Technol. 2010;1:497–519. https://doi.org/10.1146/annurev.food.102308.124143.
8. Zhang Y, Li Y. Engineering the antioxidative properties of lactic acid bacteria for improving its robustness. Curr Opin Biotechnol. 2013;24(2): 142–7. https://doi.org/10.1016/j.copbio.2012.08.013.
9. Papadimitriou K, Alegría Á, Bron PA, de Angelis M, Gobbetti M, Kleerebezem M, Lemos JA, Linares DM, Ross P, Stanton C, Turroni F, van Sinderen D, Varmanen P, Ventura M, Zúñiga M, Tsakalidou E, Kok J. Stress physiology of lactic acid bacteria. Microbiol Mol Biol Rev. 2016;80(3):837–90. https://doi.org/10.1128/MMBR.00076-15. https://mmbr.asm.org/content/80/3/837.full.pdf.
10. Gottesman S. Micros for microbes: non-coding regulatory RNAs in bacteria. Trends Genet. 2005;21(7):399–404. https://doi.org/10.1016/j.tig.2005.05.008.
11. Holmqvist E, Wagner EGH. Impact of bacterial sRNAs in stress responses. Biochem Soc Trans. 2017;45(6):1203–12. https://doi.org/10.1042/BST20160363.
12. Kok J, van Gijtenbeek LA, de Jong A, van der Meulen SB, Solopova A, Kuipers OP. The evolution of gene regulation research in *Lactococcus lactis*,. FEMS Microbiol Rev. 2017;41(Supp_1):220–43. https://doi.org/10.1093/femsre/fux028.
13. Wassarman KM, Storz G. 6S RNA regulates E. coli RNA polymerase activity. Cell. 2000;101(6):613–23.
14. Hsu L, Zagorski J, Wang Z, Fournier M. Escherichia coli 6S RNA gene is part of a dual-function transcription unit. J Bacteriol. 1985;161(3):1162–70.
15. Barrick JE, Sudarsan N, Weinberg Z, Ruzzo WL, Breaker RR. 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. RNA (New York, N.Y.) 2005;11(5):774–84. https://doi.org/10.1261/rna.7286705.
16. Wehner S, Damm K, Hartmann RK, Marz M. Dissemination of 6S RNA among bacteria. RNA Biol. 2014;11(11):1467–78. https://doi.org/10.4161/rna.29894.
17. Lechner M, Nickel AI, Wehner S, Riege K, Wieseke N, Beckmann BM, Hartmann RK, Marz M. Genomewide comparison and novel ncrnas of aquificales. BMC Genom. 2014;15:522. https://doi.org/10.1186/1471-2164-15-522.
18. Elkina D, Weber L, Lechner M, Burenina O, Weisert A, Kubareva E, Hartmann RK, Klug G. 6S RNA in *Rhodobacter sphaeroides*: 6S RNA and pRNA transcript levels peak in late exponential phase and gene deletion causes a high salt stress phenotype. RNA Biol. 2017;14(11):1627–37. https://doi.org/10.1080/15476286.2017.1342933.
19. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature. 2010;464(7286):250–5. https://doi.org/10.1038/nature08756.
20. Wassarman KM. 6S RNA: a small RNA regulator of transcription. Curr Opin Microbiol. 2007;10(2):164–8. https://doi.org/10.1016/j.mib.2007.03.008. Cell regulation (RNA special issue).
21. Steuten B, Hoch PG, Damm K, Schneider S, Köhler K, Wagner R, Hartmann RK. Regulation of transcription by 6S RNAs. RNA Biol. 2014;11(5):508–21. https://doi.org/10.4161/rna.28827.
22. Chen J, Wassarman KM, Feng S, Leon K, Feklistov A, Winkelman JT, Li Z, Walz T, Campbell EA, Darst SA. 6S RNA mimics b-form dna to regulate *Escherichia coli* RNA polymerase. Mol Cell. 2017;68(2):388–3976. https://doi.org/10.1016/j.molcel.2017.09.006.
23. Wassarman KM. 6S RNA, a global regulator of transcription. Microbiol Spectr. 2018;6(3):. https://doi.org/10.1128/microbiolspec.RWR-0019-2018.
24. Cavanagh AT, Klocko AD, Liu X, Wassarman KM. Promoter specificity for 6S RNA regulation of transcription is determined by core promoter sequences and competition for region 4.2 of sigma70. Mol Microbiol. 2008;67(6):1242–56. https://doi.org/10.1111/j.1365-2958.2008.06117.x.
25. Steuten B, Setny P, Zacharias M, Wagner R. Mapping the spatial neighborhood of the regulatory 6S RNA bound to *Escherichia coli* RNA polymerase holoenzyme. J Mol Biol. 2013;425(19):3649–61. https://doi.org/10.1016/j.jmb.2013.07.008.
26. Beckmann BM, Hoch PG, Marz M, Willkomm DK, Salas M, Hartmann RK. A pRNA-induced structural rearrangement triggers 6S-1 RNA release from RNA polymerase in *Bacillus subtilis*. EMBO J. 2012;31(7):1727–38. https://doi.org/10.1038/emboj.2012.23.
27. Panchapakesan SSS, Unrau PJ. *E. coli* 6S RNA release from RNA polymerase requires σ70 ejection by scrunching and is orchestrated by a conserved RNA hairpin. RNA (New York, N.Y.) 2012;18(12):2251–9. https://doi.org/10.1261/rna.034785.112.
28. Willkomm DK, Hartmann RK. 6S RNA - an ancient regulator of bacterial RNA polymerase rediscovered. Biol Chem. 2005;386(12):1273–77. https://doi.org/10.1515/BC.2005.144.
29. Cavanagh AT, Sperger JM, Wassarman KM. Regulation of 6S RNA by pRNA synthesis is required for efficient recovery from stationary phase in *E. coli* and *B. subtilis*. Nucleic Acids Res. 2012;40(5):2234–46.
30. Beckmann BM, Burenina OY, Hoch PG, Kubareva EA, Sharma CM, Hartmann RK. In vivo and in vitro analysis of 6S RNA-templated short transcripts in Bacillus subtilis. RNA Biol. 2011;8(5):839–49.
31. Cavanagh AT, Wassarman KM. 6S-1 RNA function leads to a delay in sporulation in *Bacillus subtilis*. J Bacteriol. 2013;195(9):2079–86.
32. Zheng H, Liu E, Shi T, Ye L, Konno T, Oda M, Ji Z-S. Strand-specific RNA-seq analysis of the *Lactobacillus delbrueckii subsp. bulgaricus* transcriptome. Mol bioSyst. 2016;12(2):508–19. https://doi.org/10.1039/c5mb00547g.
33. Le Rhun A, Beer YY, Reimegård J, Chylinski K, Charpentier E. RNA sequencing uncovers antisense RNAs and novel small RNAs in *Streptococcus pyogenes*. RNA Biol. 2016;13(2):177–95.

34. van der Meulen SB, de Jong A, Kok J. Transcriptome landscape of *Lactococcus lactis* reveals many novel RNAs including a small regulatory RNA involved in carbon uptake and metabolism. RNA Biol. 2016;13(3): 353–66. https://doi.org/10.1080/15476286.2016.1146855.
35. Schumacher MA, Sprehe M, Bartholomae M, Hillen W, Brennan RG. Structures of carbon catabolite protein a–(hpr-ser46-p) bound to diverse catabolite response element sites reveal the basis for high-affinity binding to degenerate dna operators. Nucleic Acids Res. 2011;39(7):2931–42.
36. Marciniak BC, Pabijaniak M, de Jong A, Duhring R, Seidel G, Hillen W, Kuipers OP. High- and low-affinity cre boxes for ccpa binding in *Bacillus subtilis* revealed by genome-wide analysis. BMC Genom. 2012;13(1):401. https://doi.org/10.1186/1471-2164-13-401.
37. Fouquier d'Hérouel A, Wessner F, Halpern D, Ly-Vu J, Kennedy SP, Serror P, Aurell E, Repoila F. A simple and efficient method to search for selected primary transcripts: non-coding and antisense RNAs in the human pathogen *Enterococcus faecalis*. Nucleic Acids Res. 2011;39(7):46. https://doi.org/10.1093/nar/gkr012.
38. Sayers EW, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2009;37(Database issue):5–15. https://doi.org/10.1093/nar/gkn741.
39. Burenina OY, Hoch PG, Damm K, Salas M, Zatsepin TS, Lechner M, Oretskaya TS, Kubareva EA, Hartmann RK. Mechanistic comparison of *Bacillus subtilis* 6S-1 and 6S-2 RNAs–commonalities and differences. RNA (New York, N.Y.) 2014;20(3):348–59. https://doi.org/10.1261/rna.042077.113.
40. Qi J, Zhang D, Wang S, Huang L, Xia L, Dong W, Zheng Q, Liu Q, Xiao J, Xu Z. Transcriptome analysis of xylo-oligosaccharides utilization systems in Weissella confusa xu1. AMS. 2020;60(5):912–23.
41. Jeong SE, Chun BH, Kim KH, Park D, Roh SW, Lee SH, Jeon CO. Genomic and metatranscriptomic analyses of Weissella koreensis reveal its metabolic and fermentative features during kimchi fermentation. Food Microbiol. 2018;76:1–10.
42. Engelhardt J, Heyne S, Will S, Reiche R. RNAclust: A Tool for Clustering of RNAs Based on Their Secondary Structures Using LocARNA. http://www.bioinf.uni-leipzig.de. Accessed 03 Aug 2018.
43. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol. 2007;3(4):65. https://doi.org/10.1371/journal.pcbi.0030065.
44. Holzapfel WH, Haberer P, Geisen R, Björkroth J, Schillinger U. Taxonomy and important features of probiotic microorganisms in food and nutrition. Am J Clin Nutr. 2001;73(2):365–73. https://doi.org/10.1093/ajcn/73.2.365s.
45. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011;12:124.
46. Barre F-X, Søballe B, Michel B, Aroyo M, Robertson M, Sherratt D. Circles: the replication-recombination-chromosome segregation connection. Proc Natl Acad Sci. 2001;98(15):8189–95.
47. Stanage TH, Page AN, Cox MM. Dna flap creation by the RarA/MgsA protein of Escherichia coli. Nucleic Acids Res. 2017;45(5):2724–35.
48. Carrasco B, Seco EM, López-Sanz M, Alonso JC, Ayora S. Bacillus subtilis RarA modulates replication restart. Nucleic Acids Res. 2018;46(14): 7206–20.
49. Liu W-T, Karavolos MH, Bulmer DM, Allaoui A, Hormaeche RDCE, Lee JJ, Khan CA. Role of the universal stress protein UspA of *Salmonella* in growth arrest, stress and virulence. Microb Pathog. 2007;42(1):2–10.
50. Gustavsson N, Diez A, Nyström T. The universal stress protein paralogues of *Escherichia coli* are co-ordinately regulated and co-operate in the defence against DNA damage. Mol Microbiol. 2002;43(1):107–17.
51. Kvint K, Nachin L, Diez A, Nyström T. The bacterial universal stress protein: Function and regulation. Curr Opin Microbiol. 2003;6:140–5. https://doi.org/10.1016/S1369-5274(03)00025-0.
52. Huang G, Li C, Cao Y. Proteomic analysis of differentially expressed proteins in *Lactobacillus brevis* ncl912 under acid stress. FEMS Microbiol Lett. 2011;318(2):177–82. https://doi.org/10.1111/j.1574-6968.2011.02257.x.
53. Kaur G, Ali SA, Kumar S, Mohanty AK, Behare P. Label-free quantitative proteomic analysis of *Lactobacillus fermentum* ncdc 400 during bile salt exposure. J Proteomics. 2017;167:36–45. https://doi.org/10.1016/j.jprot.2017.08.008.
54. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA Package 2.0. Algorithms Mol Biol AMB. 2011;6:26. https://doi.org/10.1186/1748-7188-6-26.
55. Hoch PG, Schlereth J, Lechner M, Hartmann RK. *Bacillus subtilis* 6S-2 RNA serves as a template for short transcripts *in vivo*. RNA (New York, N.Y.) 2016;22(4):614–22. https://doi.org/10.1261/rna.055616.115.
56. Donner J, Reck M, Bergmann S, Kirschning A, Müller R, Wagner-Döbler I. The biofilm inhibitor Carolacton inhibits planktonic growth of virulent pneumococci via a conserved target. Sci Rep. 2016;6(1):1–15.
57. Lécrivain A-L, Le Rhun A, Renault TT, Ahmed-Begrich R, Hahnke K, Charpentier E. In vivo 3′-to-5′ exoribonuclease targetomes of Streptococcus pyogenes. Proc Natl Acad Sci. 2018;115(46):11814–9.
58. Muscariello L, Marasco R, De Felice M, Sacco M. The functional ccpa gene is required for carbon catabolite repression in *Lactobacillus plantarum*. Appl Environ Microbiol. 2001;67(7):2903–7. https://doi.org/10.1128/AEM.67.7.2903-2907.2001.
59. Giaretta S, Treu L, Vendramin V, da Silva Duarte V, Tarrah A, Campanaro S, Corich V, Giacomini A. Comparative transcriptomic analysis of *Streptococcus thermophilus* th1436 and th1477 showing different capability in the use of galactose. Front Microbiol. 2018;9:1765. https://doi.org/10.3389/fmicb.2018.01765.
60. Grand M, Aubourg M, Pikis A, Thompson J, Deutscher J, Hartke A, Sauvageot N. Characterization of the gen locus involved in b-1,6-oligosaccharide utilization by *Enterococcus faecalis*. Mol Microbiol. 2019;112(6):1744–56. https://doi.org/10.1111/mmi.14390.
61. Kim H-M, Waters A, Turner ME, Rice KC, Ahn S-J. Regulation of cid and lrg expression by ccpa in *Streptococcus mutans*. Microbiology (Reading, England). 2019;165(1):113–23. https://doi.org/10.1099/mic.0.000744.
62. Kim J-H, Yang Y-K, Chambliss GH. Evidence that Bacillus catabolite control protein CcpA interacts with RNA polymerase to inhibit transcription. Mol Microbiol. 2005;56(1):155–62.
63. Ogaugwu CE, Cheng Q, Fieck A, Hurwitz I, Durvasula R. Characterization of a *Lactococcus lactis* promoter for heterologous protein production. Biotechnol Rep. 2018;17:86–92. https://doi.org/10.1016/j.btre.2017.11.010.
64. Oberto J. Synttax: a web server linking synteny to prokaryotic taxonomy. BMC Bioinformatics. 2013;14:4. https://doi.org/10.1186/1471-2105-14-4.
65. Hernández-Tamayo R, Graumann PL. Bacillus subtilis RarA forms damage-inducible foci that scan the entire cell. BMC Res Notes. 2019;12(1):1–3.
66. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. Blast+: architecture and applications. BMC Bioinformatics. 2009;10(1):421.
67. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, Bateman A, Finn RD, Petrov AI. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res. 2018;46(D1):335–42. https://doi.org/10.1093/nar/gkx1038.
68. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics (Oxford, England). 2013;29(22):2933–5. https://doi.org/10.1093/bioinformatics/btt509.
69. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol. 2011;7:539. https://doi.org/10.1038/msb.2011.75.
70. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. 2007;8(2):1–12.
71. Gardner PP, Barquist L, Bateman A, Nawrocki EP, Weinberg Z. RNIE: genome-wide prediction of bacterial intrinsic terminators. Nucleic Acids Res. 2011;39(14):5845–52.
72. Brownlee GG. Sequence of 6S RNA of *E. coli*. Nat New Biol. 1971;229(5): 147–9.
73. Darty K, Denise A, Ponty Y. Varna: Interactive drawing and editing of the RNA secondary structure. Bioinformatics (Oxford, England). 2009;25(15): 1974–5. https://doi.org/10.1093/bioinformatics/btp250.
74. Wurm R, Neusser T, Wagner R. 6S RNA-dependent inhibition of RNA polymerase is released by RNA-dependent synthesis of small *de novo*

products. Biol Chem. 2010;391(2-3):187–96. https://doi.org/10.1515/BC.2010.018.

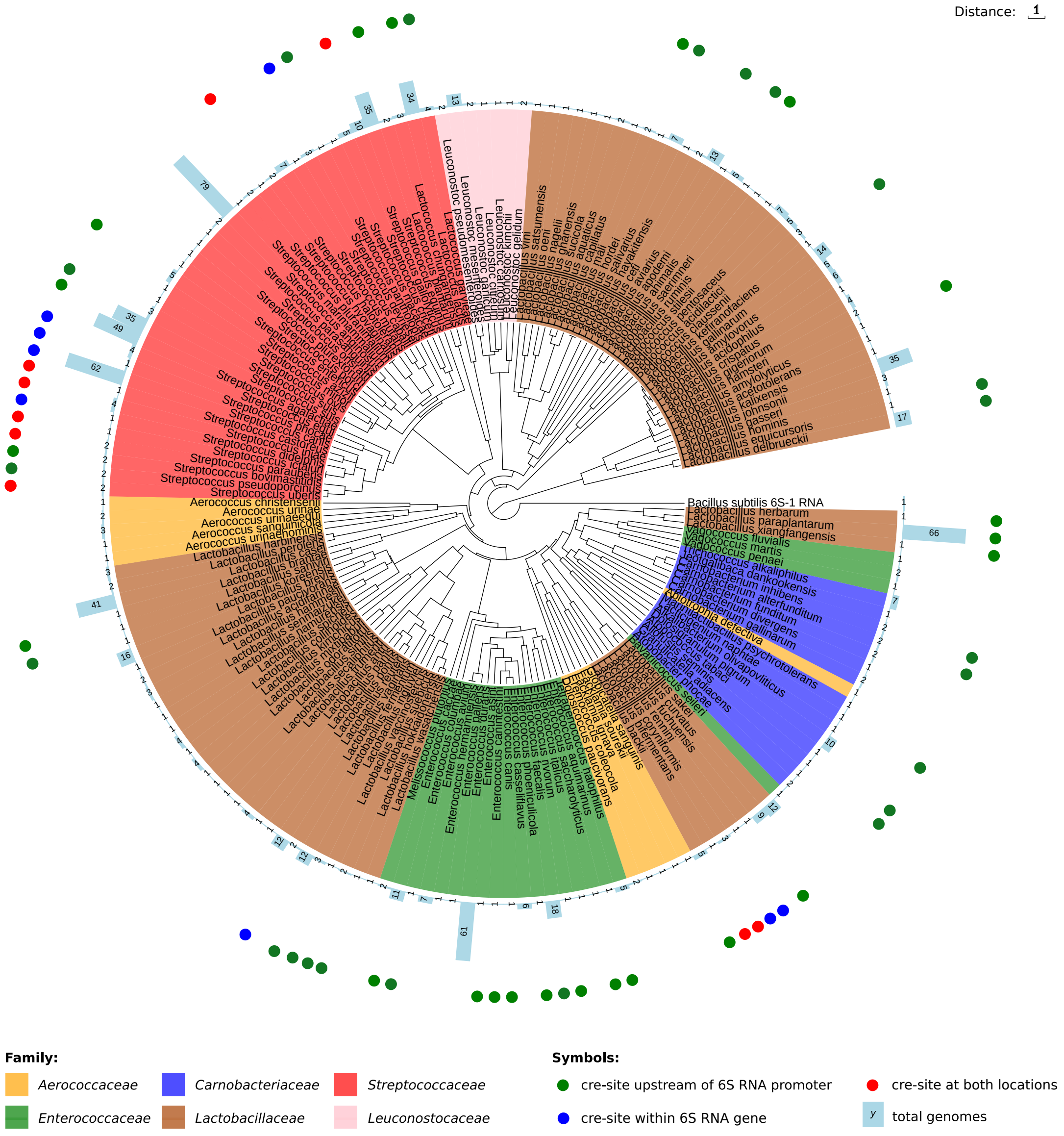
75. Crooks GE, Hon G, Chandonia J-M, Brenner SE. Weblogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90. https://doi.org/10.1101/gr.849004.
76. Subramanian B, Gao S, Lercher MJ, Hu S, Chen W-H. Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees. Nucleic Acids Res. 2019;47(W1):270–5. https://doi.org/10.1093/nar/gkz357.
77. Edgar RC. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
78. Stamatakis A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England). 2014;30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033.
79. Zomer AL, Buist G, Larsen R, Kok J, Kuipers OP. Time-resolved determination of the ccpa regulon of *Lactococcus lactis subsp. cremoris* mg1363. J Bacteriol. 2007;189(4):1366–81. https://doi.org/10.1128/JB.01013-06.
80. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics (Oxford, England). 1998;14(1):48–54.
81. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–12. https://doi.org/10.14806/ej.17.1.200.
82. Andrews S. FastQC A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 12 Dec 2017.
83. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009;5(9):1000502. https://doi.org/10.1371/journal.pcbi.1000502.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

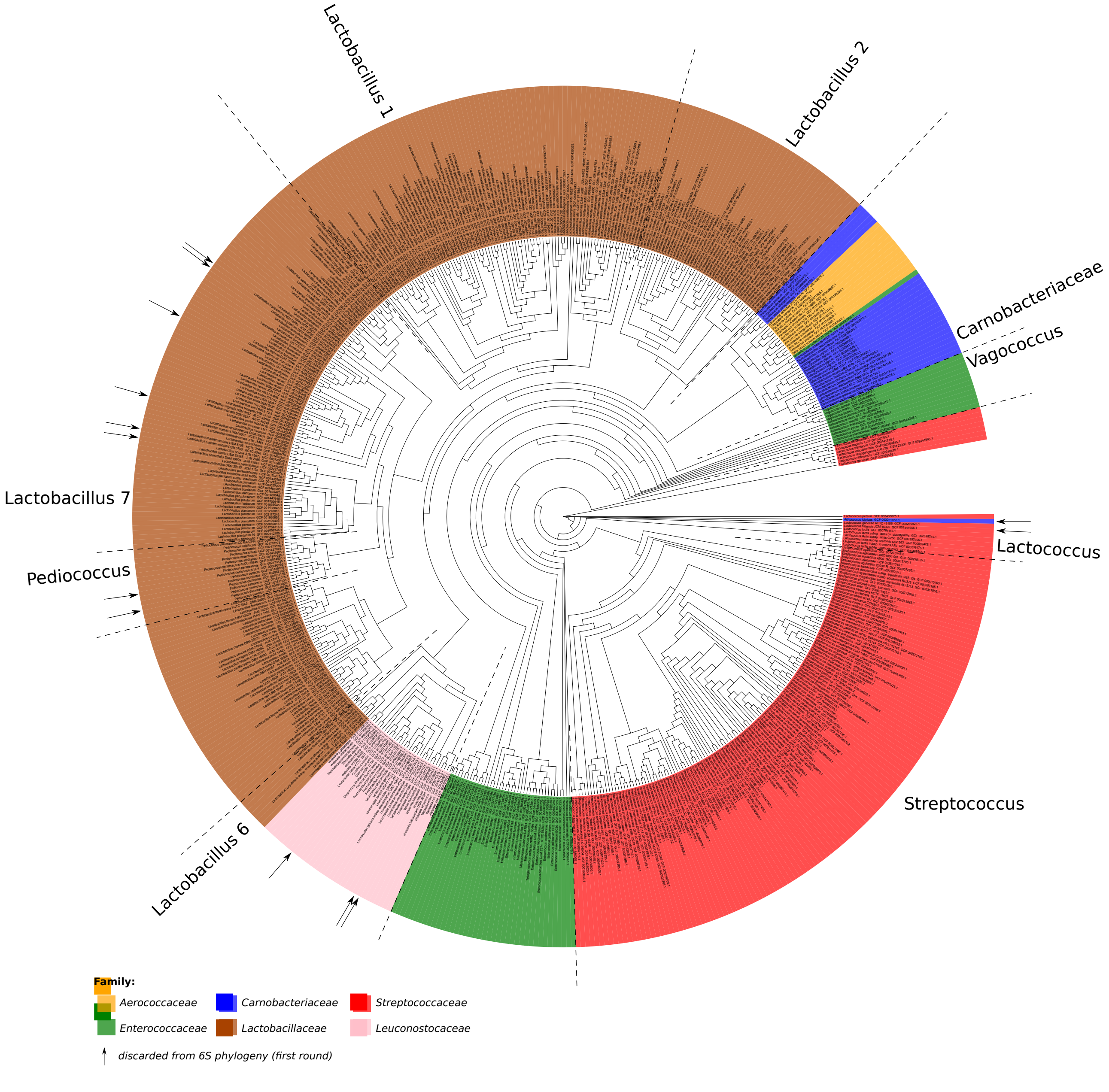
Additional File 2 — Full 6S RNA phylogeny

Sequence- and structure-based reconstruction of 6S RNA phylogeny in LAB. Canonical 6S RNAs were clustered hierarchically using `RNAclust` and `mlocarna`. Family membership is indicated by color. 6S-1 RNA from *B. subtilis* is used as outgroup. The full number of represented genomes is indicated in blue boxes in the outer ring. Circles in the outer ring indicate whether and where a potential cre-site were identified at the 6S RNA locus.



Additional File 3 — 16S rRNA phylogeny

Supplemental Figure 2: Sequence-based reconstruction of 16S rRNA phylogeny in LAB. The phylogenetic reconstruction was performed with `RAxML` using the GTR model with an optimization of substitution rates and the GAMMA model of rate heterogeneity. 1000 bootstrap iterations.

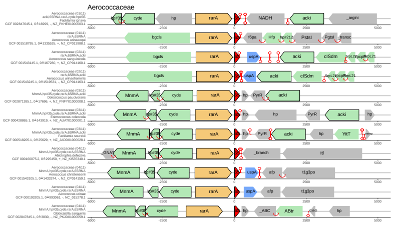


Additional File 4 — Full genomic context of 6S RNA in LAB

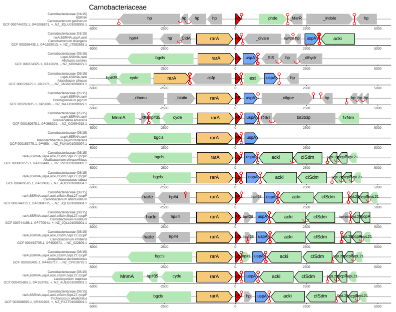
For each family the genomic context around the 6S RNA (±5000nt) is shown. `Proteinortho` was used to group the protein-coding genes. The displayed names represent the unique ortholog groups (# is used for groups with same name). If no ortholog were found the gene is marked with a underscore prefix and is thus excluded from the analysis. Genes with a solid border were found in ≥ 50% of the respective family. Genes found in multiple families are colored. *rarA* and *uspA* are colored orange and blue respectively. Putative Rho-independent terminators are indicated by red hexagons. Genes in close proximity (<20 nt) are indicated by a semicircle connecting them. These could be part of a polycistronic transcript.

Each genomic context contains 5 lines of information to the left:
(1) family name and conservation score
(2) the group of conserved gene order
(3) species name
(4) species id
(5) chromosome name

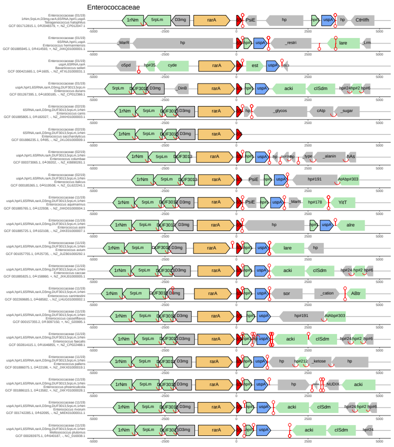
The given conservation score describes the ratio of species containing this group of conserved gene order over the corresponding family.



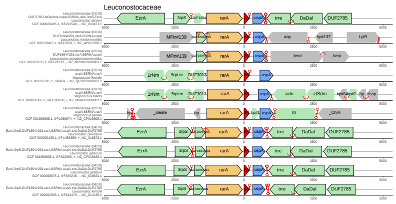
abbreviation	full description
5rpL21	50S ribosomal protein L21.
5rpL27	50S ribosomal protein L27.
_ABC	ABC transporter permease.
ABtr	ABC transporter.
acki	acetate kinase.
afp	aquaporin family protein.
_argini	arginine–tRNA ligase.
bgcls	bifunctional glutamate–cysteine ligase/glutathione synthetase.
_branch	branched-chain amino acid transport system II carrier protein.
cISdm	class I SAM-dependent methyltransferase.
_CsbD	CsbD family protein.
cyde	cysteine desulfurase.
f6pa	fructose-6-phosphate aldolase.
_GNAT	GNAT family N-acetyltransferase.
Hfp	HAD family phosphatase.
hp#212	hypothetical protein (orthology group #212).
hp#35	hypothetical protein (orthology group #35).
hp	hypothetical protein.
itl	isoleucine–tRNA ligase.
MnmA	tRNA 2-thiouridine(34) synthase MnmA.
NADH	NADH oxidase.
PgtsI	PTS glucitol/sorbitol transporter subunit IIC.
PstsI	PTS sorbitol transporter subunit IIB.
PyrR	bifunctional pyr operon transcriptional regulator/uracil phosphoribosyltransferase PyrR.
rarA	replication-associated recombination protein A.
rpcpP	ribosomal-processing cysteine protease Prp.
t1g3po	type 1 glycerol-3-phosphate oxidase.
_transc	transcriptional regulator.
_transp	transposase.
uspA	universal stress protein.
YitT	YitT family protein.



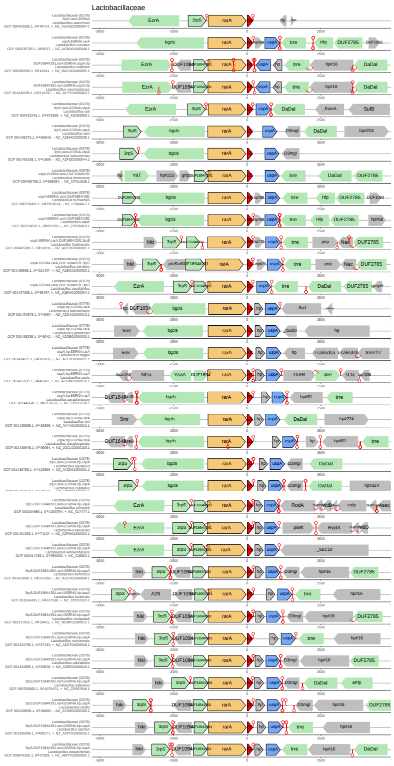
abbreviation	full description
1rNm	16S rRNA (uracil(1498)-N(3))-methyltransferase.
5rpL21	50S ribosomal protein L21.
5rpL27	50S ribosomal protein L27.
acki	acetate kinase.
asfp	alanine:cation symporter family protein.
bgcls	bifunctional glutamate–cysteine ligase/glutathione synthetase.
_biotin	biotin–[acetyl-CoA-carboxylase] ligase.
bs3b3p	bifunctional (p)ppGpp synthetase/guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase.
_CidA	CidA/LrgA family protein.
cISdm	class I SAM-dependent methyltransferase.
cyde	cysteine desulfurase.
_dihydr	dihydrodipicolinate synthase family protein.
_divale	divalent metal cation transporter.
Dttd	D-tyrosyl-tRNA(Tyr) deacylase.
est	esterase.
hade	haloacid dehalogenase.
hp#169	hypothetical protein (orthology group #169).
hp#1	hypothetical protein (orthology group #1).
hp#28	hypothetical protein (orthology group #28).
hp#35	hypothetical protein (orthology group #35).
hp#4	hypothetical protein (orthology group #4).
hp	hypothetical protein.
_indole	indole-3-pyruvate decarboxylase.
_MarR	MarR family transcriptional regulator.
MnmA	tRNA 2-thiouridine(34) synthase MnmA.
_oligoe	oligoendopeptidase.
phde	phosphoglycerate dehydrogenase.
rarA	replication-associated recombination protein A.
_ribonu	ribonuclease J.
rpcpP	ribosomal-processing cysteine protease Prp.
_SIS	SIS domain-containing protein.
uspA	universal stress protein.
_XRE	XRE family transcriptional regulator.

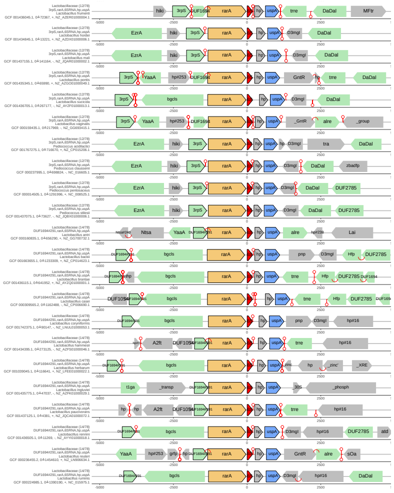


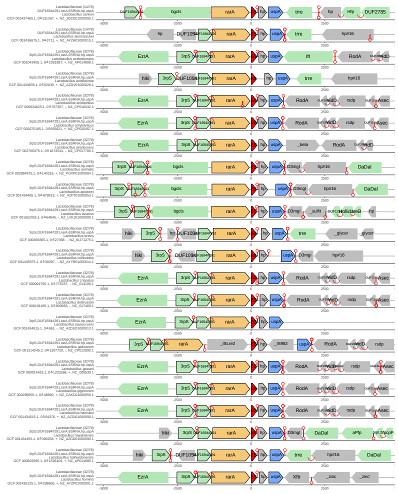
abbreviation	full description
1rNm	16S rRNA (uracil(1498)-N(3))-methyltransferase.
5rpLm	50S ribosomal protein L11 methyltransferase.
ABtr	ABC transporter.
acki	acetate kinase.
_alanin	alanine racemase.
alre	aldo/keto reductase.
AtAbp#303	ABC transporter ATP-binding protein (orthology group #303).
_cation	cation transporter.
cAtp	carbohydrate ABC transporter permease.
cISdm	class I SAM-dependent methyltransferase.
CtHIfh	Cof-type HAD-IIB family hydrolase.
cyde	cysteine desulfurase.
D3mg	DNA-3-methyladenine glycosylase.
_DinB	DinB family protein.
_DUF292	DUF2922 domain-containing protein.
DUF3013	DUF3013 domain-containing protein.
est	esterase.
_glycos	glycoside hydrolase family 127 protein.
hAs	holo-ACP synthase.
hp#178	hypothetical protein (orthology group #178).
hp#191	hypothetical protein (orthology group #191).
hp#1	hypothetical protein (orthology group #1).
hp#212	hypothetical protein (orthology group #212).
hp#24	hypothetical protein (orthology group #24).
hp#2	hypothetical protein (orthology group #2).
hp#35	hypothetical protein (orthology group #35).
hp#6	hypothetical protein (orthology group #6).
hp	hypothetical protein.
_ISL3	ISL3 family transposase.
_ketose	ketose-bisphosphate aldolase.
lare	lactaldehyde reductase.
_Lrm	L-rhamnose mutarotase.
_MarR	MarR family transcriptional regulator.
_NUDIX	NUDIX domain-containing protein.
o5pd	orotidine-5'-phosphate decarboxylase.
_PemK	PemK family transcriptional regulator.
PsiE	phosphate-starvation-inducible protein PsiE.
rarA	replication-associated recombination protein A.
_restri	restriction endonuclease.
sor	sortase.
_sugar	sugar ABC transporter permease.
_type	type II toxin-antitoxin system PemK/MazF family toxin.
uspA	universal stress protein.
YitT	YitT family protein.



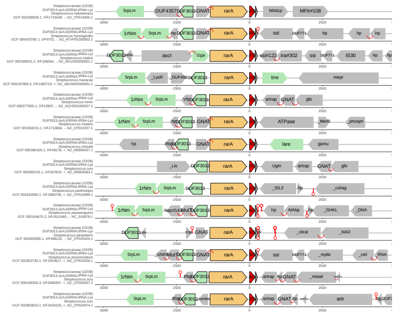
abbreviation	full description
1rNm	16S rRNA (uracil(1498)-N(3))-methyltransferase.
3rpS	30S ribosomal protein S4.
5rpLm	50S ribosomal protein L11 methyltransferase.
aap	amino acid permease.
acki	acetate kinase.
_beta'	beta-galactosidase.
_beta	beta-galactosidase small subunit.
_ChrA	ChrA protein.
cISdm	class I SAM-dependent methyltransferase.
DaDal	D-alanine–D-alanine ligase.
DUF1694#291	DUF1694 domain-containing protein (orthology group #291).
DUF1694	DUF1694 domain-containing protein.
DUF2785	DUF2785 domain-containing protein.
DUF3013	DUF3013 domain-containing protein.
EzrA	septation ring formation regulator EzrA.
hp#137	hypothetical protein (orthology group #137).
hp#1	hypothetical protein (orthology group #1).
hp#2	hypothetical protein (orthology group #2).
hp#54	hypothetical protein (orthology group #54).
hp	hypothetical protein.
Ift	IS5/IS1182 family transposase.
LytR	LytR family transcriptional regulator.
MFtr#139	MFS transporter (orthology group #139).
_oleate	oleate hydratase.
rarA	replication-associated recombination protein A.
tIssp	type II secretion system protein.
trre	transcriptional regulator.







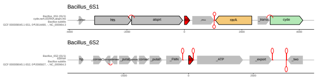
abbreviation	full description
_30S	30S ribosomal protein S14.
_30S	30S ribosomal protein S14.
3rpS	30S ribosomal protein S4.
5mr	5,10-methylenetetrahydrofolate reductase.
5rpL27	50S ribosomal protein L27.
_8od	8-oxoguanine deaminase.
A2ft	AI-2E family transporter.
alre	aldo/keto reductase.
aPfp	aminopeptidase P family protein.
ArsR	ArsR family transcriptional regulator.
Asec	ATP synthase epsilon chain.
atd	aminoacyl-tRNA deacylase.
_beta	beta-N-acetylhexosaminidase.
bgcls	bifunctional glutamate–cysteine ligase/glutathione synthetase.
D3mgI	DNA-3-methyladenine glycosylase I.
DaDal	D-alanine–D-alanine ligase.
DdRpsb	DNA-directed RNA polymerase subunit beta.
DUF1054	DUF1054 domain-containing protein.
DUF1093	DUF1093 domain-containing protein.
DUF1146	DUF1146 domain-containing protein.
DUF1292	DUF1292 domain-containing protein.
DUF1648	DUF1648 domain-containing protein.
DUF1694#291	DUF1694 domain-containing protein (orthology group #291).
DUF1694	DUF1694 domain-containing protein.
DUF2785	DUF2785 domain-containing protein.
DUF2969	DUF2969 domain-containing protein.
_EamA	EamA family transporter.
EzrA	septation ring formation regulator EzrA.
gcspH	glycine cleavage system protein H.
_glycer	glycerophosphodiester phosphodiesterase.
_glycer’	glycerophosphodiester phosphodiesterase family protein.
_GntR	GntR family transcriptional regulator.
GntR	GntR family transcriptional regulator.
grfp	glyoxalase/bleomycin resistance/dioxygenase family protein.
_group	group II intron reverse transcriptase/maturase.
Hfp	HAD family phosphatase.
hiki	histidine kinase.
Hollida	Holliday junction resolvase RuvX.
hp#16	hypothetical protein (orthology group #16).
hp#224	hypothetical protein (orthology group #224).
hp#238	hypothetical protein (orthology group #238).
hp#253	hypothetical protein (orthology group #253).
hp#68	hypothetical protein (orthology group #68).
hp#69	hypothetical protein (orthology group #69).
hp#73	hypothetical protein (orthology group #73).
hp#75	hypothetical protein (orthology group #75).
hp#81	hypothetical protein (orthology group #81).
hp	hypothetical protein.
Ift	IS5/IS1182 family transposase.
IreB	IreB family regulatory phosphoprotein.



abbreviation	full description
_replic	replication initiation protein.
_RNA	RNA-binding protein.
RseP	RIP metalloprotease RseP.
sap	surface-anchored protein.
_signal	signal peptidase I.
_site	site-specific integrase.
_SrtB	SrtB family sortase.
ssi#122	site-specific integrase (orthology group #122).
ssi	site-specific integrase.
_strept	streptomycin resistance protein.
_SulP	SulP family inorganic anion transporter.
t1ga	type 1 glutamine amidotransferase.
_TetR	TetR/AcrR family transcriptional regulator.
TIGR007	TIGR00730 family Rossman fold protein.
tra#302	transposase (orthology group #302).
_transc	transcriptional regulator.
_transc	transcription antiterminator.
_transp	transposase.
TrmB	tRNA (guanosine(46)-N7)-methyltransferase TrmB.
trre#124	transcriptional regulator (orthology group #124).
trre	transcriptional regulator.
_two	two-component sensor histidine kinase.
_UDP	UDP-glucose–hexose-1-phosphate uridylyltransferase.
Ugm	UDP-galactopyranose mutase.
Vfp	VOC family protein.
_XRE	XRE family transcriptional regulator.
YaaA	peroxide stress protein YaaA.
_Zdp	Zn-dependent protease.
_zinc	zinc transporter ZupT.

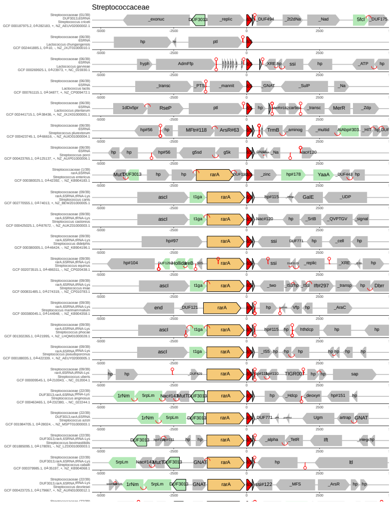
abbreviation	full description
_ISLre2	ISLre2 family transposase.
Lai	L-arabinose isomerase.
Lsalisdsa	L-serine ammonia-lyase, iron-sulfur-dependent, subunit alpha.
Lsalisdsb	L-serine ammonia-lyase, iron-sulfur-dependent, subunit beta.
MFtr	MFS transporter.
Nac	N-acetyltransferase.
Ntsa#240	NAD(P) transhydrogenase subunit alpha (orthology group #240).
Ntsa	NAD(P) transhydrogenase subunit alpha.
_phosph	phosphoenolpyruvate synthase.
pmSor	peptide-methionine (S)-S-oxide reductase.
pnp	phosphate–nucleotide phosphotransferase.
pseft	putative sulfate exporter family transporter.
rarA	replication-associated recombination protein A.
RodA	rod shape-determining protein RodA.
rpcpP	ribosomal-processing cysteine protease Prp.
rsdp	rod shape-determining protein.
_SEC10	SEC10/PgrA surface exclusion domain-containing protein.
sOa	sugar O-acetyltransferase.
SufB	Fe-S cluster assembly protein SufB.
_sulfit	sulfite exporter TauE/SafE family protein.
t1ga	type 1 glutamine amidotransferase.
_transp	transposase.
tra	transposase.
trre#27	transcriptional regulator (orthology group #27).
trre	transcriptional regulator.
uspA	universal stress protein.
Xftr	XRE family transcriptional regulator.
_XRE	XRE family transcriptional regulator.
YaaA	peroxide stress protein YaaA.
YidD	membrane protein insertion efficiency factor YidD.
YitT	YitT family protein.
zbadfp	zinc-binding alcohol dehydrogenase family protein.
_zinc	zinc ribbon domain-containing protein.
_zinc’	zinc ribbon domain-containing protein.

abbreviation	full description
1dDx5pr	1-deoxy-D-xylulose-5-phosphate reductoisomerase.
1rNm	16S rRNA (uracil(1498)-N(3))-methyltransferase.
_2t2dNa	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-acetyltransferase.
_50S	50S ribosomal protein L28.
5fcl	5-formyltetrahydrofolate cyclo-ligase.
5rpLm	50S ribosomal protein L11 methyltransferase.
AdmFfp	ATP-dependent metallopeptidase FtsH/Yme1/Tma family protein.
_alpha	alpha/beta hydrolase.
_aminog	aminoglycoside phosphotransferase.
_AraC	AraC family transcriptional regulator.
ArsR#63	ArsR family transcriptional regulator (orthology group #63).
_ArsR	ArsR family transcriptional regulator.
artr	anaerobic ribonucleoside-triphosphate reductase.
artrap	anaerobic ribonucleoside-triphosphate reductase activating protein.
ascI	aminodeoxychorismate synthase, component I.
AtAbp#303	ABC transporter ATP-binding protein (orthology group #303).
AtAbp	ABC transporter ATP-binding protein.
ATPase	heavy metal translocating P-type ATPase.
_ATP	ATP-binding protein.
_ATP	ATP-grasp domain-containing protein.
_carbox	carboxymuconolactone decarboxylase family protein.
_cell	cell division protein FtsK.
_cell	cell surface protein.
_citrat	citrate:sodium symporter.
_collag	collagen-binding protein.
_Cro	Cro/CI family transcriptional regulator.
csp	cold-shock protein.
Dbrr	DNA-binding response regulator.
_deoxyn	deoxynucleoside kinase.
_DNA	DNA-binding response regulator.
_DUF121	DUF1211 domain-containing protein.
DUF1292	DUF1292 domain-containing protein.
_DUF130	DUF1304 domain-containing protein.
_DUF175	DUF1751 domain-containing protein.
DUF3013	DUF3013 domain-containing protein.
_DUF302	DUF3021 domain-containing protein.
DUF3173	DUF3173 domain-containing protein.
_DUF429	DUF4298 domain-containing protein.
DUF4357	DUF4357 domain-containing protein.
_DUF443	DUF4430 domain-containing protein.
_DUF486	DUF4865 domain-containing protein.
_DUF494	DUF4947 domain-containing protein.
_DUF59	DUF59 domain-containing protein.
_DUF771	DUF771 domain-containing protein.
DUF771	DUF771 domain-containing protein.
end	endonuclease.
_exonuc	exonuclease SbcC.
FlaR	DNA topology modulation protein FlaR.
g5k	glutamate 5-kinase.
g5sd	glutamate-5-semialdehyde dehydrogenase.


Bacillus_6S1

Bacillus_6S2

abbreviation	full description
atsprt	aspartyl-tRNA synthetase, promiscuous (also recognizes tRNAasn).
_conser	conserved protein of unknown function.
cyde	cysteine desulfurase.
hts	histidyl-tRNA synthetase.
rarA	replication-associated recombination protein A.
_transc	transcriptional regulator of cysteine biosynthesis.
_tRNA	tRNA threonylcarbamoyladenosine dehydratase (t(6)A37 dehydratase).
_ATP	ATP-dependent helicase.
_conser”	conserved hypothetical phage protein.
_conser	conserved protein of unknown function.
_conser’	conserved protein of unknown function.
_conser”’	conserved protein of unknown function (mother cell in sporulation).
_conser””	conserved protein of unknown function (sporulation-related).
_export	exported cell wall lytic enzyme.
_FMN	FMN-dependent NADH-azoreductase.
hp	conserved hypothetical protein.
_putati’	putative chaperone.
_putati	putative general stress protein.
_two	two-component response regulator [DesK].

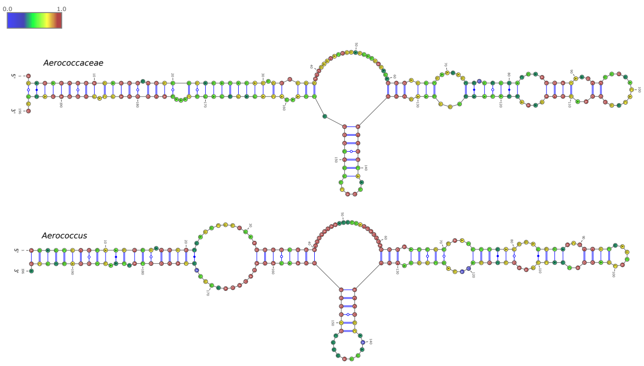

Streptococcaceae
Streptococcaceae (22/39)
DUF3013,rarA,6SRNA,tRNA-Lys
Streptococcus anginosus
GCF 000463465.1, 0≙232380, -, NC_022244.1
1rNm
5rpLm
Nac#143
MutT
DUF3013
rarA
hp
_Hdcp
_deoxyn
Streptococcus azizii
GCF 001984705.1, 0≙28024, -, NZ_MSPT01000003.1
_DUF771
Ugm
Streptococcus bovimastitidis
GCF 001885095.1, 0≙178091, -, NZ_LZDD01000003.1
hp#110
hp#111
_alpha
_TetR
Ift
Streptococcus caballi
GCF 000379985.1, 0≙35197, +, NZ_KB904068.1
GNAT
Streptococcus devriesei
GCF 000423725.1, 0≙179667, +, NZ_AUIN01000012.1
trre#124
hp#125
ssi#122
_MFS
_ArsR
ascI
t1ga
AdmFfp
RseP
hp#178
YaaA
-5000
-2500
0
2500

abbreviation	full description
hp#104	hypothetical protein (orthology group #104).
hp#110	hypothetical protein (orthology group #110).
hp#111	hypothetical protein (orthology group #111).
hp#115	hypothetical protein (orthology group #115).
hp#121	hypothetical protein (orthology group #121).
hp#125	hypothetical protein (orthology group #125).
hp#146	hypothetical protein (orthology group #146).
hp#151	hypothetical protein (orthology group #151).
hp#178	hypothetical protein (orthology group #178).
hp#56	hypothetical protein (orthology group #56).
hp#97	hypothetical protein (orthology group #97).
hthdcp	helix-turn-helix domain-containing protein.
hyph	hypoxanthine phosphoribosyltransferase.
Ift#297	IS3 family transposase (orthology group #297).
Ift	IS3 family transposase.
_integr	integrase.
IreB	IreB family regulatory phosphoprotein.
_IS30	IS30 family transposase.
IS30	IS30 family transposase.
_IS3	IS3 family transposase.
_IS3’	IS3 family transposase.
_IS5	IS5/IS1182 family transposase.
_ISL3	ISL3 family transposase.
lare	lactaldehyde reductase.
_Llo	L-lactate oxidase.
ltl	leucine–tRNA ligase.
_LysR	LysR family transcriptional regulator.
_mannit	mannitol-1-phosphate 5-dehydrogenase.
mepr	membrane protein.
MerR#182	MerR family transcriptional regulator (orthology group #182).
_MerR	MerR family DNA-binding transcriptional regulator.
MerR	MerR family transcriptional regulator.
_metall	metallophosphatase.
_MFS	MFS transporter.
MFtr#108	MFS transporter (orthology group #108).
MFtr#118	MFS transporter (orthology group #118).
_multid	multidrug ABC transporter permease.
MutT	DNA mismatch repair protein MutT.
Nac#120	N-acetyltransferase (orthology group #120).
Nac#143	N-acetyltransferase (orthology group #143).
_Nad	N-acetyldiaminopimelate deacetylase.
_NAD	NAD-dependent malic enzyme.
_Na	N-acetyltransferase.
_NINE	NINE protein.
_nucleo	nucleotide pyrophosphohydrolase.
phlt#152	putative holin-like toxin (orthology group #152).
phlt	putative holin-like toxin.
_phosph	phosphopantetheinyl transferase.
ptl	proline–tRNA ligase.
_PTS	PTS mannitol transporter subunit IIA.
_QVPTGV	QVPTGV class sortase B protein-sorting domain-containing protein.

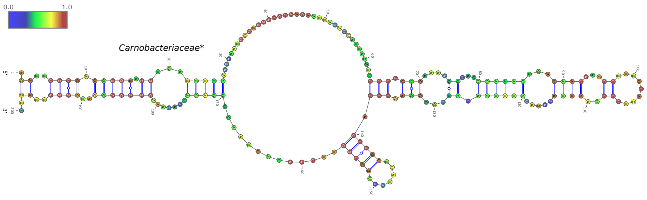
Additional File 5 — 6S RNA grouped consensus alignment (pdf)

Folded consensus structure of the 6S RNA groups generated with `RNAclust` (the names next to the colored square boxes indicate the family). The consensus secondary structures were calculated with `RNAalifold` and visualized using `VARNA`. The name right above the structure indicates the `RNAclust` group name in line with Fig. 1. Colors indicate sequence conservation within the respective LAB family.

🟧 *Aerococcaceae*

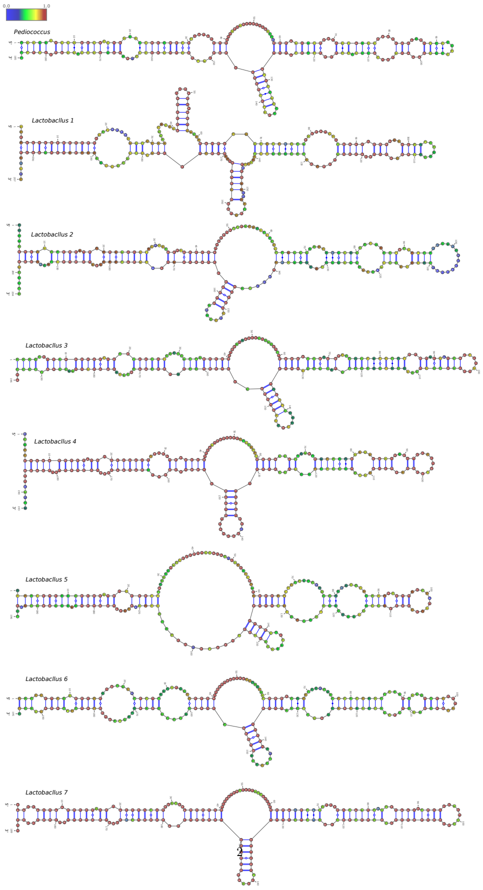


🟦 *Carnobacteriaceae*



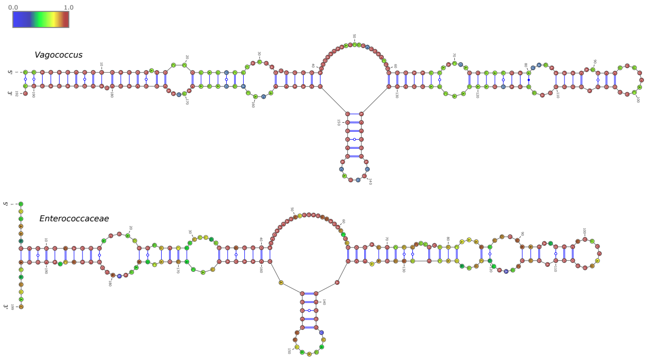
1

🟫 *Lactobacillaceae*

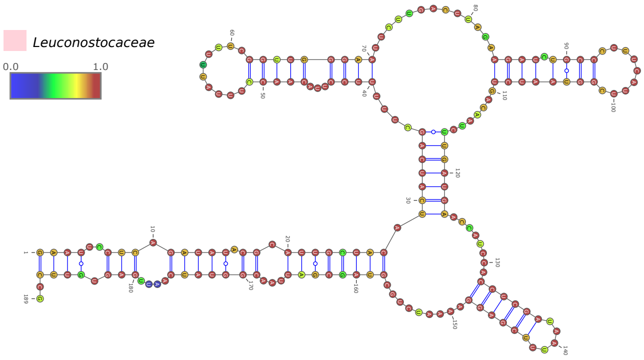


2

🟩 *Enterococcaceae*

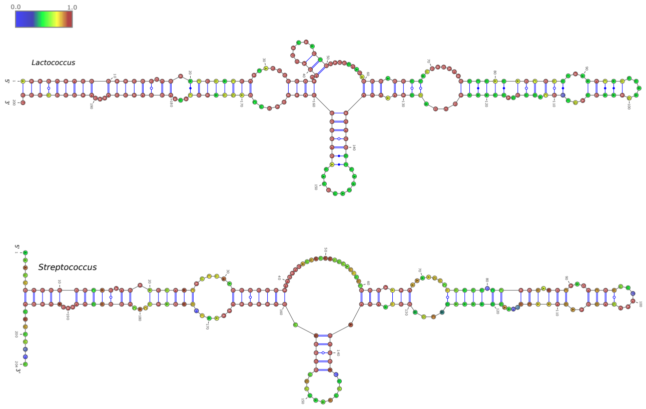


Leuconostocaceae



3

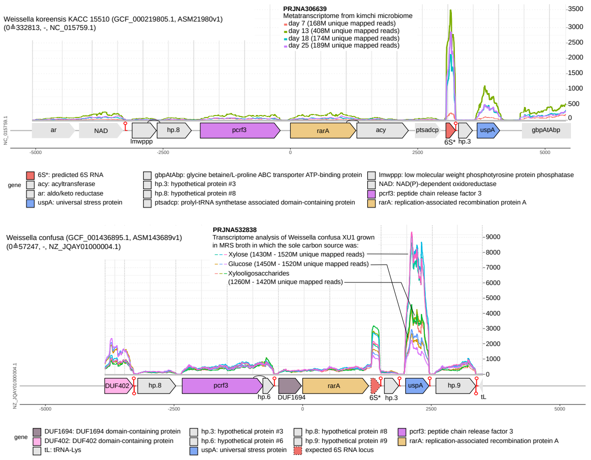
🟥 *Streptococcaceae*



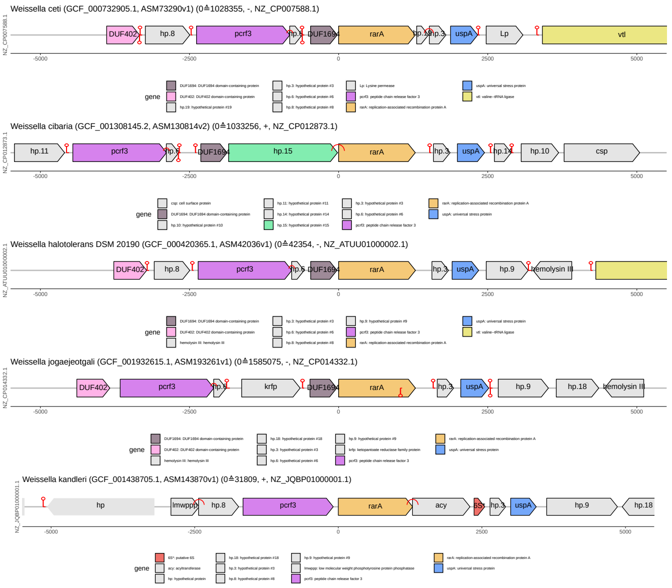
4

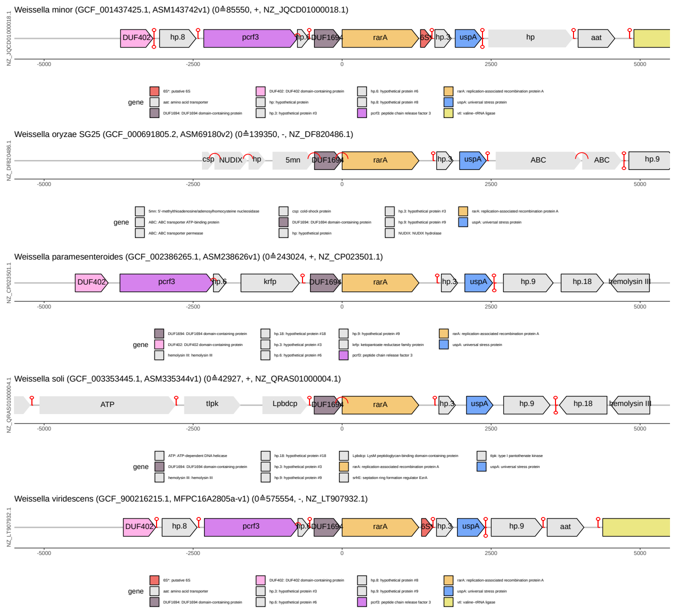
Additional File 8 — 6S RNA evidence in *Weissella*

Supplemental Figure 1: Genomic context of *rarA* in *Weissella koreensis* and *Weissella confusa* mapped RNA-Seq data from bioprojects PRJNA306639 and PRJNA532838. The number of mapped reads is indicated on the right. Conditions are overlayed in different colors. As for main Figure 2, putative Rho-independent terminators are indicated by red hexagons. Genes in close proximity (<20 nt) are indicated by a semicircle connecting them. The data verifies active transcription of the predicted 6S RNA in *W. koreensis*. No prediction was found for *W. confusa*. However, similar transcriptional activity is observed for the expected locus immediately downstream of *rarA*.

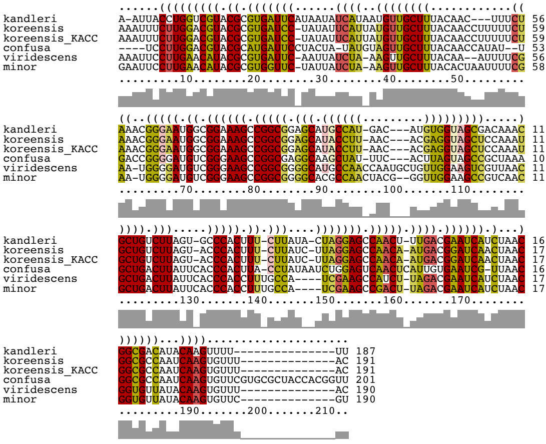


Supplemental Figure 2: Genomic context of *rarA* in further *Weissella* species. For each species, one representative strain is shown. Typically, *rarA* is followed by an intergenic region that is closed by a Rho-independent terminator. In three species, a low-scoring 6S RNA candidate was predicted in this locus (highlighted in red). We assume that a similar transcript is produced from the remaining intergenic regions of the other species.


Weissella ceti (GCF_000732905.1, ASM73290v1) (0≙1028355, -, NZ_CP007588.1)
DUF402
hp.8
pcrf3
DUF1694
rarA
uspA
Lp
vtl
Weissella cibaria (GCF_001308145.2, ASM130814v2) (0≙1033256, +, NZ_CP012873.1)
hp.11
pcrf3
DUF1694
hp.15
rarA
uspA
hp.10
csp
Weissella halotolerans DSM 20190 (GCF_000420365.1, ASM42036v1) (0≙42354, -, NZ_ATUU01000002.1)
DUF402
hp.8
pcrf3
DUF1694
rarA
uspA
hp.9
hemolysin III
Weissella jogaejeotgali (GCF_001932615.1, ASM193261v1) (0≙1585075, -, NZ_CP014332.1)
DUF402
pcrf3
krfp
DUF1694
rarA
uspA
hp.9
hp.18
hemolysin III
Weissella kandleri (GCF_001438705.1, ASM143870v1) (0≙31809, +, NZ_JQBP01000001.1)
hp
lmwppp
hp.8
pcrf3
rarA
acy
6S
uspA
hp.9
hp.18
-5000
-2500
0
2500
5000
gene


Weissella minor (GCF_001437425.1, ASM143742v1) (0≙85550, +, NZ_JQCD01000018.1)
DUF402
hp.8
pcrf3
DUF1694
rarA
6S
uspA
hp
aat
Weissella oryzae SG25 (GCF_000691805.2, ASM69180v2) (0≙139350, -, NZ_DF820486.1)
csp
NUDIX
hp
5mn
DUF1694
rarA
uspA
ABC
ABC
hp.9
Weissella paramesenteroides (GCF_002386265.1, ASM238626v1) (0≙243024, +, NZ_CP023501.1)
DUF402
pcrf3
krfp
DUF1694
rarA
uspA
hp.9
hp.18
hemolysin III
Weissella soli (GCF_003353445.1, ASM335344v1) (0≙42927, +, NZ_QRAS01000004.1)
ATP
tlpk
Lpbdcp
DUF1694
rarA
uspA
hp.9
hp.18
hemolysin III
Weissella viridescens (GCF_900216215.1, MFPC16A2805a-v1) (0≙575554, -, NZ_LT907932.1)
DUF402
hp.8
pcrf3
DUF1694
rarA
6S
uspA
hp.9
aat
-5000
-2500
0
2500
5000
gene

Supplemental Figure 3: Structural alignment of predicted 6S RNAs in *Weissella* species. *W. confusa* was predicted only based on RNA-Seq data (see Supplemental Figure 3 above).



```
                 ......((((((((.((.((((((......((((..((((((((...........((
kandleri         A-AUUACCUGGUCGUACGCGUGAUUCAUAAUAUCAUAAUGUUGCUUUACAAC---UUUCU 56
koreensis        AAAUUUCUUGGACGUACGCGUGAUCC-UAUAUUCAUUAUGUUGCUUUACAACCUUUUUCU 59
koreensis_KACC   AAAUUUCUUGGACGUACGCGUGAUCC-UAUAUUCAUUAUGUUGCUUUACAACCUUUUUCU 59
confusa          ----UCCUUGGACGUACGCAUGAUUCCUACUA-UAUGUAGUUGCUUUACAACCAUAU--U 53
viridescens      AAAUUCCUUGAACAUACGCGUGAUUC-AAUUAUCUA-AAGUUGCUUUACAA--AUUUUCG 56
minor            GAAUUCCUUGAACAUACGCGUGGUUC-UAUUAUCUA-AAGUUGCUUUACACUAAUUUUCG 58
                 .........10........20........30........40........50.........

                 ((..(((((.((.((((((.(((((.(((.((((((.........)))))))))).....)
kandleri         AAACGGGAAUGGCGGAAAGCCGGCGGAGCAUGCCAU-GAC---AUGUGGUAGCGACAAAC 112
koreensis        AAACGGGAAUGGCGGAAAGCCGGCGGAGCAUACCUU-AAC---ACGAGGUAGCUCCAAAU 115
koreensis_KACC   AAACGGGAAUGGCGGAAAGCCGGCGGAGCAUACCUU-AAC---ACGAGGUAGCUCCAAAU 115
confusa          GACCGGGGAUGUCGGGAAGCCGGCGAGGCAAGCUAU-UUC---ACUUAGUAGCCGCUAAA 109
viridescens      AA-UGGGGAUGUCGGGAAGCCGGCGGGGCAUGCCAACCAAUGCUGUUGGAAGUCGUUAAC 115
minor            AA-UGGGGAUGUCGGGAAGCCGGCGGGGCACGCCAACUACG--GGUUGGAAGCCGUCAAC 115
                 .........70........80........90........100.......110........

                 )))).))).....))))).)).))).....)))))))).))))).))))).)))))).)...)
kandleri         GCUGUCUUAGU-GCCCACUUU-CUUAUA-CUAGGAGCCAACU-UUGACGAAUCAUCUAAC 168
koreensis        GCUGUCUUAGU-ACCCACUUU-CUUAUC-UUAGGAGCCAACA-AUGACGGAUCAACUAAC 171
koreensis_KACC   GCUGUCUUAGU-ACCCACUUU-CUUAUC-UUAGGAGCCAACA-AUGACGGAUCAACUAAC 171
confusa          GCUGACUUAUUCACCCACUUA-CCUAUAAUCUGGAGUCAACUCAUUGUGAAUCG-UUAAC 167
viridescens      GCUGACUUAUUCACCCACCUUUGCCA----UCGAAGCCAUCU-UAGACGAAUCAUCUAAC 170
minor            GCUGACUUAUUCACCCACCUUUGCCA----UCGAAGCCGACU-UAGACGAAUCAUCUAAC 170
                 .........130.......140.......150.......160.......170........

                 ))))))...)))).....................
kandleri         GGCGACAUACAAGUUUU---------------UU 187
koreensis        GGCGCCAAUCAAGUGUUU--------------AC 191
koreensis_KACC   GGCGCCAAUCAAGUGUUU--------------AC 191
confusa          GGCGCCAAUCAAGUGUUCGUGCGCUACCACGGUU 201
viridescens      GGUGUUAUACAAGUGUUU--------------AC 190
minor            GGUGUUAUACAAGUGUUC--------------GU 190
                 .........190.......200.......210..
```

2.3 Proteinortho

Original Research Article: **Klemm, P.**, Stadler, P. F., & Lechner, M. (2023). Proteinortho6: Pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs. Frontiers in Bioinformatics. In revision.

frontiers

Proteinortho6: Pseudo-reciprocal best alignment heuristic for graph-based detection of (co-)orthologs

Paul Klemm [1], **Peter F. Stadler** [2,3,4,5,6] **and Marcus Lechner** [1,*]

[1] *Philipps-Universität Marburg, Center for Synthetic Microbiology, 35032 Marburg, Germany*
[2] *Bioinformatics Group, Institute of Computer Science and Interdisciplinary Center for Bioinformatics, Leipzig University, Härtelstraße 16-18, 04107 Leipzig, Germany*
[3] *Max-Planck-Institute for Mathematics in the Sciences, Inselstraße Leipzig, 04103 Leipzig, Germany*
[4] *Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, 1090 Wien, Austria*
[5] *Facultad de Ciencias, Universidad National de Colombia, Sede Bogotá, Colombia*
[6] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

Correspondence*:
Marcus Lechner
lechner@staff.uni-marburg.de

ABSTRACT

Proteinortho is a widely used tool to predict (co)-orthologous groups of genes for any set of species. It finds application in comparative and functional genomics, phylogenomics, and evolutionary reconstructions. With a rapidly increasing number of available genomes, the demand for large-scale predictions is also growing. In this contribution, we evaluate and implement major algorithmic improvements that significantly enhance the speed of the analysis without reducing precision. Graph-based detection of (co-)orthologs is typically based on a reciprocal best alignment heuristic that requires an all *vs.* all comparison of proteins from all species under study. The initial identification of similar proteins is accelerated by introducing an alternative search tool along with a revised search strategy – the pseudo-reciprocal best alignment heuristic – that reduces the number of required sequence comparisons by one-half. The clustering algorithm was reworked to efficiently decompose very large clusters and accelerate processing. Proteinortho6 reduces the overall processing time by an order of magnitude compared to its predecessor while maintaining its small memory footprint and good predictive quality.

*** Keywords: orthology, homology, sequence similarity, spectral clustering, algebraic connectivity**

INTRODUCTION

Comparative analyses of nucleic and amino acid sequences have become routine approaches in modern biology. A problem frequently encountered in comparative and functional genomics as well as in phylogenomics and evolutionary reconstructions is the detection of homologous genes that share an evolutionary ancestry. These genes are orthologs if they have derived from a common ancestor by means of a speciation event. Paralogs, in contrast, have derived from a duplication event and thus represent gene copies (Fitch, 1970). Orthologs are of particular interest as their function is likely conserved due to selective pressure (ortholog conjecture (Koonin, 2005)). In contrast, paralogs diverge faster, specialize, acquire new functions, or become dysfunctional (Ohno, 1999; Lynch and Conery, 2000). Gene duplications followed by subsequent speciation events create two or more genes in one lineage that are, collectively, orthologous to one or more genes in another lineage. These sets of genes are termed co-orthologs (Koonin, 2005). Even though orthology is not a transitive relation (Johnson, 2007), large-scale orthology assessment is often treated as a clustering problem, resulting in *clusters of (co)-orthologous genes* (COGs), see e.g. (Setubal and Stadler, 2018) for a review. Proteinortho (Lechner et al., 2011) in its previous version 5 (Proteinortho5) is a well-established tool for the detection of (co-)orthologs in large-scale analysis that also adheres to this approach. It has demonstrated its utility in various studies within the field of comparative genomics including e.g. evolutionary analyses (Peter et al., 2018), genomic signatures (Kapheim et al., 2015), functional annotation (Pinho et al., 2013), phylogenetic reconstructions (Klemm et al., 2022), and so on. Proteinortho also found integration into tools and databases, such as Echinobase (Arshinoff et al., 2022) or Funannotate (Palmer and Stajich, 2023).

Sequence-based orthology inference is based on pairwise sequence comparisons. This stage requires scoring the similarity of all proteins in order to determine groups with high similarity. To simplify the terminology, we use the term “protein” to designate the amino acid representation of protein-coding gene sequences in the following. The well-known reciprocal best alignment heuristic (RBAH) (Bork et al., 1998), can be used to retrieve at least a good approximation of the correct ortholog set. We refer to Schaller et al. (2021) for a comprehensive mathematical analysis of the relation between best matches and orthology. Proteinortho extends the RBAH to an adaptive version, which includes alternative matches to the set of potential orthologs if they closely resemble the similarity of the best match. For details, refer to the original implementation (Lechner et al., 2011). Pairwise sequence comparisons are typically the most time-consuming stage as the computational effort scales quadratically to the number of proteomes analyzed.

When all pairwise sets of reciprocal best hits are known, this information is merged. In this process, all proteins are represented as nodes in a graph that are connected by edges whenever their similarity score is within the adaptive RBAH criterion. A set of proteins linked to each other by any path is called a connected component (CC). Each CC represents a potential co-orthologous group. However, the small world phenomenon (Milgram, 1967) also applies to empirical orthology graphs: Even though the number of possible protein sequences is practically limitless, there are relatively few basic folding shapes, of which some folds and superfamilies are extremely abundant (Koonin et al., 2002). CCs quickly become large and thereby non-informative. This effect increases with the number of proteins analyzed at once. Therefore, a clustering step is required. CCs are divided into smaller, more informative CCs by iteratively isolating well-connected subsets. The results are clusters of mutually similar proteins reported as co-orthologous groups.

In this contribution, we evaluate major algorithmic improvements for Proteinortho and present version 6 of the tool (Proteinortho6). All improvements primarily aim towards a significant speedup of orthology analyses while keeping the quality of its results and the small memory footprint that makes it applicable on large HPC systems and average off-the-shelf desktop systems.

1 METHODS

1.1 Alternative sequence search tools

The first stage of Proteinortho analyses is a pairwise sequence comparison. Proteinortho5 relies on BLAST (Camacho et al., 2009) which is still considered the gold standard for any homology search (Ward and Moreno-Hagelsieb, 2014). BLAST implements a seed-and-extend-paradigm. Meanwhile, it has inspired numerous alternative algorithms that can be used as direct replacements. Here, we evaluate these alternatives for use in the context of the adaptive RBAH strategy in order to speed up the sequence comparisons performed for orthology inference.

Proteinortho6 directly supports the following BLAST alternatives: ucsc BLAT is optimized for quickly finding very similar sequences of closely related species. It uses an index of non-overlapping k-mers (Kent, 2002) to speed up the search. UBLAST uses spaced seeds and a reduced alphabet to facilitate the comparison of distant gene sequences with a low identity (Edgar, 2010). USEARCH instead requires exact matches and was designed for comparisons of sequences with a high identity (Edgar, 2010). LAST implements a suffix array for a variable seed length, spaced seeds, and a reduced alphabet. A design goal was to handle repeat-rich sequences more efficiently than other tools (Kiełbasa et al., 2011). The parameter m (default 10) controls the maximum initial matches per query position comparable to the max_target_seqs parameter of BLAST. The higher m, the more hits are reported at the cost of increased running time and memory usage. RAPSearch2 is based on a collision-free hash table of sorted 6-mers and a reduced alphabet for amino acid sequences (Zhao et al., 2012). DIAMOND implements a double index alignment, spaced seeds, and a reduced database alphabet (Buchfink et al., 2015). It provides several sensitivity modes depending on the expected sequence identity of reported hits. The default is optimized for hits $> 60\%$ identity and short read alignment. The fast mode aims for highly similar hits with $> 90\%$. The sensitive mode is recommended for comparisons above $> 40\%$ sequence identity, while the highest sensitivity setting ultra-sensitive is supposed to perform well even below 40% identity, although with largely increased running time. MMSeqs2 uses a memory-efficient inexact k-mer matching optimized for multi-core systems (Steinegger and Söding, 2017). Speed and sensitivity can be controlled with the s parameter. A reasonable range starts from 1, corresponding with fast but coarse results, to 7.5, which is highly sensitive but slow. The default value is 5.7 and thus aims towards sensitivity over speed. Topaz is the most recent addition of BLAST replacements. It uses an advanced version of the SANS algorithm (Koskinen and Holm, 2012) that generalizes the symmetric suffix array neighborhood search to an asymmetric search in combination with scored seeds, a variation of spaced seeds (Medlar and Holm, 2018). Similarly to the tools above, a fast mode is implemented that decreases running time at the expense of sensitivity.

The results obtained using BLAST were considered as the point of reference. Based on these, we computed sensitivity and precision, where sensitivity = TP/(TP+FN) and precision = TP/(TP+FP) and TP is the number of true positive reported edges, that coincide with BLAST, FP is the number of false-positive reported edges, that do not coincide with BLAST, and FN is the number of false-negatives edges, that are only reported by BLAST. Computational efficiency was quantified in terms of total running time (wall time), scalability (running time in relation to the number of species), and maximal memory allocation (peak memory consumption). The evaluation was performed using the following tool versions: BLAST+ (v2.13.0), ucsc BLAT (v377), UBLAST and USEARCH (v11.0.667), LAST (v1318), RAPSearch2 (v2.24), DIAMOND (v2.0.15), MMSeqs2 (v14.7e284), and topaz (commit 24bdb61).

Note that the free 32-bit versions of USEARCH and UBLAST were used instead of the 64-bit versions that are available only commercially. Even though these versions are likely faster, we do not expect that the sensitivity and precision of the tool are affected by the build architecture.

1.2 Pseudo-reciprocal sequence comparison strategy

Pairwise similarity scores between all proteins in the dataset are the foundation of sequence-based orthology inference via adaptive RBAH. For reasons of complexity, only scores below a certain expectation value (E-value) are considered. In Proteinortho, sets of proteins $S_1, S_2, \cdots, S_n$ are presented for each species of interest. Similarity scores are then calculated using a sequence search tool st, like BLAST. This is performed reciprocally for all pairs of sets, e.g., $st(S_1, S_2), st(S_2, S_1), st(S_1, S_3), st(S_3, S_1), \cdots$ in order to obtain all scores required for RBAH. Notably, the alignments of any two proteins $a \in S_n$ and $b \in S_m$ are calculated twice if their match is below the E-value threshold in the comparisons $st(S_n, S_m)$ and $st(S_m, S_n)$.

The new feature pseudo (pseudo-reciprocal) in Proteinortho6, calculates only one pair $st(S_n, S_m)$ and approximates the results of the $st(S_m, S_n)$. The missing E-values of $st(S_m, S_n)$ are calculated based on the query sequence length l, and the database size $|S_n|$ of the respective set of proteins in order to resemble E-values comparable to a pair-wise search:

$$e = \frac{l \cdot |S_n|}{2^{bitscore}}$$

1.3 Clustering algorithm

Eigenvector decomposition

Proteinortho uses a spectral clustering algorithm. It recursively divides connected components into two connected subcomponents that are maximally connected with respect to their algebraic connectivity (Fiedler, 1975). Spectral clustering has a long history in multivariate statistics, image processing, and machine learning, see e.g. Shi and Malik (2000) for detailed descriptions. The implementation is based on the eigenvector decomposition of subgraphs, which are calculated via the power iteration in Proteinortho5 (Boutsidis et al., 2015). As large components usually build up due to bridge and hub clusters, most nodes within a connected component are not connected by an edge which is exploited by representing the data via a space-efficient edge list rather than a largely unoccupied adjacency matrix. This data structure is also well utilized by the power iteration. In contrast to alternative implementations based on adjacency matrices, non-existing edges do not require memory nor do they require consideration during the calculations. The strategy enables large-scale clustering by minimizing memory requirements and computational effort (Lechner et al., 2011).

In addition to the power iteration, Proteinortho6 implements ssyevr (single precision, symmetric eigenvalue problem, RRR algorithm). It is based on the "Relatively Robust Representation" algorithm (Parlett and Dhillon, 2000) that can compute an eigenpair in linear time (Bientinesi et al., 2005) which is provided via the highly optimized Fortran 77 library Lapack (v3.8.0) (Anderson et al., 1999). Although ssyevr outperforms the power iteration by orders of magnitude in many scenarios, the Lapack routine cannot be applied for large clusters of protein as it is bound by quadratic memory requirements due to the reliance on adjacency matrices.

Flooding heuristic

With a growing number of species that are analyzed at once, connected components in orthology graphs grow exponentially in size due to the small world phenomenon. The resulting CCs can quickly cover a large proportion of the whole protein set. An example of this observation is shown in Supplemental Data 1. Theoretically, these huge CCs are easily broken down into informative subsets by spectral clustering. However, with an increasing number of species, their size poses a computational problem. The power iteration algorithm is not able to process them in a reasonable time while the memory requirements for `ssyevr` are not feasible. Hence, orthologs in these large CCs cannot be recovered.

To salvage the issue with large CCs, `Proteinortho6` employs an iterative approach that removes batches of outlier edges based on their associated bitscore when spectral clustering is not possible. Therefore a cutoff threshold is raised until a significant number of outliers is covered with respect to the one-sided Grubb-Smirnov outlier test. If necessary, this process is repeated until spectral clustering is possible.

Multithreading

`Proteinortho6` introduces support for parallel computing at the clustering stage. The main thread employs a breadth-first search (BFS) approach to identify CCs. The worker threads then calculate the algebraic connectivity in parallel for each CC. Split components are added back to the processing queue if necessary. This feature also facilitates distribution across multiple computing nodes by processing batches of connected components in parallel. An overview can be found in Supplemental Data 1.

Adaptive clustering

The spectral clustering approach follows a bisecting paradigm. Groups are successively divided until a predefined algebraic connectivity threshold is met. The choice of this threshold directly affects the size and quality of reported (co-)orthologous groups. A high connectivity threshold will only return sets of mutually similar proteins but can lead to excessive fragmentation of the orthology graph in numerous small CCs. Orthologous groups might fall apart into several subsets. A low threshold, on the other hand, might return non-informative large CCs with multiple putative co-orthologs for each species that actually represent unions of several orthologous groups. The default threshold applied by `Proteinortho` was defined empirically and represents a reasonable trade-off between both extremes.

Different protein families have different overall similarities. Therefore, a connectivity threshold that works well for one protein family, might be suboptimal for another. To address this, `Proteinortho6` offers an adaptive clustering with the `core` option. It assumes that members of orthologous groups should be found in all species. Iterative spectral clustering is applied irrespective of the graph's connectivity until the graph would split into two subgraphs of which neither covers all species that were covered by the original CC. The CC is only clustered further if it appears too big, e.g., comprises many (co-)orthologous genes per species. This threshold is defined by the parameter `coreMaxProts` (default 10), which continues clustering if more than 10 proteins are present per species.

1.4 Evaluation

Datasets

Several real-world datasets were used as a biologically relevant basis for representative comparisons. These are summarized in Supplemental Data 2. It shows the number of species and proteins for each dataset and how this translates into a reciprocal best hit graph (RBH) using `Proteinortho6` with `BLAST` (E-Value threshold 10^{-5}).

The dataset $\texttt{QfO}_{2020/04}$ was provided by the `QfO` benchmark service (Altenhoff et al., 2016). It comprises a curated set of proteomes from 23 Bacteria, 7 Archaea, and 48 Eukaryota sampled from `UniProt` (UniProt-Consortium, 2018). Note that `QfO` provides two versions of this dataset, and we used the newer version with updated UP000008143 sequences.

The `Bac` dataset comprised all bacterial reference proteomes from `UniProt`, release 2022/03 (UniProt-Consortium, 2018). This set was downsampled to incremental subsets of random proteomes. For instance, $\texttt{Bac}_{10}$ contains 10 randomly selected bacterial proteomes, $\texttt{Bac}_{20}$ extends this set by 10 additionally randomly selected proteomes, and so on. A full list is shown in Supplemental Data 2.

The `BigCC` set comprises connected components of $1,800$ bacteria for which an origin of replication was identified (related study not published so far). Due to a huge connected component, this dataset represents a challenge for the clustering algorithm. So far, it was not solvable using regular spectral clustering. A subset of this is `BigCC100` which focuses on larger CCs with at least 100 nodes. To evaluate edge cases that were not covered by this real-world dataset, such as components with high density and a large number of nodes, a set of 300 simulated graphs was generated. The set will be referred to as `simulated`. Its connected components were generated in three steps: An unweighted path graph was generated with the given number of nodes n and $n-1$ edges connecting each node in a series to ensure connectivity. Edges were added one by one, randomly assigning unconnected nodes until the given graph density was satisfied. Bitscores were defined randomly (between 1 to 2000). E-Values were trivially set to $1/\texttt{bitscore}$.

Benchmark system

All benchmarks were conducted on the HPC cluster MaRC3 located at the University of Marburg using AMD EPYC 7702P processors with 64 cores and 256 GB RAM.

Clustering algorithms

The spectral clustering algorithms were applied to the datasets `BigCC100` and `simulated`, representing particularly large connected components. A total of 8,881 connected components were evaluated in this way, see Supplemental Data 2. If the relative clustering time differed by less than 5 minutes or one $\log_2$ fold, the algorithms were considered to be equally fast. To evaluate the comparability of both clustering approaches, the adjusted rand index (ARI) was used (Hubert and Arabie, 1985). The higher the ARI value, the more similar the partitioning.

Precision of orthology predictions

The `QfO` benchmark service was used to evaluate the orthology predictions (Altenhoff et al., 2016). The Nextflow implementation of the benchmark system was used as provided in the corresponding `GitHub` repository Altenhoff (2023). All benchmarks were performed using the $\texttt{QfO}_{2020/04}$ (2020.2) dataset. In this analysis, the precision metrics of the three categories of benchmarks were employed:

1. Phylogeny-based benchmarks `GSTD2` (4 tests), the generalized species tree discordance, as well as the `STD` (3 tests), the species tree discordance, using the Average Robinson-Foulds (RF) distance between predicted gene trees based on the set of orthologs and the underlying species tree (the lower the better). The RF metric is a dissimilarity measure that quantifies the difference between two trees by counting the number of partitions that can be observed in one phylogenetic tree but not the other and

vice versa. This metric can be seen as an approximation of the false discovery rate or the inverse of precision Altenhoff et al. (2016).

2. Function-based benchmarks used EC, the Enzyme Classification Conservation, and GO, the Gene Ontology Conservation, through the Average Schlicker Similarity as a proxy for precision (the higher, the better). The Average Schlicker Similarity is a semantic similarity measure used to assess the terms.
3. Reference Orthology-based benchmarks examined the agreement with the SwissTree, VGNC or TreeFam-A gene phylogeny, measured by the Positive Predictive Value (PPV, the higher, the better).

Full details on the test statistics can be found in (Altenhoff et al., 2016).

To combine the precision metrics of the different benchmarks, we define improvement as the mean $\log_2$ fold ratio between all scores. The scores of Proteinortho v5.16b with default settings serve as the baseline and for example, an improvement of 0.5 correspond to scores that are on average 41% better than the results of Proteinortho5.

Scalability

The evaluation was performed with the Bac datasets of increasing size (Bac$_{10}$, Bac$_{20}$, $\cdots$). The following tools were evaluated in the comparison: OrthoFinder v2.5.4 (Emms and Kelly, 2019) in the graph-based modus (-og) using MMSeqs2 v14.7e284, SonicParanoid2 v1.3.8 (Cosentino and Iwasaki, 2023) using DIAMOND v2.0.15 in sensitive mode, OMA v2.5.0 (Altenhoff et al., 2019) without an out-group set, Proteinortho5 v5.16b utilizing BLAST v2.13.0, and Proteinortho6 v6.3.0 with DIAMOND v2.0.15 in sensitive mode, as well as the pseudo-reciprocal variation. A full list of all dependencies, versions, and parameters is provided in Supplemental Data 2.

2 RESULTS

2.1 Sequence search tools

Pairwise similarity data is fundamental to graph-based orthology inference. The computation of all *vs.* all comparisons using a sequence search tool is also the most costly step. Typically, BLAST was the search tool of choice. It offers excellent performance compared to directly calculating scores from pairwise alignments and is considered the gold standard in terms of sensitivity and precision (Ward and Moreno-Hagelsieb, 2014). To our knowledge, more modern search tools are less accurate in general but perform much better in respect to processing time and memory consumption (see Tab. 1). We systematically compared potential alternatives to BLAST in the context of Proteinortho's adaptive reciprocal best-hit heuristic (Lechner et al., 2011). The evaluation is based on the QfO 2020/04 dataset (Altenhoff et al., 2016), which comprises a representative mix of eukaryotic, bacterial, and archaeal proteomes.

The original implementation of Proteinortho5 relies on BLAST. It required 97 GB of memory and about three days (78h) of processing time in total. Table 1 shows that both running time as well as memory consumption improve significantly if alternative search tools are used. In terms of precision, ucsc BLAT stands out with 94% and is best in total processing time (21 minutes, 7.8 $\log_2$ fold improvement) as well as memory footprint (2 GB, 5.5 $\log_2$ fold improvement over BLAST). However, this tool returns the lowest number of edges and achieves the by far worst sensitivity of all options (20%). Similarly, RAPSearch2, and USEARCH also fall behind in terms of sensitivity (47% and 52%, respectively). The remaining tools are close regarding precision (around 90%) and sensitivity (usually between 80% and 90%). With respect to both measures of quality, DIAMOND, LAST, MMSeqs2, topaz, and UBLAST could serve as suitable

Table 1. Performance and resource consumption of sequence search tools in the context of Proteinortho based on the QfO benchmark dataset 2020/04. Alternative search modes are listed below the tool's names. The default option is indicated (def.). Sensitivity and precision are given relative to the BLAST results in line 1. Edges: number of edges in the initial orthology graph; wall time: total processing time; memory: peak memory usage; l_2FC: $\log_2$ fold change relative to Proteinortho5 results; $*$: default option of Proteinortho6. Ranks are indicated: 🟩 top 25%, 🟧 top 50%.

algorithm	edges	sensitivity	precision	wall time		memory	
		%	%	l_2FC	h	l_2FC	GB
Proteinortho5	5435k	100	100	0	77.8	0	97
DIAMOND							
default	4701k	77	89	7.2	0.5	4.1	6
sensitive	5366k	88.4	89.5	6.4	0.9	4	6
+ pseudo*	5417k	88.7	88.9	7.5	0.4	4.3	5
ultrasens	5457k	89.7	89.3	4.6	3.2	3.8	7
fast	3894k	63.8	89	7.3	0.5	4.4	4
LAST							
m10 (def.)	4853k	79.5	89	6.6	0.8	3.5	9
m100	5118k	84.2	89.4	5.1	2.3	2.3	20
m1000	5239k	86.2	89.5	2.2	16.5	1.9	25
MMSeqs2							
s1	3877k	64	89.7	6.9	0.7	3.4	9
s5.7 (def.)	5149k	85.6	90.4	4.5	3.5	2.9	13
s7.5	5235k	87.1	90.5	2.7	12	2.9	13
topaz							
default	5025k	82.3	89	4	4.9	3.2	10
fast	5025k	82.3	89	4.1	4.5	3.2	11
USEARCH							
ublast	5167k	81.1	85.3	5.5	1.8	2.1	23
usearch	3215k	51.8	87.5	7.6	0.4	5.5	2
ucsc BLAT	1158k	20	94	7.8	0.3	5.5	2
RAPSearch2	2781k	46.6	91.1	2.2	16.9	3.1	11

BLAST replacements when applying the right search mode. Factoring in processing time and memory requirements, DIAMOND with the sensitive option was evaluated to be the most optimal approach.

Using DIAMOND with the sensitive option as the search tool improved the running time by a $\log_2$ fold of 6.4 (to 56 minutes instead of 77.8 h) and the memory consumption by 4 $\log_2$ units (peak memory usage of 6 GB instead of 97 GB). In addition, we applied the pseudo-reciprocal sequence comparison strategy, pseudo. Here, protein alignments are calculated only in one direction while the reverse direction is estimated. See the Methods section for details. This approach additionally speeds up the calculation by half. Compared to the classic search strategy, the measures of quality are hardly affected. Precision decreases from 89.5 to 88.9% while sensitivity increases from 88.4 to 88.7%.

Comparable outcomes were noted for a group of closely related species and for randomly selected bacterial proteomes from the Bac_n dataset. For additional details, please refer to Supplemental Data 1. The pseudo-reciprocal best alignment heuristic using DIAMOND with the sensitive option, therefore, became the new default for Proteinortho6.

2.2 Clustering algorithm

Once pairwise similarity data was merged into an overarching graph structure, spectral clustering is applied to reduce it to an orthology graph. `Proteinortho` recursively divides connected components into two connected subcomponents that are maximally connected with respect to their algebraic connectivity. For this process, the space-efficient power iteration is used in `Proteinortho5`. With `Proteinortho6`, the `ssyevr` algorithm is available as an alternative. It relies on full matrices and is thus less space-efficient. We conducted a comprehensive evaluation of the running time differences between the power iteration and `ssyevr` algorithms using the real-world dataset `BigCC100` was used together with a `simulated` set that comprises components with high density and a large number of nodes. See the Methods section for details.

Lapacks `ssyevr` has a significantly larger memory footprint for large connected components. The maximal requirement for processing a CC in the reference datasets was around 18 MB (`ssyevr`) *vs.* 0.1 MB (power iteration), see Supplemental Data 2 for details. Given the availability of system memory in modern computer systems, these additional requirements are largely outweighed by the improvement in performance. The maximal relative improvement in running time was 9.3 $\log_2$ folds for a graph with 1,921 nodes (4 seconds using `ssyevr` *vs.* 40 minutes the power iteration), and the maximum absolute running time difference was 1.36 hours for a simulated graph with 7,731 nodes. 217 out of the 8,881 connected components were processed significantly faster using `ssyevr` over the power iteration. The improvement was 5.1 $\log_2$ folds on average. Our evaluation shows that the `ssyevr` implementation is consistently faster for large components and on par with the power iteration for small components. For this reason, the power iteration was replaced by the `ssyevr` as the default clustering algorithm in `Proteinortho6`.

Notably, the chosen algorithm scales quadratically in memory with the number of nodes. The `BigCC100` dataset already comprises a connected component with 4 million nodes which exceed feasible computing capacities. While the power iteration would be able to handle this component from a memory perspective, the processing time would largely exceed any reasonable value. We stopped the comparative evaluation of clustering this component after 10 days. The increasing appearance of large connected components with an increase of species that are analyzed for (co-)orthologous proteins is expected due to the small world phenomenon (Milgram, 1967). We found a number of additional components in real-world datasets that are close in size. Hence, a large proportion of the proteins cannot be assigned to any (co-)orthologous group, if the components are ignored. To avoid a loss of information due to this effect, `Proteinortho6` employs a flooding heuristic. Low-scoring edges are iteratively removed from large components until they are decomposed to sufficiently small subcomponents that are suitable for spectral clustering. See Methods sections for details.

2.3 Pseudo-reciprocal best alignment heuristic

To assess the validity of the pseudo approach, the reciprocal best hit graph from the QfO 2020/04 data sets was evaluated using the classic RBAH and the pseudo approach. Here, bitscores calculated by `DIAMOND` differ by 1.1% in median and 1.9% on average for any pairs of proteins (`Proteinortho6` with default parameters), see Supplemental Data 2. It is not surprising, given that the same sequences are aligned just with differing starting points. Although $st(S_n, S_m) \neq st(S_m, S_n)$ in general, the reciprocal bitscores for any two proteins of these sets are highly similar. With that, one can assume $st(S_n, S_m) \sim st(S_m, S_n)$, hence the calculation of $st(S_m, S_n)$ can be omitted by estimating the scores based on $st(S_n, S_m)$ as described in the Method section. This reduces the algorithmic effort by a factor of two. E-values calculated in this way strongly correlate with the reciprocal E-values (R^2_{adj}=0.99). This correlation between the pseudo and classic approach is stronger compared to any comparison between two homology search tools using the classic approach. For more details see Supplemental Data 1.

2.4 Adaptive Clustering

Table 2. Key performance indicators of different clustering parameters applied to the QfO benchmark dataset 2020/04 (78 species). similarity: ARI compared to the Proteinortho5 clustering with default parameters, classic: classic adaptive reciprocal best hit algorithm, *: default, α: algebraic connectivity threshold.

	ortho-groups					core-groups		
	total	0-25% species	25-50% species	50-75% species	75-100% species	total	max(proteins/group)	similarity
Proteinortho5:								
default (BLAST)	84k	80k	3k	988	97	0	0	1
default clustering (DIAMOND)	79k	75k	3k	984	97	0	0	.816
Proteinortho6 with DIAMOND sensitive:								
pseudo	72k	67k	3k	1k	105	0	0	.822
$\alpha = 0.1$*	72k	68k	3k	1k	106	1	97	.821
$\alpha = 0.05$	65k	61k	3k	1k	147	1	97	.819
$\alpha = 0.2$	79k	75k	3k	994	60	0	0	.797
$\alpha = 0.3$	83k	79k	3k	831	40	0	0	.769
$\alpha = 0.01$	53k	48k	3k	1k	231	9	149	.706
$\alpha = 0.5$	90k	87k	2k	484	11	0	0	.692
$\alpha = 0.75$	99k	97k	2k	186	10	0	0	.606
core	44k	40k	2k	1k	392	51	706	.352
$\alpha = 0.005$	48k	43k	3k	1k	262	12	152	.15
$\alpha = 0.001$	42k	37k	3k	1k	319	30	315	.141
$\alpha = 0.00001$	36k	32k	2k	1k	377	50	3k	.0923

Regular clustering of a CC is performed by bisecting it into two sub-CCs of maximized connectivity until a predefined algebraic connectivity threshold is met. The default threshold applied by Proteinortho was defined empirically. Instead of working with a fixed threshold, the adaptive clustering strategy (core) assumes that members of orthologous groups should be found in all species. Iterative spectral clustering is applied until the component would split into two subcomponents of which neither covers all species that the original CC covered. The algorithm is aimed to keep orthologous groups as big as they need to be to cover all initially present species, even if the connectivity criterion is not met yet. This strategy is meant to identify the pan-genome as e.g. as the basis for reconstructing phylogenetic supertrees based on the reconstruction of trees from multiple orthologs.

Table. 2 shows an overview of the number of reported orthologous groups relative to the percentage of species covered in the dataset. We found a high number of core-groups, i.e., orthology groups that span all input species, using the adaptive clustering, especially compared to the default connectivity threshold for the QfO dataset 2020/04. A comparable number of core-groups is found with a very low threshold of $1e^{-5}$ but at the same time increasing the maximal number of proteins per group dramatically. Overall the core module shows the best trade-off between the number of core-groups and size. It is worth noting that the results of the core approach differ drastically from the results from Proteinortho5 (ARI: 0.35).

2.5 Scalability

Proteinortho6 implements a number of upgrades that improve the processing time to a level that matches recent tools for the identification of orthologs such as SonicParanoid2 (Cosentino and Iwasaki, 2023) without compromising the quality of the predictions. As large-scale orthology assessment relies on pairwise sequence comparisons, processing time grows quadratically with the number of proteins to be compared. This number correlates with the number of species in an orthology analysis. To portray the scalability and thus the processing time relative to the size of analyses, we used a real-world dataset. It is based on randomly sampled proteomes of the bacteria kingdom provided by UniProt (UniProt-Consortium, 2018). Details can be found in the 1.4 section.

Fig. 1 shows the processing time and Supplemental Data 1 the memory consumption for an orthology analysis as a function of the number of species. OMA and Proteinortho5 exhibited the poorest scaling in terms of processing time, with a quadratic coefficient of $1.9 \cdot 10^{-3}$ and $2.9 \cdot 10^{-4}$ respectively, making an application to large species sets unfavorable. OrthoFinder, SonicParanoid2 and Proteinortho6 scale significantly better with the number of species. Proteinortho6 applying the classic reciprocal best alignment heuristic scales similarly to OrthoFinder in terms of processing time and outperforms the alternatives in terms of memory consumption. The pseudo reciprocal best alignment heuristic of Proteinortho6 and SonicParanoid2 show the best overall scaling results in regards to both metrics.

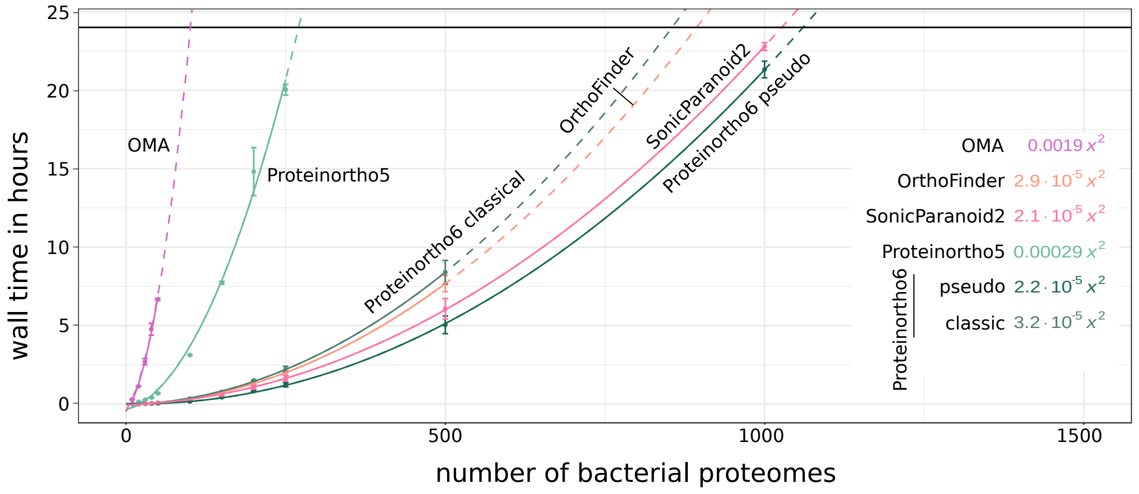


Figure 1. Scalability of total orthology prediction, including the all-versus-all sequence comparison and clustering, relative to dataset size of randomly selected bacterial proteomes of UniProt 2022_03 (Bac$_{10,20,\ldots,1000}$). Average processing times are indicated by circles and fitted using a quadratic function (solid line, $R^2_{adj} \geq 0.99$) for extrapolation (dashed lines). Details on parameters and versions can be found in the Supplemental Data 1 and 2.

2.6 QfO Sensitivity Bias

An assessment of orthology prediction quality can be performed using Quest for Orthologs (QfO). The evaluation tool offers various tests to measure the precision and recall of predictions from different perspectives (for more information, refer to the Materials section). As exemplified below, we noticed a bias in the evaluation tool regarding the recall metric. True orthology relations can only be estimated based on existing data e.g. via shared GO terms or congruence to curated species trees (Altenhoff and Dessimoz, 2009). Some QfO tests use the number of predicted orthologs as a proxy for sensitivity or recall, which translates into the number of edges in the orthology graph. In turn, the metric prefers large graphs. We exemplify this based on the results of “OrthoMCL” (Hickman, 2021) and “SonicParanoid_sensitive” (Consentino, 2021), which are among the highest recall scores across the different benchmarks. The referenced results of “SonicParanoid_sensitive” from 78 species include an orthology group that comprises over 5,000 proteins per species. Similarly, the results of “OrthoMCL” contain a group with around 42 proteins per species. The biological informativeness of such a large group, in particular in relation to the small number of input species is questionable at best. In comparison, the largest group `Proteinortho` reports contain 3.5 proteins per species (with default clustering). In our observations, there appears to be a consistent trend where an increased count of edges generally results in higher sensitivity or recall scores across most benchmarks.

To further exemplify this bias we constructed “group reference” with `Proteinortho6` with `DIAMOND` and default parameters with the exception of a relaxed clustering ($\alpha = 0.00001$). To magnify the effect, we opted to work with groups instead of a list of pairs, where every pair of proteins within a group was predicted to be orthologous. In total “group reference” contained approximately ten times as many orthologs as “SonicParanoid”. This approach achieved a Pareto optimal solution with high recall, as shown in Fig. 2. Similar effects could be observed for almost all benchmark results (see Supplemental Data 1 and 2 for more details). We are questioning this metric used and the strength of the Pareto optimal solution as a benchmark system as it is tied to this metric. For this reason, we will focus on the precision measurements of the benchmarks.

2.7 QfO Assessment

We found that using `Proteinortho6 classic`, which utilizes the classic adaptive best hit algorithm, with `default` clustering (`ssyevr`) mostly achieves precision scores within the top 25% and otherwise among top 50% for all benchmark tests with the exception of the VGNC benchmark as summarized in Tab. 3. In general, all `Proteinortho` parameterizations and variations produce below-average precision scores in the VGNC benchmark. `Proteinortho6`, including the `pseudo` extension, showed similar precision scores to those obtained with `Proteinortho5`, with the majority of the benchmarks ranking within the top 25%. Overall, precision scores are similar to the `Proteinortho5` results with mean $\log_2$ ratios below 0.005. Exchanging `BLAST` with `DIAMOND` in `Proteinortho5` results in similar but slightly improved scores. Regarding `Proteinortho` parameterizations, the adaptive clustering (`core`) performs slightly worse overall with an `improvement` of -0.028.

We found that the flooding heuristic performed similarly to the case without any clustering, highlighting the validity of this approach as a fallback system for the clustering if the size of a CC extends the capabilities of the spectral clustering algorithm. The conceptually simplified versions `pseudo` mode exhibited slightly better precision scores that are very similar to the results of `Proteinortho5`.

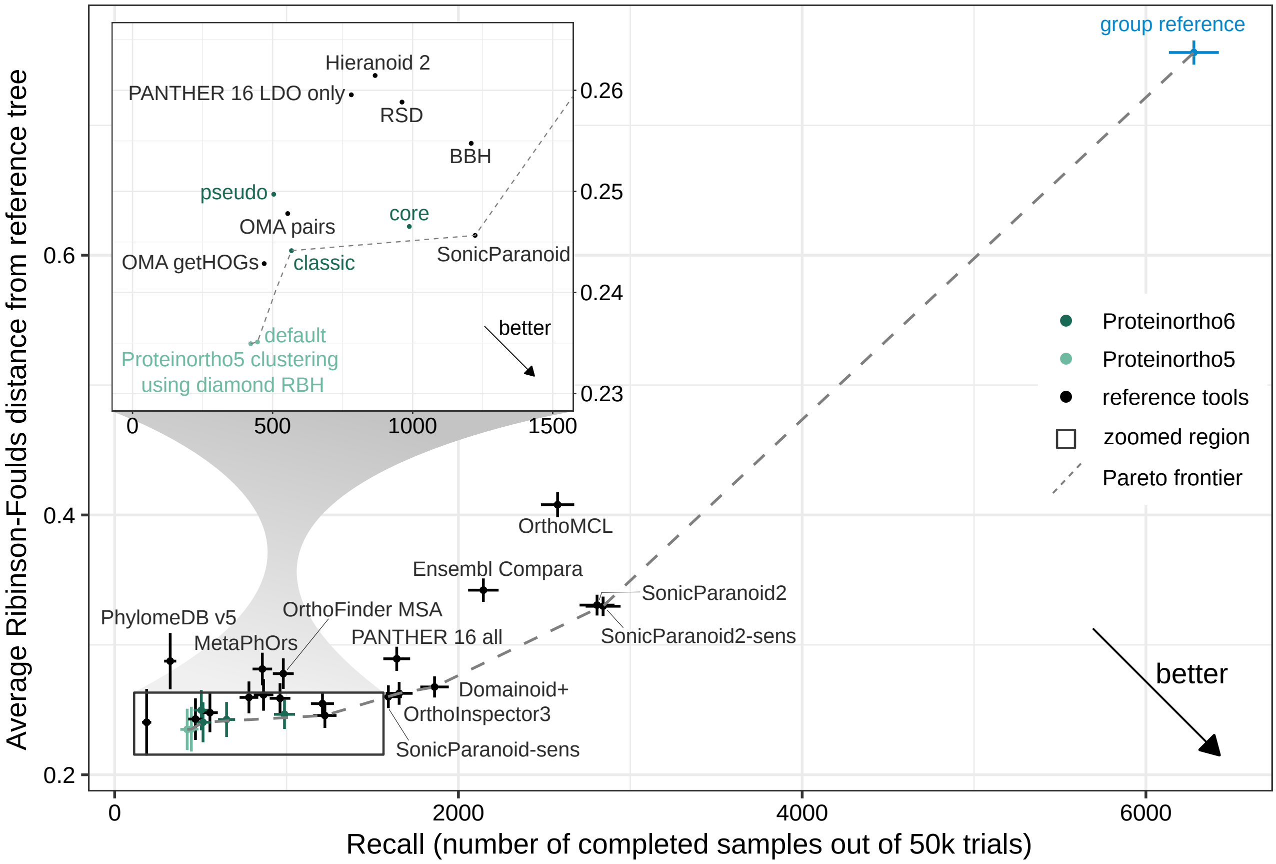


Figure 2. Assessment of Proteinortho and selected orthology tools provided by the `QfO` benchmark service using the 2020/04 dataset. The Generalized Species Tree Discordance benchmark Luca (G-STD2-Luca) with zoomed region. `Proteinortho` provides high precision at the cost of recall when used with default settings and slight variations between the different variants (pseudo, version 5 with `BLAST`, `DIAMOND`, `core`). The blue outlier on the right was generated using `Proteinortho6` with `DIAMOND` and a relaxed clustering step ($\alpha = 0.00001$, group reference).

The orthology prediction results of “OMA Pairs”, “SonicParanoid”, and “SonicParanoid-fast” showed high precision specific to the phylogenetic benchmarks, where at least 6/7 benchmarks are among the top 25%. The largest average differences were found in comparison to “OrthoMCL” (-0.648 `improvement`), “Ensembl Compara” (-0.272 `improvement`) and “SonicParanoid2” (-0.224 `improvement`). Additionally, “OMA Pairs” produces the overall closest results compared to `Proteinortho`. A full assessment of all benchmarks can be found in the Supplemental Data 2.

In the context of sensitivity scores, `Proteinortho` consistently yields some of the lowest scores, as demonstrated in detail in Supplemental Data Section 1. For example, the number of ortholog relations in the function-based Gene Ontology (GO) benchmark, is depicted in Fig. 3. `Proteinortho6` generates approximately 10k orthologs, comparable to that produced by `Proteinortho5` and “OMA pairs”. In contrast, ”SonicParanoid” generates around 20k orthologs, while the highest sensitivity scores are achieved by ”Ensembl Compara” and ”OMA GETHOGs,” which produce between 30k and 40k orthologs.

2.8 Usability

`Proteinotho6` is now readily available across various operating systems through multiple repositories, namely Bioconda (Conda), Homebrew (Brew), and the Debian apt repository. Additionally, a containerized

Table 3. Quantifying Orthology Inference Precision: Assessing Proteinortho and Other Tools Using precision metrics of QfO benchmark dataset 2020/04. Three categories of benchmarks were employed: phylogeny-based benchmarks, function-based benchmarks, and reference orthology-based benchmarks, see the Method section for more details. A full description of all tools and the detailed benchmark results can be found in Supplemental Data 2. Proteinortho parameters are given in the form X+Y, where X specifies variation in the reciprocal best hit algorithm and Y the clustering modus. improvement: average $\log_2$ improvement relative to Proteinortho5 default+default. classic: classic adaptive reciprocal best hit algorithm. $*$: new default configuration of Proteinortho6. group reference: Proteinortho6 with DIAMOND and a relaxed clustering step ($\alpha = 0.00001$). ∇: RBH output of Proteinortho6 using DIAMOND in sensitive mode. 🟩: top 25%, 🟨: top 50% of published tools.

benchmark type	functional		phlyogeny							reference				
metric	avg. Schlicker		avg. Robinson-Foulds							PPV				
benchmark	EC	GO	GSTD2 Eukaryota	GSTD2 Fungi	GSTD2 Luca	GSTD2 Vertebrata	STD Bacteria	STD Eukaryota	STD Fungi	SwissTrees	TreeFam-A	VGNC	# top 25%	improvement
Proteinortho5:														
classic + default	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩		10	0
DIAMOND RBH$^\nabla$ + default	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩		10	0.02
Proteinortho6 with DIAMOND sensitive:														
default + default	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩		10	-0.001
pseudo + default $*$	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩		10	0.003
classic + core	🟩	🟩	🟨	🟩	🟩	🟩	🟨	🟨	🟩	🟩	🟩		8	-0.028
classic without clustering	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟩	🟩	🟩		9	-0.014
classic + flooding	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨	🟩	🟩	🟩		9	-0.017
group reference													0	-1.473
published tools:														
Domainoid+		🟨		🟨				🟨		🟨			0	-0.082
Ensembl Compara													0	-0.272
Hieranoid 2	🟩	🟨	🟩	🟩	🟨	🟩	🟨	🟩	🟩	🟩	🟩	🟩	9	-0.028
MetaPhOrs v.2.5	🟨	🟩					🟩			🟨	🟨	🟨	2	-0.135
OMA GETHOGs			🟨	🟩	🟩	🟩		🟨	🟩				4	-0.059
OMA Pairs	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟨	🟩				7	-0.007
OrthoFinder MSA v2.5.2	🟨	🟨					🟨			🟩			1	-0.125
OrthoInspector 3	🟨	🟨	🟨	🟨	🟨	🟨		🟨			🟩	🟨	1	-0.056
OrthoMCL	🟨												0	-0.648
PANTHER 16 all	🟨									🟨			0	-0.183
phylomedb v5	🟩	🟩	🟨			🟨	🟩		🟨	🟨	🟩	🟨	4	-0.099
RSD	🟩	🟩		🟨	🟨		🟨		🟨		🟩	🟩	4	-0.107
RBH/BBH	🟩	🟩	🟨	🟨	🟩	🟩	🟨	🟨	🟨		🟩	🟩	6	-0.052
SonicParanoid			🟩	🟨	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	8	-0.032
SonicParanoid-fast			🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟨	8	-0.015
SonicParanoid-mostsensitive			🟨		🟨	🟨		🟩	🟨	🟨	🟨	🟩	2	-0.059
SonicParanoid-sens		🟨	🟩	🟩	🟨	🟨	🟨	🟩	🟨	🟩	🟨	🟩	5	-0.043
SonicParanoid2													0	-0.224
SonicParanoid2-sens													0	-0.228

version of Docker can be obtained from quay.io. Proteinotho6 is now actively developed on GitLab fostering collaborative development and providing a transparent platform for community involvement. Furthermore, we implemented continuous integration and continuous deployment (CI/CD) routines through GitLab, ensuring efficient and seamless updates and frequent releases.

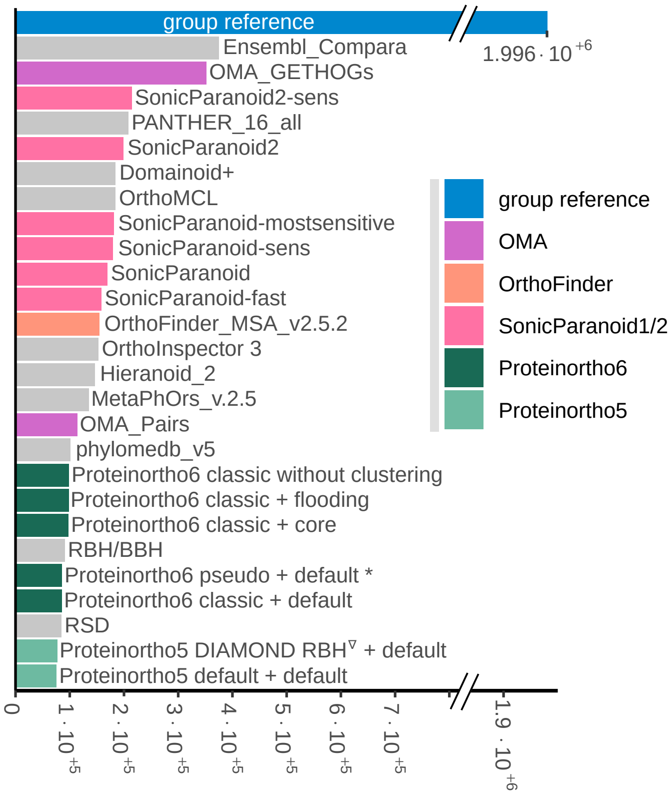


Figure 3. Number of ortholog relations in the function-based GO benchmark. Proteinortho parameters are given in the form X+Y, where X specifies variation in the reciprocal best hit algorithm and Y the clustering modus. classic: classic adaptive reciprocal best hit algorithm. $*$: new default configuration of Proteinortho6. group reference: Proteinortho6 with DIAMOND and a relaxed clustering step ($\alpha = 0.00001$). ∇: RBH output of Proteinortho6 using DIAMOND in sensitive mode.

In order to assist researchers with limited programming experience, a graphical interface has been developed, which facilitates the generation of command lines and allows for the exploration of the output related to the orthology groups. Moreover, Proteinotho6 is now accessible in usegalaxy.eu (tools-iuc), providing a graphical interface and free computing resources for users. For large datasets, Proteinortho6 now includes a convenient interface to deploy jobs to multiple computing nodes in an HPC (High-Performance Computing) environment for Slurm systems. Furthermore, the clustering algorithm of Proteinotho6 is now not limited to Proteinotho output formats and now can be used on any undirected graph in the widespread ABC format.

Proteinortho6 was implemented with a focus on minimizing dependencies to ensure portability and avoid conflicts between multiple installed programs (“dependency hell”). In the Bioconda repository, Proteinotho6 has only 10 direct dependencies, while similar programs such as SonicParanoid2 and OrthoFinder have 15 and 14 dependencies, respectively.

3 DISCUSSION

Proteinortho was designed to predict (co-)ortholog groups, with a focus on large datasets. Previous implementations have been unable to keep up with the deluge of newly sequenced genomes that calls for the analysis of millions of proteins and pairwise best-match graphs with billions of edges. With Proteinortho6, we present a comprehensive algorithmic update for both, the similarity comparisons, and the clustering step.

Based on the detailed evaluation, the sensitive variation of DIAMOND replaces BLAST in the sequence comparison step. This leads to a considerable speedup with an acceptable loss of sensitivity in the initial reciprocal best-hit graph. Proteinortho6 offers the use of all similarity search tools listed above as an alternative. An example is ucsc BLAT. It offers an even higher speedup at the cost of sensitivity. It primarily reports very similar sequences. This might be desirable if the dataset comprises only closely related species. We further explored an improved search strategy for the reciprocal best hit calculations, the pseudo approach. Results proved similar to classic strategies while consistently yielding an additional significant speed up. To optimize the performance, the pseudo option has been selected as the new default *modus operandi*. This method has the potential for broader adoption in other tools in the field.

In the clustering procedure of Proteinortho6, a new strategy is implemented to compute the algebraic connectivity and the associated Fiedler vector using the Fortran library Lapack, which is significantly faster for connected components of larger sizes. The analysis of real-world connected components in combination with artificially generated ones shows the superiority of Lapack's ssyevr approach over the original power iteration in terms of running time. The precision evaluation showed no major changes. A downside, however, is the quadratic memory requirement of ssyevr. Very large connected components are inevitable when analyzing large datasets. Technically these would be workable through the power iteration. However, at the enormous cost of CPU time. Hence, the flooding heuristic was introduced. The reworked clustering implementation also makes efficient use of multiple CPU cores and can even be distributed among multiple computing nodes.

A regular application of orthology tools is the calculation of robust phylogenetic reconstructions via a supertree analysis based on single-copy orthologs among a given set of species. The new adaptive clustering facilitates better results in this context as it automatically optimizes the clustering parameters for each group to cover as many species as possible without overestimating the amount of paralogs. Besides this specific research question, core falls behind the default clustering approach in terms of precision and thus is not chosen as the default.

For the comparison with other orthology prediction tools and databases, the standardized QfO benchmark system was used. Despite the usefulness of the benchmark system, we encountered some shortcomings that may affect the comparisons. In particular, the recall metric of the system is biased towards large inputs. Execution parameters and tool versions are typically not documented. Nevertheless, the precision estimates provided by QfO gave valuable insights regarding changes in the quality of our predictions when introducing alternative algorithms. Results generated by Proteinortho are consistently among the highest-performing tools in terms of precision and archived scores are generally close to the results of OMA. In terms of sensitivity, Proteinortho produces among the lowest scores compared to the other tools, highlighting a distinct trade-off. Proteinortho6 notably excels in terms of execution time and provides a considerable speedup over its previous implementation. This substantially increases the size of datasets that can be processed and makes efficient use of the hardware provided. This is especially notable in comparison to OMA.

Author Contributions

ML conceived the study. ML supervised the project and drafted the manuscript. PK carried out the bioinformatic analyses. PK, and ML evaluated and verified the results. PK, and ML revised the manuscript. All authors wrote, read, and approved the final manuscript.

FUNDING

ACKNOWLEDGMENTS

We thank Clemens Thölken and Fareha Masood for fruitful discussions. Computationally intensive calculations were possible thanks to the HPC cluster MaRC3 located at the Philipps-University Marburg.

SUPPLEMENTAL DATA

Supplemental Data 1: Additional figures and results (PDF)

- Cluster Algorithm Overview
- Scalability
- Tab. 1 with alternative datasets
- Sensitivity Assessment
- Small World Phenomenon Example
- QfO Evaluation All Plots
- E-value Comparison

Supplemental Data 2: Listings, including datasets, proteome identifier, execution times, and raw data (XLSX)

- `datasets overview` : sizes of all datasets used
- `scalability uniprot_n` : raw data of Fig. 1
- `clustering evaluation` : raw data of Tab. 2
- `sensitivity, precision of BLAST alternatives` : raw data of Tab. 1
- `qfo 2020_04 benchmark results` : raw data of Tab. 3
- `power iteration vs lapack for 1 thread` : raw data of the analysis in section 2.2
- `shuffled Bac_n 2022_03` : proteome identifier of the Bac$_n$ dataset
- `BigCC dataset` : proteome identifier of the `BigCC` dataset
- `QfO_release_2020_04_with_updated_UP000008143 dataset` : proteome identifier of the `QfO` 2020/04 dataset

Data Availability Statement

Project name: `Proteinortho6`
Project home page: `www.bioinf.uni-leipzig.de/Software/proteinortho/`
Operating system(s): Linux, Mac

Programming language: Perl, C++
License: GNU GPLv3
Any restrictions to use by non-academics: none

Abbreviation

l_2FC : $\log_2$ fold-change, RBH: reciprocal best hit graph, RF: Robinson-Foulds, GSTD: generalized species tree discordance test, PPV: positive predictive value, h: hours, GB: gigabyte, ARI: adjusted rand index, k: thousand, QfO: Quest for Orthologs

REFERENCES

[Dataset] Altenhoff, A. (2023). benchmark-webservice. `https://github.com/qfo/benchmark-webservice`

Altenhoff, A. M., Boeckmann, B., Capella-Gutierrez, S., Dalquen, D. A., DeLuca, T., Forslund, K., et al. (2016). Standardized benchmarking in the quest for orthologs. *Nature methods* 13, 425–430. doi: 10.1038/nmeth.3830

Altenhoff, A. M. and Dessimoz, C. (2009). Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLoS Comput Biol* 5, e1000262. doi: 10.1371/journal.pcbi.1000262

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., et al. (2019). Oma standalone: orthology inference among public and custom genomes and transcriptomes. *Genome research* 29, 1152–1163. doi: 10.1101/gr.243212.118.

Anderson, E., Bai, Z., Bischof, C., Blackford, L. S., Demmel, J., Dongarra, J., et al. (1999). *LAPACK Users' guide*, vol. 9 (Philadelphia, PA, USA: SIAM). doi: 10.1137/1.9780898719604

Arshinoff, B. I., Cary, G. A., Karimi, K., Foley, S., Agalakov, S., Delgado, F., et al. (2022). Echinobase: leveraging an extant model organism database to build a knowledgebase supporting research on the genomics and biology of echinoderms. *Nucleic acids research* 50, D970–D979. doi: 10.1093/nar/gkab1005

Bientinesi, P., Dhillon, I. S., and Van De Geijn, R. A. (2005). A parallel eigensolver for dense symmetric matrices based on multiple relatively robust representations. *SIAM Journal on Scientific Computing* 27, 43–66. doi: 10.1137/030601107

Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J Mol Biol* 283, 707–725. doi: 10.1006/jmbi.1998.2144

Boutsidis, C., Kambadur, P., and Gittens, A. (2015). Spectral clustering via the power method-provably. In *International Conference on Machine Learning*. 40–48

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods* 12, 59–60. doi: 10.1038/nmeth.3176

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). Blast+: architecture and applications. *BMC bioinformatics* 10, 421. doi: 10.1186/1471-2105-10-421

[Dataset] Consentino, S. (2021). Participant dataset submitted by Sonicparanoid-sens. doi: 10.23728/B2SHARE.49EA3C4298624330A5866E8A7E364238

Cosentino, S. and Iwasaki, W. (2023). Sonicparanoid2: fast, accurate, and comprehensive orthology inference with machine learning and language models. *bioRxiv* 2023–05. doi: 10.1101/2023.05.14.540736

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Emms, D. M. and Kelly, S. (2019). Orthofinder: phylogenetic orthology inference for comparative genomics. *Genome biology* 20, 1–14. doi: 10.1186/s13059-019-1832-y

Fiedler, M. (1975). A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Math. J.* doi: 10.21136/CMJ.1975.101357

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst Zool* 19, 99–113. doi: 10.2307/2412448

[Dataset] Hickman, M. (2021). Participant dataset submitted by OrthoMCL. doi: 10.23728/B2SHARE.B281E1797C4A409ABCA1791E09902C59

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification* 2, 193–218. doi: 10.1007/BF01908075

Johnson, T. (2007). Reciprocal best hits are not a logically sufficient condition for orthology. *arXiv preprint arXiv:0706.0117* doi: 10.48550/arXiv.0706.0117

Kapheim, K. M., Pan, H., Li, C., Salzberg, S. L., Puiu, D., Magoc, T., et al. (2015). Genomic signatures of evolutionary transitions from solitary to group living. *Science* 348, 1139–1143. doi: 10.1126/science.aaa4788

Kent, W. J. (2002). Blat–the BLAST-like alignment tool. *Genome research* 12, 656–664. doi: 10.1101/gr.229202

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome research* 21, 487–493. doi: 10.1101/gr.113985.110

Klemm, P., Christ, M., Altegoer, F., Freitag, J., Bange, G., and Lechner, M. (2022). Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family. *Frontiers in plant science* 13, 4832. doi: 10.3389/fpls.2022.1034708

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39, 309–38. doi: 10.1146/annurev.genet.39.073003.114725

Koonin, E. V., Wolf, Y. I., and Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature* 420, 218–223. doi: 10.1038/nature01256

Koskinen, J. P. and Holm, L. (2012). Sans: high-throughput retrieval of protein sequences allowing 50% mismatches. *Bioinformatics (Oxford, England)* 28, i438–i443. doi: 10.1093/bioinformatics/bts417

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-) orthologs in large-scale analysis. *BMC bioinformatics* 12, 1–9. doi: 10.1186/1471-2105-12-124

Lynch, M. and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151–1155. doi: 10.1126/science.290.5494.1151

Medlar, A. and Holm, L. (2018). Topaz: asymmetric suffix array neighbourhood search for massive protein databases. *BMC Bioinformatics* 19, 278. doi: 10.1186/s12859-018-2290-3

Milgram, S. (1967). The small world problem. *Psychology Today* , 61–67

Ohno, S. (1999). Gene duplication and the uniqueness of vertebrate genomes circa 1970-1999. *Seminars in Cell and Developmental Biology* 10, 517–522. doi: 10.1006/scdb.1999.0332

[Dataset] Palmer, J. M. and Stajich, J. E. (2023). Funannotate

Parlett, B. N. and Dhillon, I. S. (2000). Relatively robust representations of symmetric tridiagonals. *Linear Algebra and its applications* 309, 121–151. doi: 10.1016/S0024-3795(99)00262-1

Peter, J., De Chiara, M., Friedrich, A., Yue, J.-X., Pflieger, D., Bergström, A., et al. (2018). Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* 556, 339–344. doi: 10.1038/s41586-018-0030-5

Pinho, M. G., Kjos, M., and Veening, J.-W. (2013). How to get (a) round: mechanisms controlling growth and division of coccoid bacteria. *Nature reviews microbiology* 11, 601–614. doi: 10.1038/nrmicro3088

Supplementary Material

1 CLUSTER ALGORITHM OVERVIEW

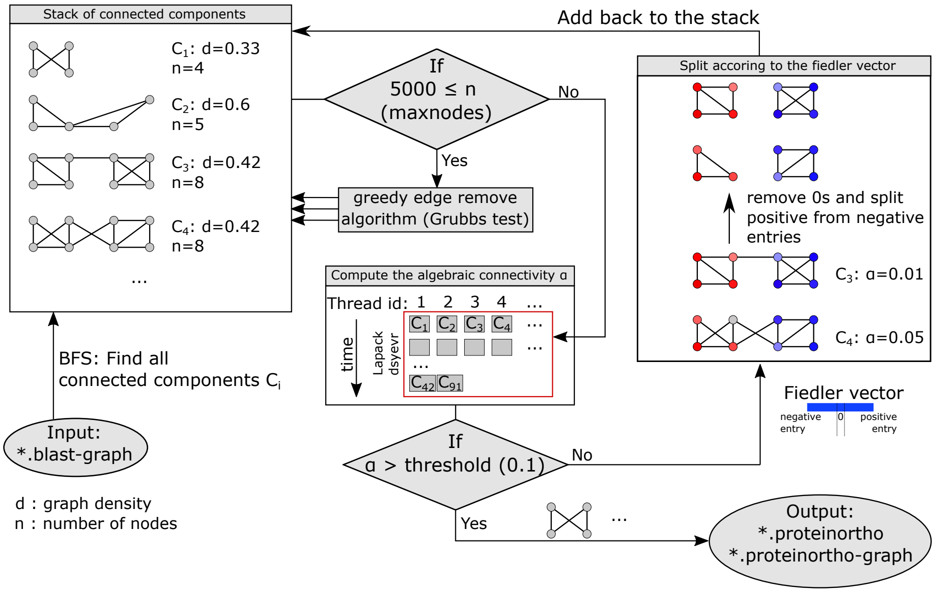


Figure S1: Updated multi-threading system for the clustering step. First, all connected components are identified in the input graph using the branch-first search algorithm (BFS). Suitable small components are processed in parallel using `Lapack ssyevr`. The remaining larger components are processed using the greedy split algorithm. Resulting components with insufficient algebraic connectivity are split according to the associated Fiedler vector and marked for an additional round of processing.

2 SCALABILITY

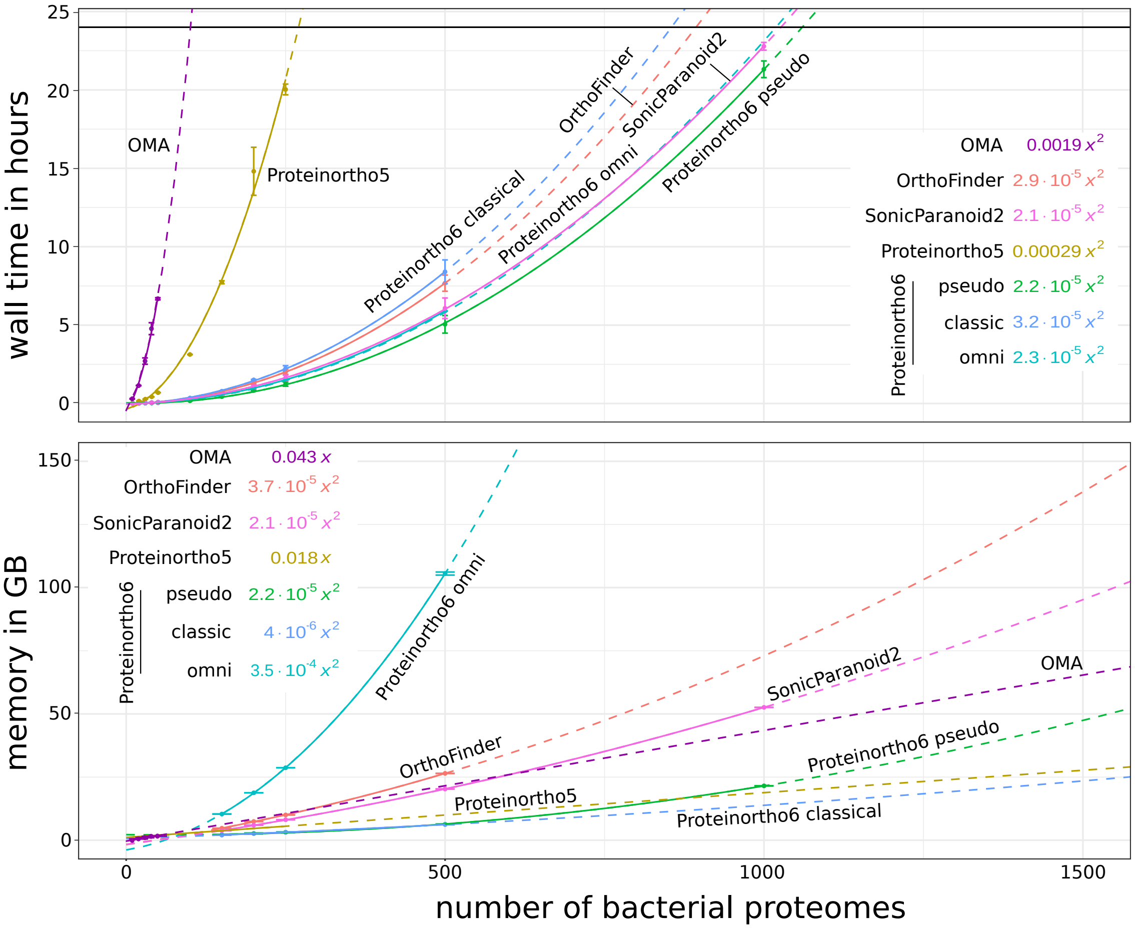
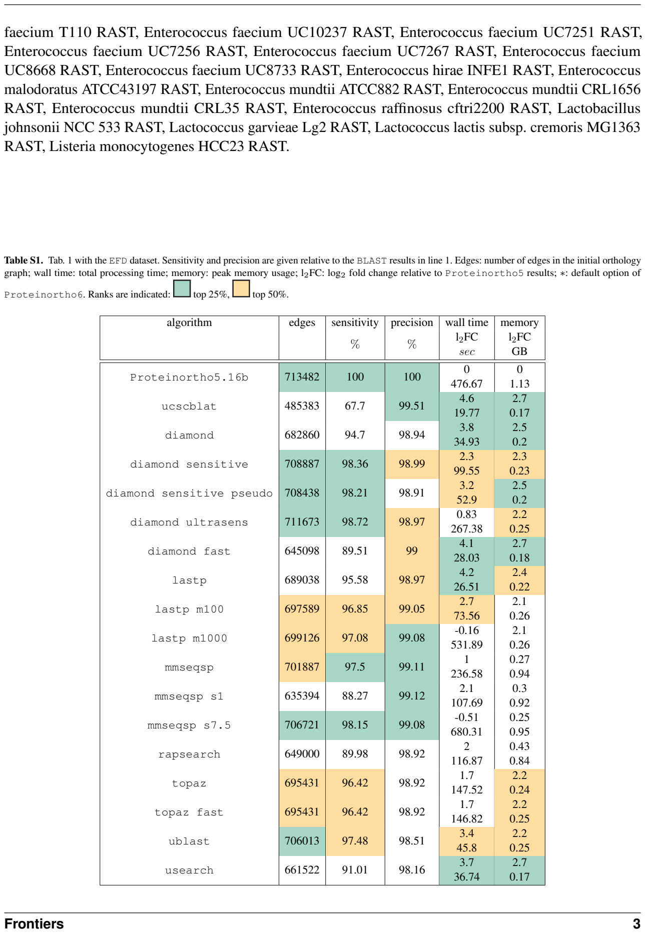


Figure S2: Scalability of total orthology prediction, including the all-versus-all sequence comparison and clustering, relative to dataset size of randomly selected bacterial proteomes of UniProt 2022_03 ($\texttt{Bac}_{10,20,\ldots,1000}$). Average processing times and peak memory consumption are indicated by circles and fitted using a quadratic function (solid line, $R^2_{adj} \geq 0.99$ for wall time and $R^2_{adj} \geq 0.89$ for memory consumption) for extrapolation (dashed lines). The peak memory consumption was restricted to a dataset of size ≥ 150. Because of a negative quadratic term, the memory consumption of `Proteinortho5` and `OMA` was fitted using a linear function instead. Coefficients of the term with the highest degree are indicated for each tool. Details on parameters and versions can be found in the Supplemental Table

3 TAB. 1 ALTERNATIVE DATASETS

3.1 EFD

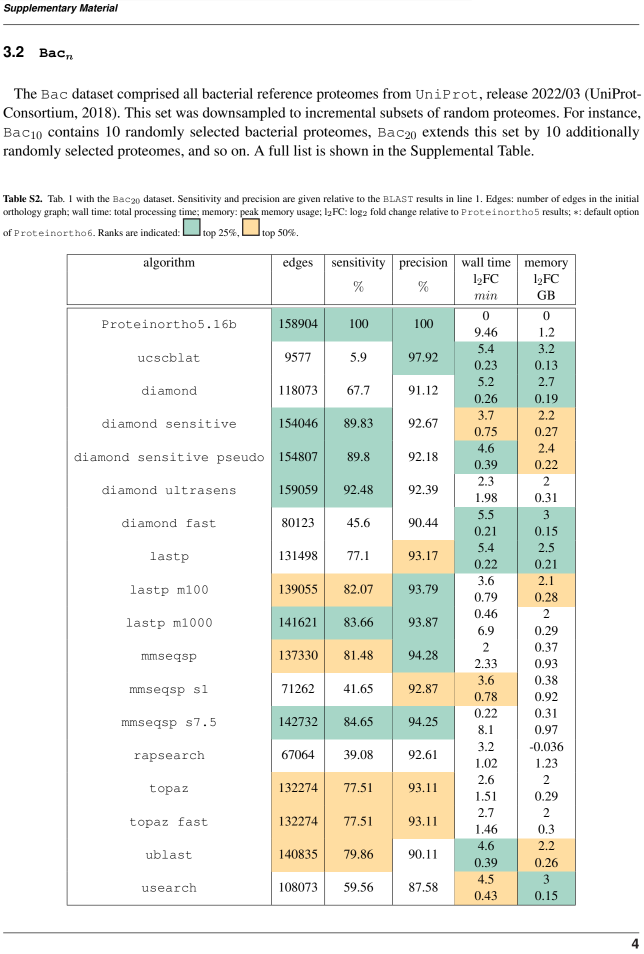
`EFD` is a dataset of 29 food-related and probiotic strains of the Lactobacillus genus Bonacina et al. (2017). It represents a small set of very similar species: Enterococcus durans IPLA655 RAST, Enterococcus faecalis 19116 RAST, Enterococcus faecalis 2924 RAST, Enterococcus faecalis MB5259 RAST, Enterococcus faecalis PC1.1 RAST, Enterococcus faecalis str. Symbioflor 1 RAST, Enterococcus faecium CRL1879 RAST, Enterococcus faecium E1604 RAST, Enterococcus faecium E1613 RAST, Enterococcus faecium L-3 RAST, Enterococcus faecium L-X RAST, Enterococcus faecium NRRLB-2354 RAST, Enterococcus



faecium T110 RAST, Enterococcus faecium UC10237 RAST, Enterococcus faecium UC7251 RAST, Enterococcus faecium UC7256 RAST, Enterococcus faecium UC7267 RAST, Enterococcus faecium UC8668 RAST, Enterococcus faecium UC8733 RAST, Enterococcus hirae INFE1 RAST, Enterococcus malodoratus ATCC43197 RAST, Enterococcus mundtii ATCC882 RAST, Enterococcus mundtii CRL1656 RAST, Enterococcus mundtii CRL35 RAST, Enterococcus raffinosus cftri2200 RAST, Lactobacillus johnsonii NCC 533 RAST, Lactococcus garvieae Lg2 RAST, Lactococcus lactis subsp. cremoris MG1363 RAST, Listeria monocytogenes HCC23 RAST.

Table S1. Tab. 1 with the EFD dataset. Sensitivity and precision are given relative to the BLAST results in line 1. Edges: number of edges in the initial orthology graph; wall time: total processing time; memory: peak memory usage; l_2FC: $\log_2$ fold change relative to Proteinortho5 results; $*$: default option of Proteinortho6. Ranks are indicated: 🟩 top 25%, 🟨 top 50%.

algorithm	edges	sensitivity	precision	wall time	memory
		%	%	l_2FC *sec*	l_2FC GB
Proteinortho5.16b	713482	100	100	0 476.67	0 1.13
ucscblat	485383	67.7	99.51	4.6 19.77	2.7 0.17
diamond	682860	94.7	98.94	3.8 34.93	2.5 0.2
diamond sensitive	708887	98.36	98.99	2.3 99.55	2.3 0.23
diamond sensitive pseudo	708438	98.21	98.91	3.2 52.9	2.5 0.2
diamond ultrasens	711673	98.72	98.97	0.83 267.38	2.2 0.25
diamond fast	645098	89.51	99	4.1 28.03	2.7 0.18
lastp	689038	95.58	98.97	4.2 26.51	2.4 0.22
lastp m100	697589	96.85	99.05	2.7 73.56	2.1 0.26
lastp m1000	699126	97.08	99.08	-0.16 531.89	2.1 0.26
mmseqsp	701887	97.5	99.11	1 236.58	0.27 0.94
mmseqsp s1	635394	88.27	99.12	2.1 107.69	0.3 0.92
mmseqsp s7.5	706721	98.15	99.08	-0.51 680.31	0.25 0.95
rapsearch	649000	89.98	98.92	2 116.87	0.43 0.84
topaz	695431	96.42	98.92	1.7 147.52	2.2 0.24
topaz fast	695431	96.42	98.92	1.7 146.82	2.2 0.25
ublast	706013	97.48	98.51	3.4 45.8	2.2 0.25
usearch	661522	91.01	98.16	3.7 36.74	2.7 0.17



3.2 Bac$_n$

The Bac dataset comprised all bacterial reference proteomes from UniProt, release 2022/03 (UniProt-Consortium, 2018). This set was downsampled to incremental subsets of random proteomes. For instance, Bac$_{10}$ contains 10 randomly selected bacterial proteomes, Bac$_{20}$ extends this set by 10 additionally randomly selected proteomes, and so on. A full list is shown in the Supplemental Table.

Table S2. Tab. 1 with the Bac$_{20}$ dataset. Sensitivity and precision are given relative to the BLAST results in line 1. Edges: number of edges in the initial orthology graph; wall time: total processing time; memory: peak memory usage; l_2FC: $\log_2$ fold change relative to Proteinortho5 results; $*$: default option of Proteinortho6. Ranks are indicated: 🟩 top 25%, 🟨 top 50%.

algorithm	edges	sensitivity	precision	wall time	memory
		%	%	l_2FC *min*	l_2FC GB
Proteinortho5.16b	158904	100	100	0 9.46	0 1.2
ucscblat	9577	5.9	97.92	5.4 0.23	3.2 0.13
diamond	118073	67.7	91.12	5.2 0.26	2.7 0.19
diamond sensitive	154046	89.83	92.67	3.7 0.75	2.2 0.27
diamond sensitive pseudo	154807	89.8	92.18	4.6 0.39	2.4 0.22
diamond ultrasens	159059	92.48	92.39	2.3 1.98	2 0.31
diamond fast	80123	45.6	90.44	5.5 0.21	3 0.15
lastp	131498	77.1	93.17	5.4 0.22	2.5 0.21
lastp m100	139055	82.07	93.79	3.6 0.79	2.1 0.28
lastp m1000	141621	83.66	93.87	0.46 6.9	2 0.29
mmseqsp	137330	81.48	94.28	2 2.33	0.37 0.93
mmseqsp s1	71262	41.65	92.87	3.6 0.78	0.38 0.92
mmseqsp s7.5	142732	84.65	94.25	0.22 8.1	0.31 0.97
rapsearch	67064	39.08	92.61	3.2 1.02	-0.036 1.23
topaz	132274	77.51	93.11	2.6 1.51	2 0.29
topaz fast	132274	77.51	93.11	2.7 1.46	2 0.3
ublast	140835	79.86	90.11	4.6 0.39	2.2 0.26
usearch	108073	59.56	87.58	4.5 0.43	3 0.15

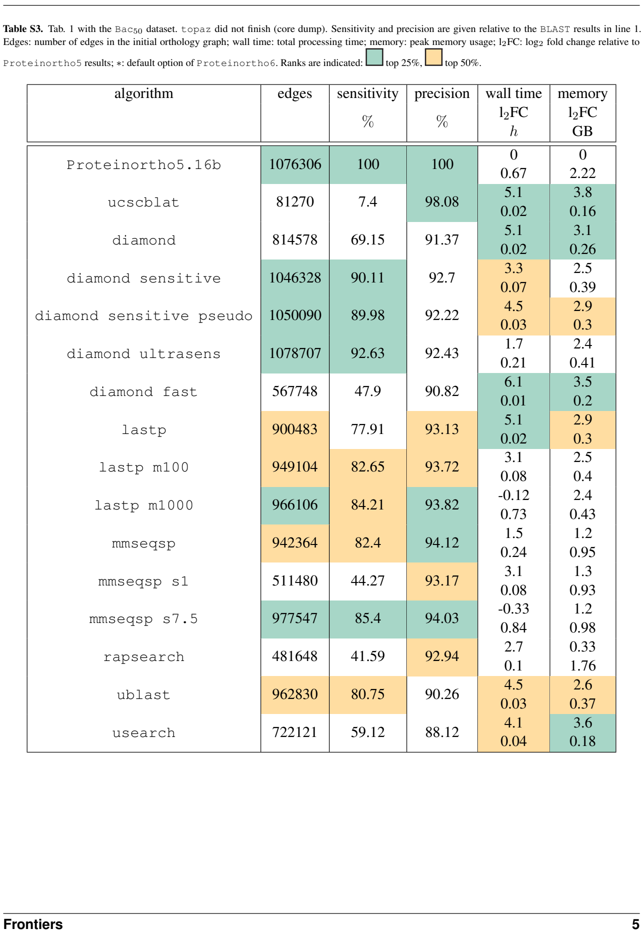


Table S3. Tab. 1 with the Bac$_{50}$ dataset. topaz did not finish (core dump). Sensitivity and precision are given relative to the BLAST results in line 1. Edges: number of edges in the initial orthology graph; wall time: total processing time; memory: peak memory usage; l_2FC: $\log_2$ fold change relative to Proteinortho5 results; $*$: default option of Proteinortho6. Ranks are indicated: 🟩 top 25%, 🟨 top 50%.

algorithm	edges	sensitivity	precision	wall time	memory
		%	%	l_2FC *h*	l_2FC GB
Proteinortho5.16b	1076306	100	100	0 0.67	0 2.22
ucscblat	81270	7.4	98.08	5.1 0.02	3.8 0.16
diamond	814578	69.15	91.37	5.1 0.02	3.1 0.26
diamond sensitive	1046328	90.11	92.7	3.3 0.07	2.5 0.39
diamond sensitive pseudo	1050090	89.98	92.22	4.5 0.03	2.9 0.3
diamond ultrasens	1078707	92.63	92.43	1.7 0.21	2.4 0.41
diamond fast	567748	47.9	90.82	6.1 0.01	3.5 0.2
lastp	900483	77.91	93.13	5.1 0.02	2.9 0.3
lastp m100	949104	82.65	93.72	3.1 0.08	2.5 0.4
lastp m1000	966106	84.21	93.82	-0.12 0.73	2.4 0.43
mmseqsp	942364	82.4	94.12	1.5 0.24	1.2 0.95
mmseqsp s1	511480	44.27	93.17	3.1 0.08	1.3 0.93
mmseqsp s7.5	977547	85.4	94.03	-0.33 0.84	1.2 0.98
rapsearch	481648	41.59	92.94	2.7 0.1	0.33 1.76
ublast	962830	80.75	90.26	4.5 0.03	2.6 0.37
usearch	722121	59.12	88.12	4.1 0.04	3.6 0.18

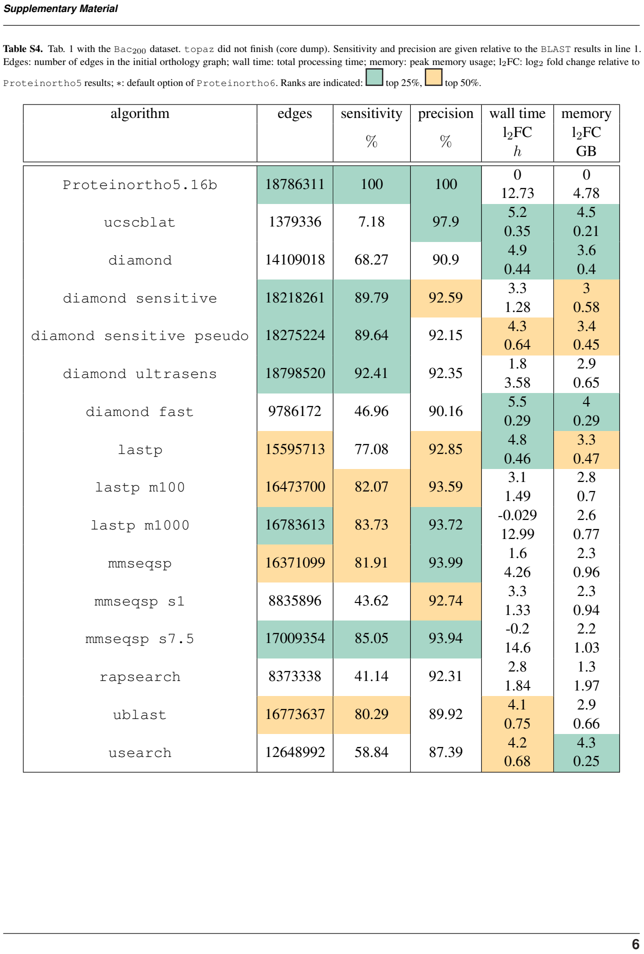


Table S4. Tab. 1 with the Bac$_{200}$ dataset. topaz did not finish (core dump). Sensitivity and precision are given relative to the BLAST results in line 1. Edges: number of edges in the initial orthology graph; wall time: total processing time; memory: peak memory usage; l_2FC: $\log_2$ fold change relative to Proteinortho5 results; $*$: default option of Proteinortho6. Ranks are indicated: 🟩 top 25%, 🟨 top 50%.

algorithm	edges	sensitivity	precision	wall time	memory
		%	%	l_2FC *h*	l_2FC GB
Proteinortho5.16b	18786311	100	100	0 12.73	0 4.78
ucscblat	1379336	7.18	97.9	5.2 0.35	4.5 0.21
diamond	14109018	68.27	90.9	4.9 0.44	3.6 0.4
diamond sensitive	18218261	89.79	92.59	3.3 1.28	3 0.58
diamond sensitive pseudo	18275224	89.64	92.15	4.3 0.64	3.4 0.45
diamond ultrasens	18798520	92.41	92.35	1.8 3.58	2.9 0.65
diamond fast	9786172	46.96	90.16	5.5 0.29	4 0.29
lastp	15595713	77.08	92.85	4.8 0.46	3.3 0.47
lastp m100	16473700	82.07	93.59	3.1 1.49	2.8 0.7
lastp m1000	16783613	83.73	93.72	-0.029 12.99	2.6 0.77
mmseqsp	16371099	81.91	93.99	1.6 4.26	2.3 0.96
mmseqsp s1	8835896	43.62	92.74	3.3 1.33	2.3 0.94
mmseqsp s7.5	17009354	85.05	93.94	-0.2 14.6	2.2 1.03
rapsearch	8373338	41.14	92.31	2.8 1.84	1.3 1.97
ublast	16773637	80.29	89.92	4.1 0.75	2.9 0.66
usearch	12648992	58.84	87.39	4.2 0.68	4.3 0.25

4 SENSITIVITY ASSESSMENT

Table S5. Quantifying Orthology Inference Sensitivity: Assessing Proteinortho and Other Tools Using sensitivity metrics of QfO benchmark dataset 2020/04. Three categories of benchmarks were employed: phylogeny-based benchmarks, function-based benchmarks, and reference orthology-based benchmarks, see the Method section for more details. A full description of all tools and the detailed benchmark results can be found in Supplemental Table. `improvement`: average $\log_2$ improvement relative to `Proteinortho5 default+default`. `Proteinortho` parameters are given in the form X+Y, where X specifies variation in the reciprocal best hit algorithm and Y the clustering modus. `classic`: classic adaptive reciprocal best hit algorithm. $*$: new default configuration of `Proteinortho6`. group reference: `Proteinortho6` with `DIAMOND` and a relaxed clustering step ($\alpha = 0.00001$). ∇: RBH output of `Proteinortho6` using `DIAMOND` in `sensitive` mode. TPR: true positive rate. num: number of orthologs. 🟩: top 25%, 🟨: top 50% of published tools.

benchmark type	functional	phlyogeny								reference				
metric	num		completed tree samples								TPR			
benchmark	EC	GO	GSTD2 Eukaryota	GSTD2 Fungi	GSTD2 Luca	GSTD2 Vertebrata	STD Bacteria	STD Eukaryota	STD Fungi	SwissTrees	TreeFam-A	VGNC	# top 25%	`improvement`
`Proteinortho5`:														
`default` + `default`													0	0
`DIAMOND` RBH$^\nabla$ + `default`													0	0.035
`Proteinortho6` with `DIAMOND sensitive`:														
`default` + `default`													0	0.251
`classic` + `core`													0	0.481
`pseudo` + `default` $*$													0	0.246
`classic` without clustering													0	0.495
`classic` + `flooding`													0	0.482
group reference	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩		10	2.047
published tools:														
Domainoid+	🟨	🟨	🟨		🟩	🟨	🟩	🟩	🟨	🟩	🟨	🟩	5	1.013
Ensembl Compara	🟩	🟩	🟩	🟨	🟩	🟨	🟨	🟩	🟨				5	1.172
Hieranoid 2													0	0.724
MetaPhOrs v.2.5				🟨		🟩		🟩	🟨	🟩			3	0.791
OMA GETHOGs	🟩	🟩					🟨						2	0.618
OMA Pairs													0	0.415
OrthoFinder MSA v2.5.2			🟨	🟩		🟩		🟨	🟩	🟩	🟩	🟩	6	0.959
OrthoInspector 3	🟨		🟨	🟨	🟨		🟨		🟨	🟨	🟨	🟨	0	0.944
OrthoMCL	🟩	🟨	🟩	🟩	🟩	🟩	🟩		🟩	🟨	🟩	🟩	9	1.172
PANTHER 16 all	🟨	🟩	🟩	🟩	🟨	🟨	🟩	🟩	🟩	🟩	🟩		8	1.066
phylomedb v5								🟨					0	0.446
RSD													0	0.524
RBH/BBH												🟩	1	0.634
SonicParanoid		🟨			🟨						🟨	🟨	0	0.804
SonicParanoid-fast												🟨	0	0.632
SonicParanoid-mostsensitive	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	0	0.975
SonicParanoid-sens	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨		🟨	🟨	🟨	0	0.938
SonicParanoid2	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	🟩		10	1.176
SonicParanoid2-sens	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟨	🟩	🟩	11	1.207

5 SMALL WORLD PHENOMENON

With rising numbers of species, the connected components tend to expand quickly, leading to the formation of extensive connected components. Proteinortho v6.3.0 with default parameters using diamond (v2.0.15) but without the clustering step was used to process randomly selected bacterial proteomes $\mathtt{Bac}_n$ until a size of $n = 1000$ species and the BigCC dataset with 1800 species. From the output, the largest connected component is determined and put in relation to the total number of nodes in the graph. The resulting growth is illustrated in Fig. S3.

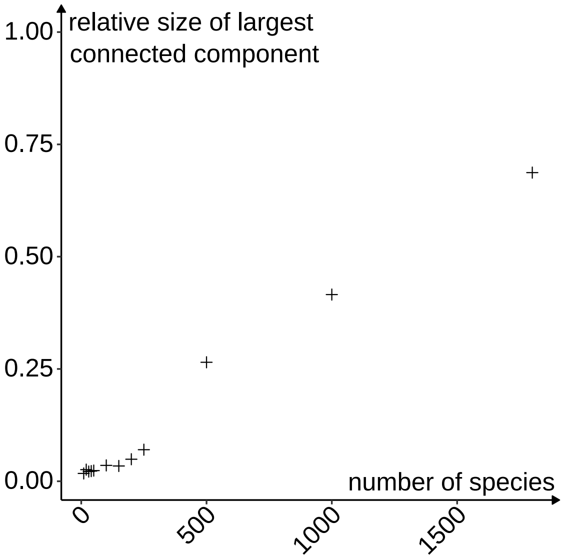
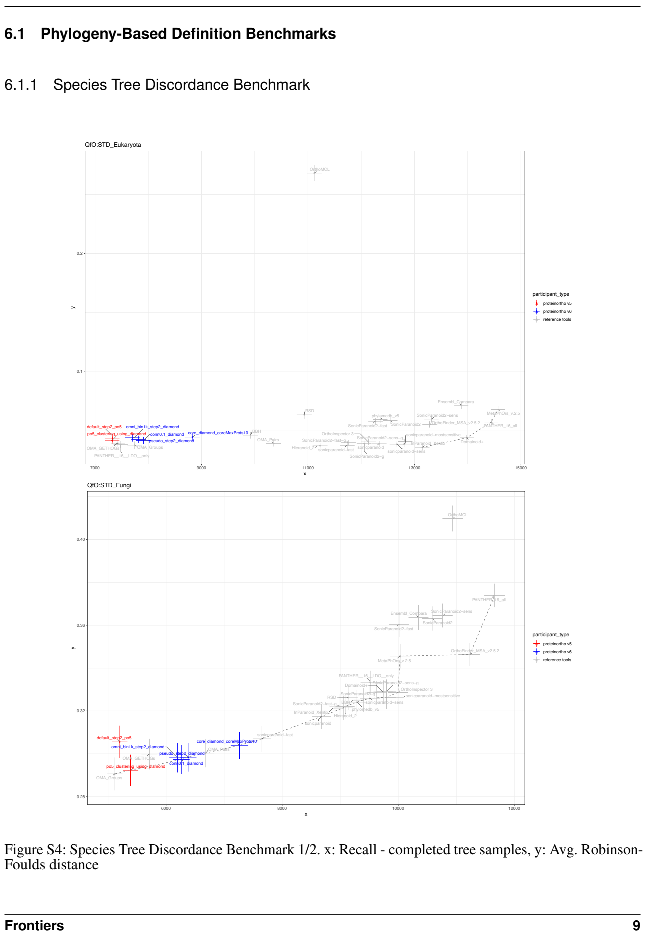


Figure S3: The size of the largest connected components relative to the total number of nodes from randomly selected bacterial proteomes of UniProt until a size of 1000 and the BigCC dataset with 1800 species. The graphs were built using Proteinortho with default parameters.

6 QFO EVALUATION

The following plots show all QfO benchmark results of the 2020_{20} dataset (2020.2) using the following configurations of Proteinortho:

1. default_step2_po5: Proteinortho5 with default settings
2. po5_clustering_using_diamond: Proteinortho5 with default clustering with an input graph that was generated using Proteinortho6 with diamond with default parameters.
3. omni_bin1k_step2_diamond : Proteinortho6 in omni modus using bin=1k (bin size) and diamond in sensitive modus.
4. pseudo_step2_diamond : Proteinortho6 in pseudo modus using diamond in sensitive modus.
5. conn0.1_diamond : Proteinortho6 in canonical modus (canonical reciprocal best hit algorithm) using diamond in sensitive modus.
6. core_diamond_coreMaxProts10 : Proteinortho6 in canonical modus (canonical reciprocal best hit algorithm) using diamond in sensitive modus and the clustering modus core with the parameter coreMaxProts=10 (maximal number of proteins of groups per species)



6.1 Phylogeny-Based Definition Benchmarks

6.1.1 Species Tree Discordance Benchmark

Figure S4: Species Tree Discordance Benchmark 1/2. x: Recall - completed tree samples, y: Avg. Robinson-Foulds distance

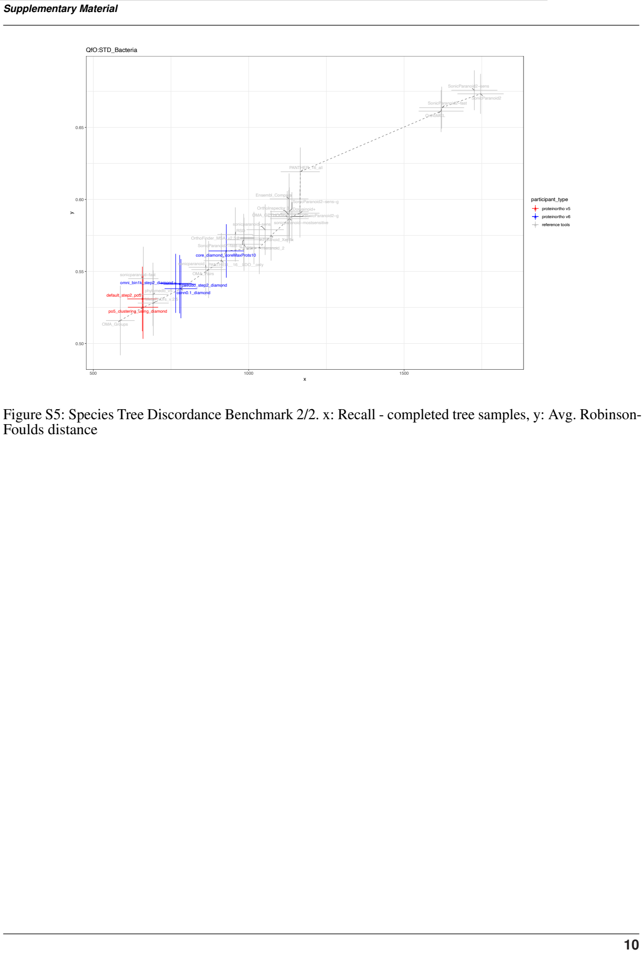
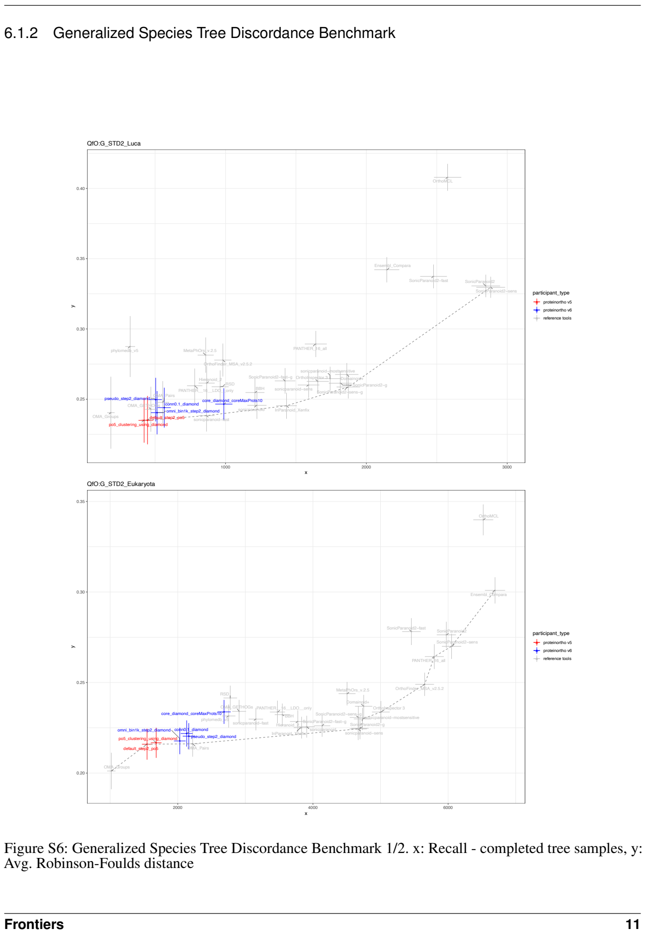


Figure S5: Species Tree Discordance Benchmark 2/2. x: Recall - completed tree samples, y: Avg. Robinson-Foulds distance



6.1.2 Generalized Species Tree Discordance Benchmark

Figure S6: Generalized Species Tree Discordance Benchmark 1/2. x: Recall - completed tree samples, y: Avg. Robinson-Foulds distance

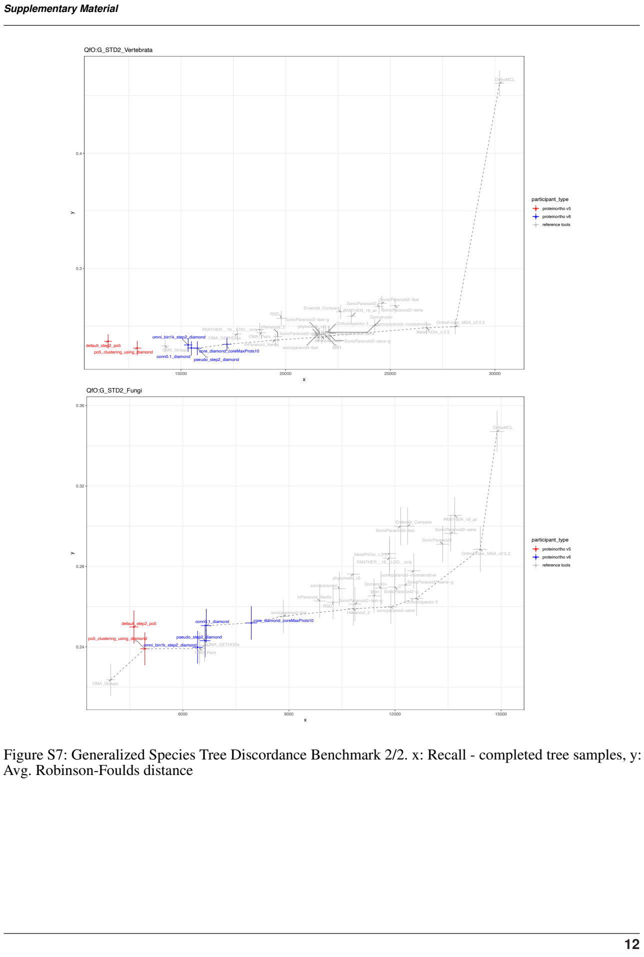
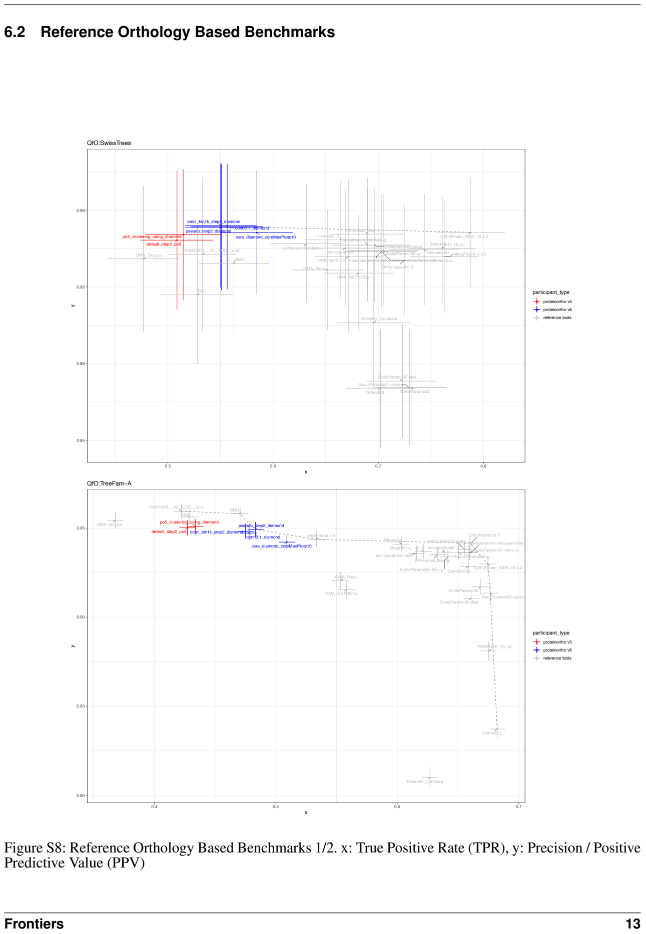


Figure S7: Generalized Species Tree Discordance Benchmark 2/2. x: Recall - completed tree samples, y: Avg. Robinson-Foulds distance

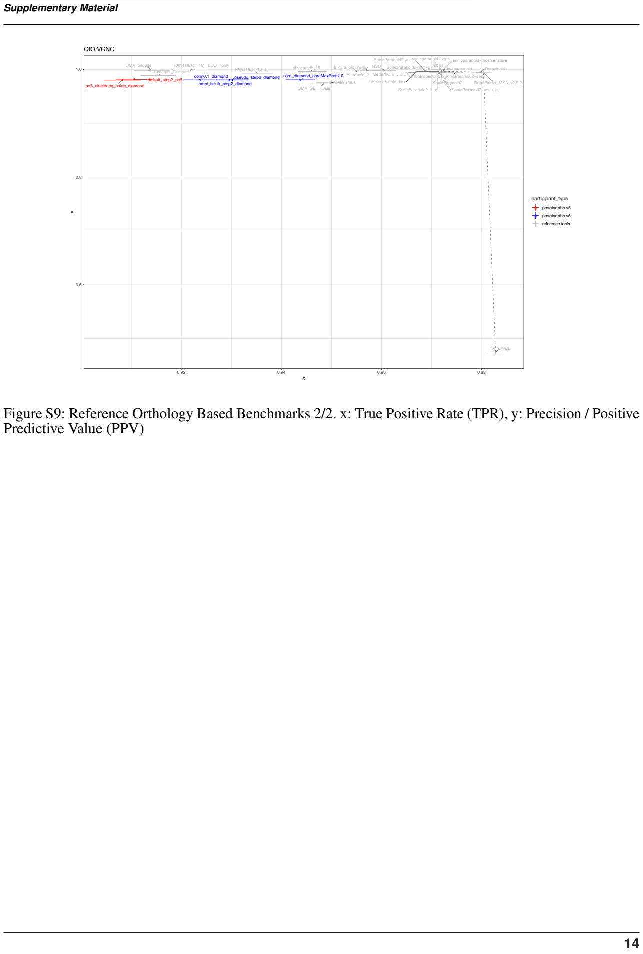


6.2 Reference Orthology Based Benchmarks

QfO:SwissTrees

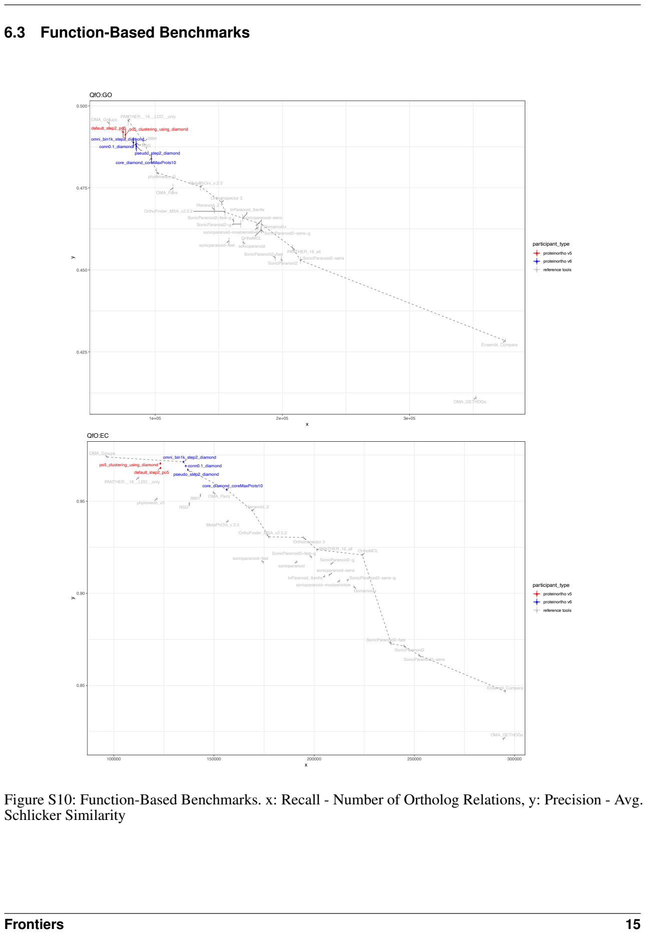
QfO:TreeFam-A

Figure S8: Reference Orthology Based Benchmarks 1/2. x: True Positive Rate (TPR), y: Precision / Positive Predictive Value (PPV)



QfO:VGNC

Figure S9: Reference Orthology Based Benchmarks 2/2. x: True Positive Rate (TPR), y: Precision / Positive Predictive Value (PPV)



6.3 Function-Based Benchmarks

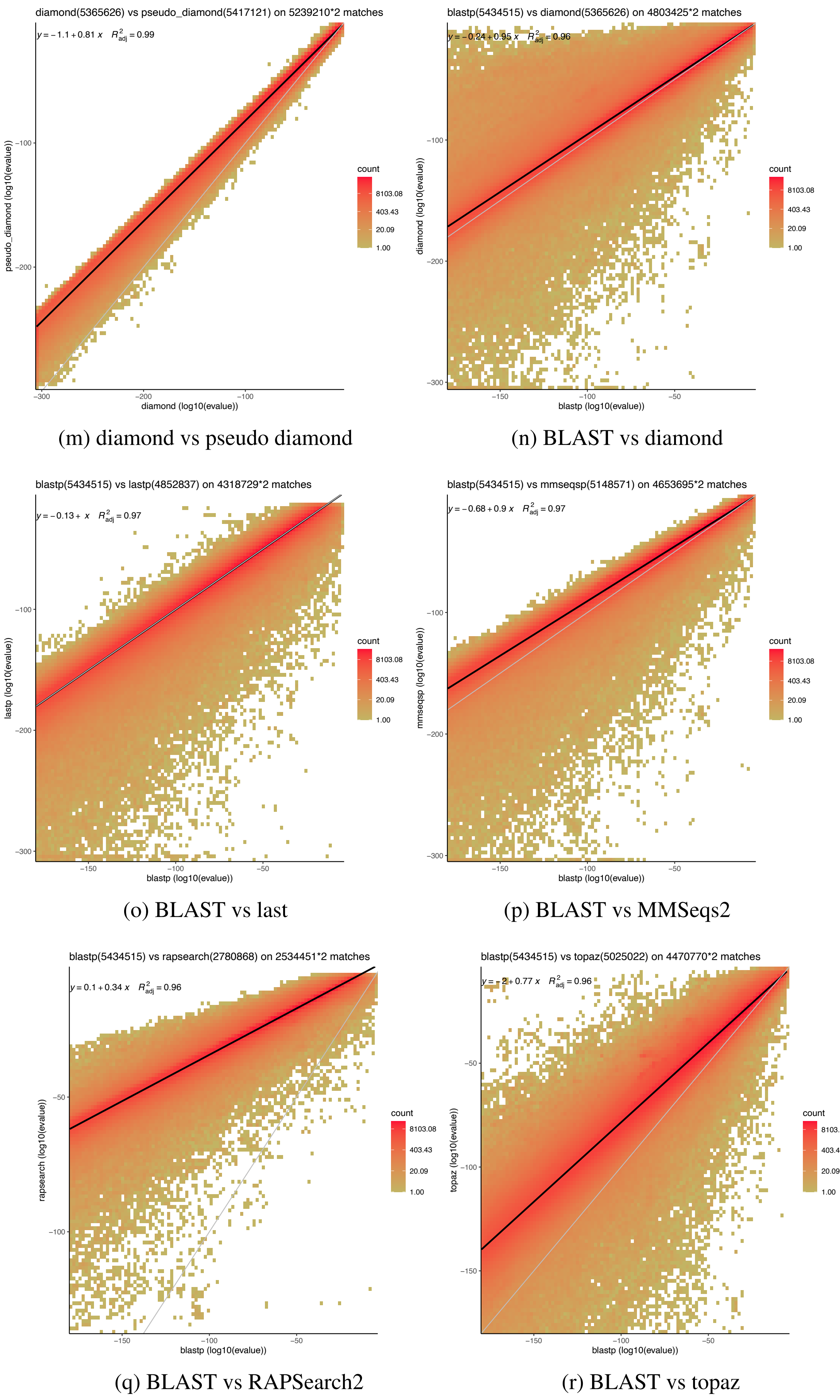
QfO:GO

QfO:EC

Figure S10: Function-Based Benchmarks. x: Recall - Number of Ortholog Relations, y: Precision - Avg. Schlicker Similarity

7 E-VALUE LINEAR REGRESSION ANALYSIS

Linear regression analysis of between different homology search programs. For two algorithms X and Y, (for example `BLAST` and `diamond`) first a classical reciprocal best hit graph is built for each program using `Proteinortho6` without clustering. The resulting BLAST graphs are then compared using R, such that for each protein pair that is found in both graphs (called ”match” in the plots) all combinations between the reported E-values are collected and correlated.



(m) diamond vs pseudo diamond

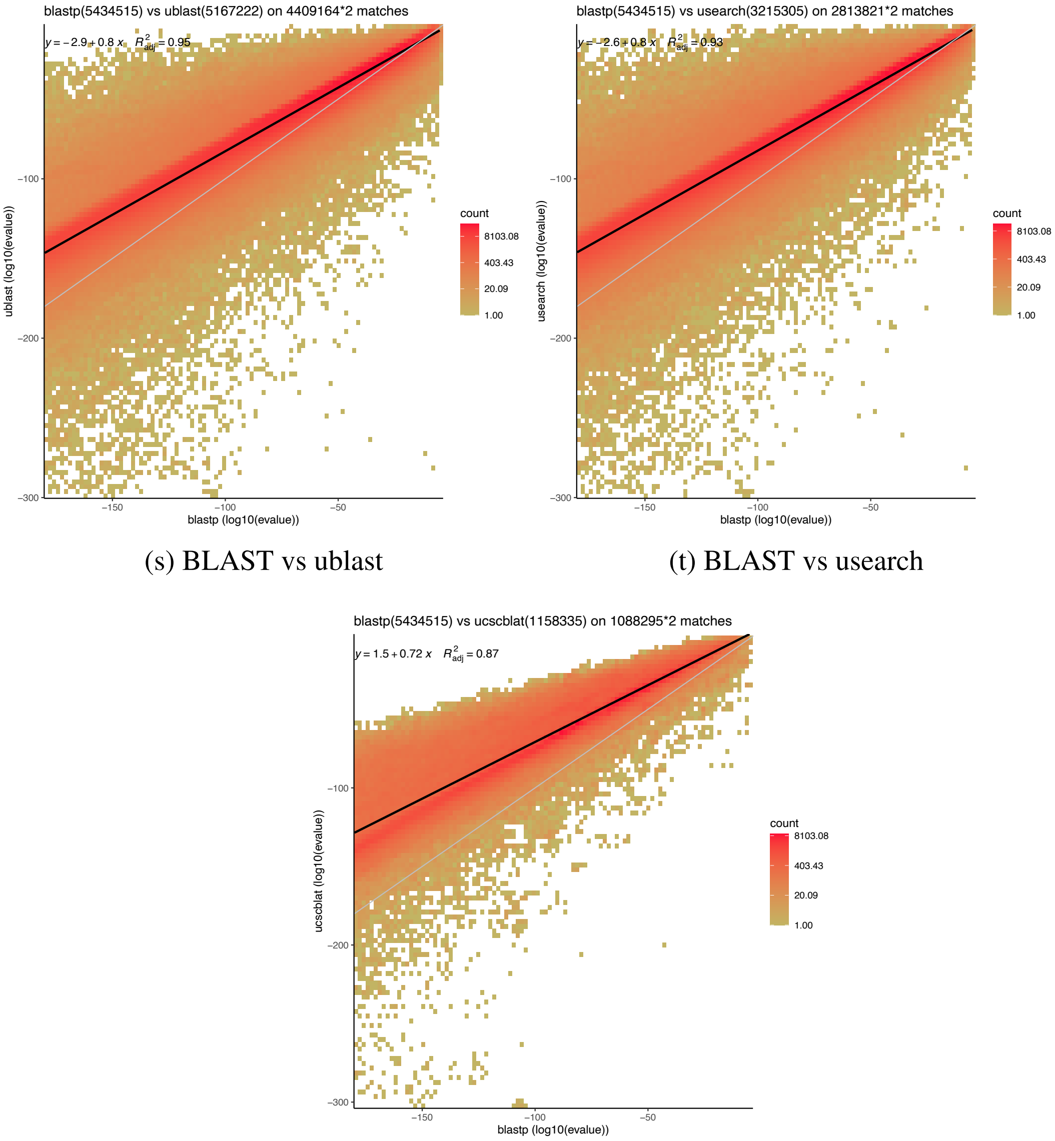
(n) BLAST vs diamond

(o) BLAST vs last

(p) BLAST vs MMSeqs2

(q) BLAST vs RAPSearch2

(r) BLAST vs topaz



(s) BLAST vs ublast

(t) BLAST vs usearch

(u) BLAST vs blat

Figure S12: Linear regression analysis of log10 transformed E-values of `pseudo` transformed values and E-values using the canonical reciprocal best hit algorithm of `Proteinortho6`. The gray line indicates the identity function y=x. diamond: diamond in sensitive mode

REFERENCES

Bonacina, J., Suárez, N., Hormigo, R., Fadda, S., Lechner, M., and Saavedra, L. (2017). A genomic view of food-related and probiotic Enterococcus strains. *DNA research* 24, 11–24. doi:10.1093/dnares/dsw043

UniProt-Consortium (2018). Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research* 47, D506–D515

III

Conclusion and Outlook

This work presents case studies that aim to disentangle phylogenetic relationships for a non-coding and a coding gene. The first study focused on the non-coding 6S RNA, while the second case study explored the Kiwellins protein family. The last article highlighted improvements made to the program `Proteinortho`, which holds significant importance in the field of comparative genomics and particularly in relation to the aforementioned case studies.

The phylogenetic analyses incorporated established as well as new approaches. While these methods can serve as valuable guidelines for future researchers, it is important to note that they should not be considered direct blueprints as the solutions are intricately tailored to the specific problems.

Both studies revealed the insufficiency of relying solely on the primary sequence for an accurate identification process of both molecules, the 6S RNA and the Kiwellin. For the 6S RNA, it is known that the primary structure is poorly conserved (Wehner, 2014), while the structure of the Kiwellins provided valuable insights to differentiate them from closely related protein families. This highlights the importance of considering additional structural characteristics, which becomes increasingly possible with the rise of `AlphaFold` for proteins (Jumper, 2021).

Furthermore, homology, orthology inference, and phylogenetic reconstruction were cornerstones of both studies. For the Kiwellins, the reconciliation of the inferred gene tree with the respective species tree led to the discovery of three distinct Kiwellin classes and the development of a sophisticated and robust nomenclature. Conversely, for the 6S RNA, the phylogenetic tree allowed us to assess gene-tree-species-tree incongruence and assess the differences between the taxonomic groups.

Although the overarching goal was similar, the two analyses' specific methodologies and research questions diverged. For example, the genomic context (synteny) and binding motifs were investigated for the 6S RNA to better understand the gene's characteristics and regulation. In contrast, for the Kiwellins, a meta-analysis of publicly available transcriptome studies was conducted to investigate the potential roles of the different Kiwellin classes.

Throughout both studies, `Proteinortho` played an essential role as a fundamental program in various aspects of the analyses. In the case of the 6S RNA, `Proteinortho` served as a vital tool in identifying the conserved genomic context. As for the Kiwellins, it enabled the construction of a supermatrix, a building block in the reconciliation analysis.

Overall, the research showcased the significance of both primary and structural characteristics while emphasizing the significance of phylogenetic reconstruction in combination with `Proteinortho`. The following sections highlight potential future research questions.

3.1 Kiwellins in Embryophyta

This article introduces a nomenclature based on a reconciled phylogeny for the Kiwellin protein family, highlighting their distinct structural characteristics and evolutionary relationship with BL proteins. The presented meta-analysis hints at a more general and intricate response to various biotic and abiotic stresses in symbiotic interactions and cultivar and tissue-specific differences. Manipulating Kiwellins or their expression could offer a new approach for developing, for example, disease-resistant plants or enhancing symbiotic capabilities. The provided classification and understanding of Kiwellins will guide future research in unraveling their functions.

3.1.1 Kiwellins are Putative Descendants of BL

It can be hypothesized that Kiwellins have evolved out of BL by acquiring new functionality with the N-terminal extension (β-hairpin) in combination with a modification of the DPBB (loop region between β_5 and β_6). BL is a well-known protein family with diverse functions, including roles in pathogenic interactions, and is widespread in Eukaryota, including fungi and plants (Scherer, 2010). In contrast, Kiwellins have a more restricted taxonomic distribution and are exclusively found in embryophyta (land plants). There is a co-occurrence of Kiwellins and BL found in the presented study, with at least half of all Kiwellin-containing plants also harboring BL proteins. However, it is important to note that the presented study was not intended to identify BL. Using the Interpro dataset 'Barwin domain' (IPR001153) with lengths filters[16] this number goes up to three quarters. More research must be done to evaluate the statistical significance of this co-occurrence.

Another piece of evidence supporting this evolutionary hypothesis can be found in the loop region between β_5 and β_6 of the DPBB that is crucial for the function is highly conserved in Kiwellins, see Fig. 11B. While BL proteins predominantly contain a shortened version of this loop, approximately one-third of the BL identified in the study exhibit this Kiwellin-like loop region. Furthermore, seven BL proteins[17] with a kissper domain were found, which could represent intermediate versions between Kissper-Kiwellins and BL proteins. Further investigation, such as a reconciliation analysis combining the BL proteins with the Kiwellin protein family, could give more insights into this hypothesis.

In addition to the BL proteins, there are various other DPBB-containing proteins, for example, the glycoside hydrolase (family 45, IPR000334), rare lipoprotein A (IPR012997), or the

[16] lengths between 100 and 150

[17] A0A1U8EI66 (Fig. 11A), A0A1U8A6K9, A0A1S3YYL2, A0A1S3Z8R5, A0A1U7YB43, A0A5C7IXJ5, P43082

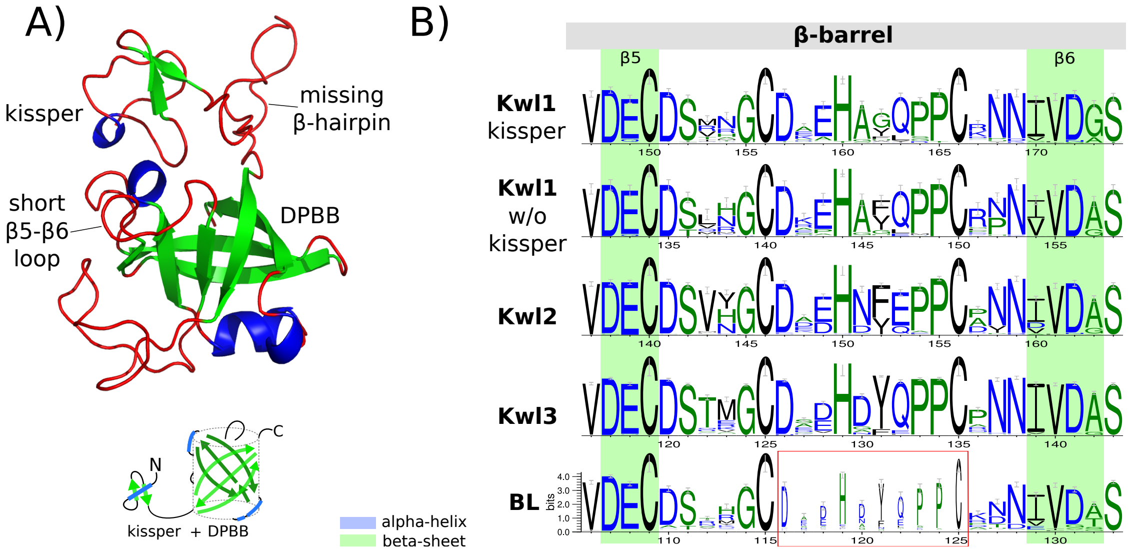


Figure 11: Hint of the evolutionary connection between the BL and the Kiwellin protein families. 3D structure prediction of a BL with kissper domain missing the β-hairpin (A0A1U8EI66) **(A)** with comic abstraction below. Aligned weblogo of the loop region between β_5 and β_6 in the DPBB **(B)**, adapted from 2.1 (Klemm, 2022). The symbol height correlates with the degree of aminio acid conservation, and the reduced symbol width (red box) indicates that a substantial fraction of BL proteins lack this loop region that is a hallmark of Kiwellins. Color code in panel A: green: beta sheet, blue: alpha helix.

pathogenesis-related protein-4 (IPR044301) (Scherer, 2010). Understanding the evolutionary relationship and dependency of these protein families in relation to Kiwellins and BL proteins would require further investigation.

3.1.2 Agricultural Applications

Crop diseases pose a severe threat to global food security, drawing attention to crop design using synthetic biology approaches (Messina, 2020; Bi, 2022; Van Dijk, 2021; Fenu, 2021; Zaidi, 2020). Kiwellins were shown to be an important defense mechanism of the plant in the interaction between *Z. mays*, and *U. maydis* (Han, 2019; Altegoer, 2020). Besides this single finding, various other pathogenic interactions with agriculturally important plants were found, for example, rice (*Oryza sativa*), soybean (*Glycine max*) and wheat (*Triticum aestivum*). Full details can be found in the supplement of the article. In general, at least one differential response with respect to pathogenic interaction was found for all three classes of Kiwellins, including the Kissper-Kiwellins. Furthermore, all classes showed differential responses to symbiotic interactions with increased concentrations in nodules and some with respect to abiotic stresses as well.

Although no direct response pattern was observable regarding infection time points or treatment conditions, overall, these findings suggest that Kiwellins may act as a general plant communication molecules with different specializations of the three classes. However, more research must be done to draw precise conclusions about the functions of the different Kiwellin classes. A limitation of this meta-analysis is the relatively small sample size of the included studies compared to the number of different plants and interaction partners. At the time, finding comparable studies with similar setups for many plants was challenging, such as using the same tissue type or infection time-point. To gain a clearer understanding, it would be beneficial to investigate more plants and different cultivars, analyzing their Kiwellin class composition and expression profiles using standardized protocols. These results could be correlated with pathogenic susceptibility, symbiotic interaction, or general stress response.

The provided meta-analysis of the article can serve as a starting point for future researchers interested in this topic. Building upon these findings, breeding or genetically modifying plants by optimizing their Kiwellin portfolio for specific needs could be a viable strategy to fight crop loss (Zaidi, 2020).

Plants belonging to the Brassicales order are particularly intriguing in this context, as the presented phylogenetic analysis has revealed that this order lacks this protein family entirely. The Brassicales order includes the model organism *Arabidopsis thaliana*, as well as agriculturally significant plants like rapeseed (*Brassica napus*). Investigating the interactions between these plants and pathogenic fungi could provide valuable insights for this taxonomic group. Furthermore, it is plausible to hypothesize that the Brassicales order could particularly benefit from introducing Kiwellins into their systems using genome-editing techniques (Zaidi, 2020). This approach holds the potential to enhance the disease resistance of these plants.

3.2 6S RNA

This study provides valuable insights into the presence and characteristics of 6S RNA in LAB species. The phylogenetic analysis revealed differences and similarities in the 6S RNA between the taxonomic groups of LAB. A comprehensive catalog of all identified 6S RNAs and structure predictions were provided to give future researchers a starting point in this field. Furthermore, the findings highlight the need for further research to unravel the functional relationships and regulatory mechanisms of this non-coding RNA with respect to the syntenic conserved *rarA*[18] and *uspA*[19] and other taxonomic specific conserved genes in close proximity. The presence of catabolite responsive elements (CREs) hints at a potential association between 6S RNA and metabolic adaptation in LAB. Expanding the knowledge of 6S RNA in LAB may open doors to its utilization in biotechnological or pharmacological applications.

3.2.1 Biotechnological Applications

In the context of fermentation, LAB species are commonly used as starter cultures in various industrial products. Manipulating the 6S RNA could accelerate the fermentation process. One approach involves targeting a cre-binding site that is present in about a third of the LAB species. It is conceivable that the carbon catabolite repression protein could bind the cre-site and subsequently inhibits 6S RNA, but further experimental validation is required. Another approach could involve using 6S RNA knock-out mutants, which would be independent of the presence of cre-sites. However, potential side effects induced by the knock-out

[18] replication-associated recombination protein A
[19] universal stress protein A, putative Interpro:IPR006015

need to be investigated, such as increased biofilm production in the 6S-2 knockout in undomesticated *B. subtilis* strain (Thüring, 2021) or early sporulation in the 6S-1 knockout in *B. subtilis* (Cavanagh, 2013). Preliminary work demonstrated that a 6S RNA knockout could yield a faster metabolization of nutrients (Cavanagh, 2013). With this in mind, the fermentation process could be potentially accelerated upon reduction of 6S RNA levels, which may result in shorter fermentation periods and thus lower production costs.

3.2.2 Pharmacological Applications

Beyond biotechnological implications, the 6S RNA could also have direct pharmacological applications. LAB species include pathogenic species primarily found in the *Streptococcus* and *Enterococcus* genera. Manipulating the expression of 6S RNA could directly impact the survivability of these pathogens. Therefore, the 6S RNA in these species could serve as a potential target. Previous studies have shown that knock-out mutants in the pathogenic *Staphylococcus aureus* have improved antibiotic susceptibility (Esberard, 2022). Using antisense oligonucleotides (ASO) like peptide nucleic acid (Gupta, 2017), 6S RNAs could be targeted with mimics of the endogenous long pRNAs to trigger degradation of 6S RNA by cellular RNases (Beckmann, 2011). It was shown that already a 8-mer Locked Nucleic Acid (LNA) construct can trigger this rearrangement in *B. subtilis* (Beckmann, 2012). However, potential off-target effects on non-pathogenic bacteria like the symbiotic bacteria of the human gut should be evaluated carefully to ensure specificity. This approach could contribute to the development of a supplementary drug that complements antibiotic treatments for specific bacterial infections.

3.3 Proteinortho

The previous versions of `Proteinortho` were well suited for the datasets at the time. But nowadays, the ever-increasing flood of data makes it challenging to keep up with the computational demands. The algorithmic updates of `Proteinortho6` in both the sequence comparison and clustering steps have greatly improved the overall performance and scalability. Moreover, `Proteinortho6` has made significant strides in terms of availability, interoperability, and usability. It has been integrated into multiple repositories including `GitLab` and `Bioconda`, making it readily accessible to a broader user base. The adoption of the standardized *OrthoXML* (Schmitt, 2011) output format facilitates seamless integration with other bioinformatics tools, promoting interoperability across different platforms. The usability of `Proteinortho6` has been enhanced through several features. It now offers a user-friendly HTML interface, streamlining the process of setting up and running orthology analyses. Furthermore, the integration of `Proteinortho6` into the `galaxy` system makes it even more accessible and convenient for users with limited programming experience. These improvements ensure that `Proteinortho6` remains a powerful and valuable resource for analyzing protein orthology in the face of the ever-expanding wealth of biological data.

Adopting the sensitive variant of the diamond program has resulted in a substantial speedup with only marginal deductions in sensitivity and precision compared to the BLAST approach. Furthermore, introducing the new pseudo variation for RBH calculation offers additional speedup without any significant drawbacks. The integration of Lapack for computing algebraic connectivity and the Fiedler vector has significantly improved the runtime of the clustering. Additionally, the clustering step now efficiently uses multiple CPU cores and computing nodes, enhancing further scalability. Furthermore, the new core modus improves workflows in a supertree analysis, specifically within the Kiwellin project (see A1.2). Overall, Proteinortho6 provides remarkable speedup, maintains the original methodological concept, and delivers highly accurate results.

3.3.1 Orthology Refinement using Structure Prediction

The protein structure prediction with AlphaFold has the potential to enhance orthology prediction significantly. Traditionally, orthology predictions are based on sequence similarity. However, there are instances where more than sequence similarity is required, as demonstrated by the two case studies presented here and in (Holm, 2023). Although the exhaustive *de-novo* prediction of AlphaFold structures for multiple proteomes is infeasible in most cases, small datasets with appropriate computing power could be a viable option. Additionally, databases of predicted structures from AlphaFold DB (Varadi, 2022) or crystal structures RCSB PDB (Berman, 2000) could be used to reduce computational cost.

In another approach, the structural predictions could be incorporated as a refinement step building on top of the results of Proteinortho. In this scenario, sufficiently large groups containing a high ratio of proteins to species may comprise multiple groups of similar sequences (like BL and Kiwellins). Potential metrics to capture the structural similarity or differences include RMSD, MA or the proposed generalized $\mathtt{RMSDPMA}_n$ (more details can be found in the appendix A1.1). Furthermore, DALI, a structural database search tool, could be explored in this context (Holm, 2023). The major drawbacks of a DALI based implementation is the required extensive database of AlphaFold prediction and the relatively high computational costs compared to the aforementioned metrics. Therefore, exploring protein secondary structure prediction like PSSpred (Yan, 2013) could also be beneficial.

The integration of structural information can significantly improve the accuracy of the orthology prediction. However, it is crucial to fine-tune these methods to maintain a feasible computational footprint.

3.3.2 Interactive Species Workflow

In this section, a new approach is sketched that integrates phylogenetic analysis at the species level with the results obtained from Proteinortho or other orthology prediction tools. The resulting interactive report facilitates comprehensive data exploration, for example, to identify novel antibiotic targets (Bisanz, 2018).

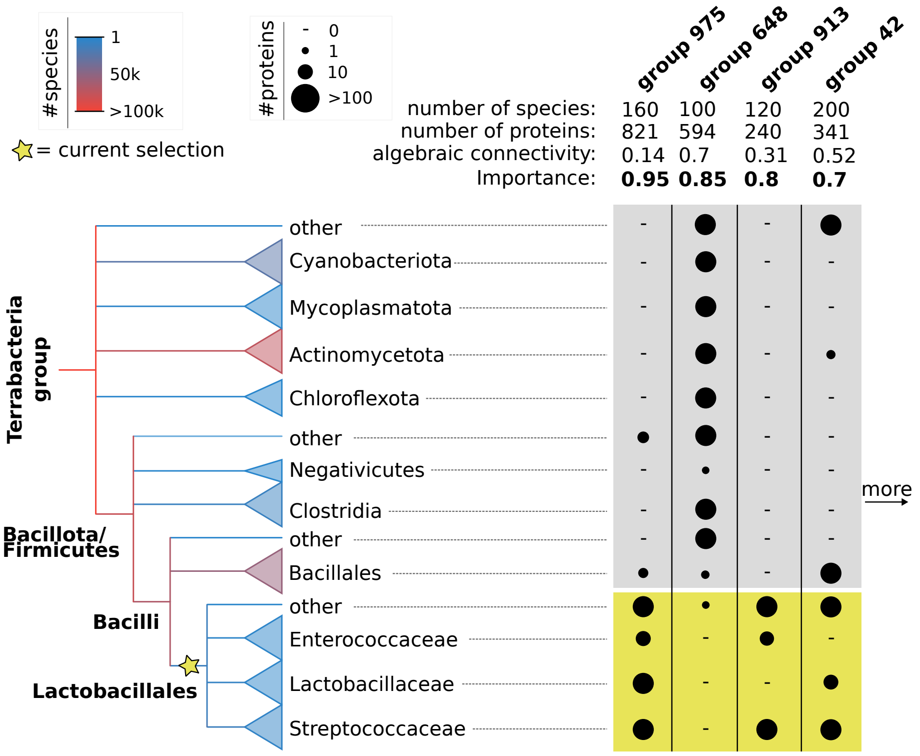


Figure 12: Concept of interactive species workflow for `Proteinortho`. Left: NCBI taxonomic tree of major groups of Terrabacteria group, colors indicate numbers of species, other: less than 1000 species. Right: `Proteinortho` groups aligned to the species tree, whereby circles encode the number of proteins contributing to the clade of species. Yellow star: selected edge highlighting the order *Lactobacialles* (yellow box). Importance: exemplary measurement of the significance of the orthology group discriminating the highlighted split (the higher the better).

A conceptual diagram illustrating this approach is presented in Fig. 12, depicting a taxonomic tree of the Terrabacteria group evolving from the left (root) to the right (species). Adjacent to the tree, a panel displays ortholog groups column-wise, where circles indicate the contribution to the corresponding clade of the tree. Each edge along the tree can be selected, inducing a split between the species below and all others. A score is required to reduce the number of displayed groups and assess the significance of orthology groups concerning the selection. One potential method that could be explored is random forest classification. In this method, the orthology groups serve as variables, represented by binary vectors where 1 indicates the presence of a protein from a given species and 0 the absence. The dependent variable also encodes the split induced by the selected edge as a binary vector. By measuring the variable importance, such as the mean decrease of accuracy or mean decrease of the GINI coefficient, a ranking of the orthology groups can be archived (Bisanz, 2018).

This integrated approach assists in identifying clade-specific orthogroups, allowing the exploration of new targets for antibiotics that specifically target certain taxonomic groups but not others. To ensure interoperability, we suggest the usage of *Newick* (Olsen, 1990) format for trees and the *OrthoXML* (Schmitt, 2011) for orthology groups, which is supported by most orthology prediction tools and databases. Additionally, this workflow could be directly integrated into the `galaxy` system (Afgan, 2022), enhancing accessibility and usability for the scientific community.

III

References

Section I

Benson, Dennis A et al. “GenBank”. In: *Nucleic acids research* 41.D1 (2012), pp. D36–D42. DOI: 10.1093/nar/gkab1135.

Bentley, David R et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *nature* 456.7218 (2008), pp. 53–59. DOI: 10.1038/nature07517.

Chaisson, Mark JP et al. “Resolving the complexity of the human genome using single-molecule sequencing”. In: *Nature* 517.7536 (2015), pp. 608–611. DOI: doi:10.1038/nature13907.

Consortium, 1000 Genomes Project et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), p. 68. DOI: 10.1038/nature15393.

Consortium, IHGS et al. “Initial sequencing and analysis of the human genome”. In: *nature* 409.6822 (2001), pp. 860–921. DOI: 10.1038/35057062.

Kunst, F j et al. “The complete genome sequence of the gram-positive bacterium Bacillus subtilis”. In: *Nature* 390.6657 (1997), pp. 249–256. DOI: 10.1038/36786.

Sanger, Frederick, Steven Nicklen, and Alan R Coulson. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467. DOI: 10.1073/pnas.74.12.5463.

Sboner, Andrea et al. “The real cost of sequencing: higher than you think!” In: *Genome biology* 12 (2011), pp. 1–10. DOI: 10.1186/gb-2011-12-8-125.

Stephens, Zachary D et al. “Big data: astronomical or genomical?” In: *PLoS biology* 13.7 (2015), e1002195. DOI: 10.1371/journal.pbio.1002195.

Turnbull, Clare et al. “The 100 000 Genomes Project: bringing whole genome sequencing to the NHS”. In: *Bmj* 361 (2018). DOI: 10.1136/bmj.k1952.

Section 1.1.1

Barrick, Jeffrey E et al. “6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter”. In: *Rna* 11.5 (2005), pp. 774–784. DOI: 10.1261/rna.7286705.

Beckmann, Benedikt M et al. “In vivo and in vitro analysis of 6S RNA-templated short transcripts in Bacillus subtilis”. In: *RNA biology* 8.5 (2011), pp. 839–849. DOI: 10.4161/rna.8.5.16151.

Beckmann, Benedikt M et al. “A pRNA-induced structural rearrangement triggers 6S-1 RNA release from RNA polymerase in Bacillus subtilis”. In: *The EMBO journal* 31.7 (2012), pp. 1727–1738. DOI: 10.1038/emboj.2012.23.

Burenina, Olga Y et al. “Involvement of E. coli 6S RNA in Oxidative Stress Response”. In: *International Journal of Molecular Sciences* 23.7 (2022), p. 3653. DOI: 10.3390/ijms23073653.

Cavanagh, Amy T, Jamie M Sperger, and Karen M Wassarman. “Regulation of 6S RNA by pRNA synthesis is required for efficient recovery from stationary phase in E. coli and B. subtilis”. In: *Nucleic acids research* 40.5 (2012), pp. 2234–2246. DOI: 10.1093/nar/gkr1003.

Cavanagh, Amy T and Karen M Wassarman. “6S-1 RNA function leads to a delay in sporulation in Bacillus subtilis”. In: *Journal of bacteriology* 195.9 (2013), pp. 2079–2086. DOI: 10.1128/jb.00050-13.

Chen, James et al. “6S RNA mimics B-form DNA to regulate Escherichia coli RNA polymerase”. In: *Molecular cell* 68.2 (2017), pp. 388–397. DOI: 10.1016/j.molcel.2017.09.006.

Hindley, J t. “Fractionation of 32P-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting”. In: *Journal of molecular biology* 30.1 (1967), pp. 125–136. DOI: 10.1016/0022-2836(67)90248-3.

Leroy, Frédéric and Luc De Vuyst. “Lactic acid bacteria as functional starter cultures for the food fermentation industry”. eng. In: *Trends in Food Science & Technology* 15.2 (Feb. 2004), pp. 67–78. DOI: 10.1016/j.tifs.2003.09.004.

Mattila-Sandholm, Tiina, Jaana Mättö, and Maria Saarela. “Lactic acid bacteria with health claims—interactions and interference with gastrointestinal flora”. In: *International Dairy Journal* 9.1 (1999), pp. 25–35. ISSN: 0958-6946. DOI: 10.1016/S0958-6946(99)00041-2.

Smid, E. J. and J. Hugenholtz. “Functional genomics for food fermentation processes.” In: *Annual review of food science and technology* 1 (2010), pp. 497–519. ISSN: 1941-1413. DOI: 10.1146/annurev.food.102308.124143.

Steuten, Benedikt et al. “Regulation of transcription by 6S RNAs: insights from the Escherichia coli and Bacillus subtilis model systems”. In: *RNA biology* 11.5 (2014), pp. 508–521.

Thüring, Marietta et al. “6S-2 RNA deletion in the undomesticated B. subtilis strain NCIB 3610 causes a biofilm derepression phenotype”. In: *RNA biology* 18.1 (2021), pp. 79–92. DOI: 10.1080/15476286.2020.1795408.

Wassarman, Karen Montzka and Gisela Storz. “6S RNA regulates E. coli RNA polymerase activity”. In: *Cell* 101.6 (2000), pp. 613–623. DOI: 10.1016/S0092-8674(00)80873-9.

Wehner, Stefanie et al. “Dissemination of 6S RNA among bacteria”. In: *RNA biology* 11.11 (2014), pp. 1467–1478. DOI: 10.4161/rna.29894.

Willkomm, Dagmar K et al. “Experimental RNomics in Aquifex aeolicus: identification of small non-coding RNAs and the putative 6S RNA homolog”. In: *Nucleic acids research* 33.6 (2005), pp. 1949–1960. DOI: 10.1093/nar/gki334.

Section 1.1.2

Altegoer, Florian et al. “The two paralogous kiwellin proteins KWL1 and KWL1-b from maize are structurally related and have overlapping functions in plant defense”. In: *Journal of Biological Chemistry* 295.23 (2020), pp. 7816–7825. DOI: 10.1074/jbc.RA119.012207.

Bebber, Daniel P, Mark AT Ramotowski, and Sarah J Gurr. “Crop pests and pathogens move polewards in a warming world”. In: *Nature climate change* 3.11 (2013), pp. 985–988. DOI: 10.1038/NCLIMATE1990.

Blum, Matthias et al. “The InterPro protein families and domains database: 20 years on”. In: *Nucleic acids research* 49.D1 (2021), pp. D344–D354. DOI: 10.1093/nar/gkaa977.

Dabravolski, Siarhei A and Zakharia Frenkel. “Diversity and evolution of pathogenesis-related proteins family 4 beyond plant kingdom”. In: *Plant Gene* 26 (2021), p. 100279. DOI: 10.1016/j.plgene.2021.100279.

Djamei, Armin et al. “Metabolic priming by a secreted fungal effector”. In: *Nature* 478.7369 (2011), pp. 395–398. DOI: 10.1038/nature10454.

Fenu, Gianni and Francesca Maridina Malloci. “Forecasting plant and crop disease: an explorative study on current algorithms”. In: *Big Data and Cognitive Computing* 5.1 (2021), p. 2. DOI: 10.3390/bdcc5010002.

Fine, Aaron J. “Hypersensitivity reaction to kiwi fruit (Chinese gooseberry, Actinidia chinensis)”. In: *Journal of Allergy and Clinical Immunology* 68.3 (1981), pp. 235–237. DOI: 10.1016/0091-6749(81)90189-5.

Hamiaux, Cyril et al. “Crystal structure of kiwellin, a major cell-wall protein from kiwifruit”. In: *Journal of Structural Biology* 187.3 (2014), pp. 276–281. DOI: 10.1016/j.jsb.2014.07.005.

Han, Xiaowei. “Structure-function analysis of Cmu1, the secreted chorismate mutase from Ustilago maydis”. In: (2017). DOI: 10.17192/z2017.0770.

Han, Xiaowei et al. “A kiwellin disarms the metabolic activity of a secreted fungal virulence factor”. In: *Nature* 565.7741 (2019), pp. 650–653. DOI: 10.1038/s41586-018-0857-9.

Hickey, Lee T et al. “Breeding crops to feed 10 billion”. In: *Nature biotechnology* 37.7 (2019), pp. 744–754. DOI: 0.1038/s41587-019-0152-9.

Jaswal, Rajdeep et al. “A kiwellin protein-like fold containing rust effector protein localizes to chloroplast and suppress cell death in plants”. In: *bioRxiv* (2021).

Klemm, Paul et al. “Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family”. In: *Frontiers in plant science* 13 (2022), p. 4832. DOI: 10.3389/fpls.2022.1034708.

Maor, Rudy and Ken Shirasu. “The arms race continues: battle strategies between plants and fungal pathogens”. In: *Current opinion in microbiology* 8.4 (2005), pp. 399–404. DOI: 10.1016/j.mib.2005.06.008.

Misas Villamil, Johana C et al. “A fungal substrate mimicking molecule suppresses plant immunity via an inter-kingdom conserved motif”. In: *Nature communications* 10.1 (2019), p. 1576. DOI: 10.1038/s41467-019-09472-8.

Roser, Max et al. “World population growth”. In: *Our world in data* (2013).

Scherer, Nicole M. “Pathogenesis-related proteins: phylogenetic characterization”. PhD thesis. Düsseldorf, Univ., Diss., 2010, 2010.

Tamburrini, Maurizio et al. “Kiwellin, a novel protein from kiwi fruit. Purification, biochemical characterization and identification as an allergen”. In: *The protein journal* 24.7 (2005), pp. 423–429. DOI: 10.1007/s10930-005-7638-7.

Tuppo, Lisa et al. “Kiwellin, a modular protein from green and gold kiwi fruits: evidence of in vivo and in vitro processing and IgE binding”. In: *Journal of agricultural and food chemistry* 56.10 (2008), pp. 3812–3817. DOI: 10.1021/jf703620m.

Van Dijk, Michiel et al. “A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050”. In: *Nature Food* 2.7 (2021), pp. 494–501. DOI: 10.1038/s43016-021-00322-9.

Wang, Jin et al. “A comprehensive review on kiwifruit allergy: pathogenesis, diagnosis, management, and potential modification of allergens through processing”. In: *Comprehensive reviews in food science and food safety* 18.2 (2019), pp. 500–513. DOI: 10.1111/1541-4337.12426.

Section 1.2.1

Altenhoff, Adrian M and Christophe Dessimoz. “Phylogenetic and functional assessment of orthologs inference projects and methods”. In: *PLoS computational biology* 5.1 (2009), e1000262. DOI: 10.1371/journal.pcbi.1000262.

Boyden, Alan. “Homology and analogy”. In: *Science* 164.3878 (1969), pp. 455–456. DOI: 10.1126/science.164.3878.455.

Darwin, Charles. *On the origin of species, 1859.* Routledge, 1859. DOI: 10.1093/owc/9780199554652.003.0077.

Fitch, Walter M. “Distinguishing homologous from analogous proteins”. In: *Systematic zoology* 19.2 (1970), pp. 99–113. DOI: 10.2307/2412448.

Fitch, Walter M. “Homology: a personal view on some of the problems”. In: *Trends in genetics* 16.5 (2000), pp. 227–231.

Gabaldón, Toni and Eugene V Koonin. “Functional and evolutionary implications of gene orthology”. In: *Nature Reviews Genetics* 14.5 (2013), pp. 360–366. DOI: 10.1038/nrg3456.

Gerlt, John A and Patricia C Babbitt. “Can sequence determine function?” In: *Genome biology* 1 (2000), pp. 1–10.

Hulsen, Tim et al. “Benchmarking ortholog identification methods using functional genomics data”. In: *Genome biology* 7 (2006), pp. 1–12. DOI: 10.1186/gb-2006-7-4-r31.

Jensen, Roy A. “Orthologs and paralogs-we need to get it right”. In: *Genome biology* 2.8 (2001), interactions1002–1. DOI: 10.1186/gb-2001-2-8-interactions1002.

Johnson, Toby. “Reciprocal best hits are not a logically sufficient condition for orthology”. In: *arXiv preprint arXiv:0706.0117* (2007). DOI: 10.48550/arXiv.0706.0117.

Koonin, Eugene V. “An apology for orthologs-or brave new memes”. In: *Genome Biology* 2.4 (2001), pp. 1–2.

Koski, Liisa B and G Brian Golding. “The closest BLAST hit is often not the nearest neighbor”. In: *Journal of molecular evolution* 52 (2001), pp. 540–542. DOI: 10.1007/s002390010184.

Nehrt, Nathan L et al. “Testing the ortholog conjecture with comparative functional genomic data from mammals”. In: *PLoS computational biology* 7.6 (2011), e1002073. DOI: 10.1371/journal.pcbi.1002073.

Ohno, Susumu. *Evolution by gene duplication.* Springer Science & Business Media, 1970. DOI: 10.1007/978-3-642-86659-3.

Panchen, Alec L. “Homology—history of a concept”. In: *Novartis Foundation Symposium 222-Homology: Homology: Novartis Foundation Symposium 222.* Wiley Online Library. 2007, pp. 5–23. DOI: 10.1002/9780470515655.ch2.

Pearson, William R. “An introduction to sequence similarity (“homology”) searching”. In: *Current protocols in bioinformatics* 42.1 (2013), pp. 3–1.

Petsko, Gregory A. “Homologuephobia”. In: *Genome biology* 2.2 (2001), pp. 1–2.

Wake, David B. “Comparative Terminology: Homology. The Hierarchical Basis of Comparative Biology. Brian K. Hall, Ed. Academic Press, San Diego, CA, 1994. xvi, 483 pp., illus. 54.95 or£ 42.” In: *Science* 265.5169 (1994), pp. 268–269. DOI: 10.1126/science.265.5169.268.

Wake, David B. “Homoplasy, homology and the problem of ‘sameness’ in biology”. In: *Novartis Foundation Symposium 222-Homology: Homology: Novartis Foundation Symposium 222.* Wiley Online Library. 2007, pp. 24–46.

Wolf, Yuri I and Eugene V Koonin. “A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes”. In: *Genome biology and evolution* 4.12

(2012), pp. 1286–1294. DOI: 10.1186/gb-2001-2-4-comment1005.

Section 1.2.2

Altenhoff, Adrian M et al. “Standardized benchmarking in the quest for orthologs”. In: *Nature methods* 13.5 (2016), pp. 425–430. DOI: 10.1038/nmeth.3830.

Altenhoff, Adrian M et al. “OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more”. In: *Nucleic acids research* 49.D1 (2021), pp. D373–D379. DOI: 10.1093/nar/gkaa1007.

Altschul, Stephen F et al. “Basic local alignment search tool”. In: *Journal of molecular biology* 215.3 (1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Bisanz, Jordan E et al. “Illuminating the microbiome’s dark matter: a functional genomic toolkit for the study of human gut Actinobacteria”. In: *BioRxiv* (2018), p. 304840.

Bork, P et al. “Predicting function: from genes to genomes and back.” In: *J Mol Biol* 283.4 (Nov. 1998), pp. 707–725. DOI: 10.1006/jmbi.1998.2144.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. “Fast and sensitive protein alignment using DIAMOND”. In: *Nature methods* 12.1 (2015), pp. 59–60. DOI: 10.1038/nmeth.3176.

Cosentino, Salvatore and Wataru Iwasaki. “SonicParanoid: fast, accurate and easy orthology inference”. In: *Bioinformatics* 35.1 (2019), pp. 149–151. DOI: 10.1093/bioinformatics/bty631.

Emms, DM and S Kelly. “OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences”. In: *BioRxiv* 466201 (2018). DOI: 10.1101/466201.

Klemm, Paul et al. “Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family”. In: *Frontiers in plant science* 13 (2022), p. 4832. DOI: 10.3389/fpls.2022.1034708.

Lechner, Marcus et al. “Proteinortho: detection of (co-) orthologs in large-scale analysis”. In: *BMC bioinformatics* 12.1 (2011), pp. 1–9. DOI: 10.1186/1471-2105-12-124.

Li, Li, Christian J Stoeckert, and David S Roos. “OrthoMCL: identification of ortholog groups for eukaryotic genomes”. In: *Genome research* 13.9 (2003), pp. 2178–2189. DOI: 10.1101/gr.1224503.

Milgram, S. “The small world problem”. In: *Psychology Today* 1 (May 1967), pp. 61–67.

Palmer, Jonathan M. and Jason E. Stajich. *Funannotate*. Version 1.8.13. Aug. 2022. URL: https://github.com/nextgenusfs/funannotate.

Peter, Jackson et al. “Genome evolution across 1,011 Saccharomyces cerevisiae isolates”. In: *Nature* 556.7701 (2018), pp. 339–344. DOI: 10.1038/s41586-018-0030-5.

Section 1.2.3

Darwin, Charles. “Notebook B”. In: (1837), p. 36. URL: http://darwin-online.org.uk/.

Darwin, Charles. *On the origin of species, 1859*. Routledge, 1859. DOI: 10.1093/owc/9780199554652.003.0077.

Dayhoff, M, R Schwartz, and B Orcutt. “22 a model of evolutionary change in proteins”. In: *Atlas of protein sequence and structure* 5 (1978), pp. 345–352.

Emms, DM and S Kelly. “OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences”. In: *BioRxiv* 466201 (2018). DOI: 10.1101/466201.

Engelhardt, Jan et al. “RNAclust. pl Documentation”. In: *Therapy* (2010), pp. 1–9.

Felsenstein, Joseph. “Cases in which parsimony or compatibility methods will be positively misleading”. In: *Systematic zoology* 27.4 (1978), pp. 401–410. DOI: 10.2307/2412923.

Hernandez, Alexandra M and Joseph F Ryan. “Six-state amino acid recoding is not an effective strategy to offset compositional heterogeneity and saturation in phylogenetic analyses”. In: *Systematic Biology* 70.6 (2021), pp. 1200–1212.

Kapli, Paschalia, Ziheng Yang, and Maximilian J Telford. “Phylogenetic tree building in the genomic age”. In: *Nature Reviews Genetics* 21.7 (2020), pp. 428–444. DOI: 10.1038/s41576-020-0233-0.

Kong, Sungsik, David Swofford, and Laura Kubatko. “Inference of Phylogenetic Networks from Sequence Data using Composite Likelihood”. In: *bioRxiv* (2022), pp. 2022–11.

Kozlov, Alexey M et al. “RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference”. In: *Bioinformatics* 35.21 (2019), pp. 4453–4455. DOI: 10.1093/bioinformatics/btz305.

Levenshtein, Vladimir I et al. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. Soviet Union. 1966, pp. 707–710.

Menet, Hugo, Vincent Daubin, and Eric Tannier. “Phylogenetic reconciliation”. In: *PLOS Computational Biology* 18.11 (2022), e1010621. DOI: 10.1371/journal.pcbi.1010621.

Michener, Charles D and Robert R Sokal. “A quantitative approach to a problem in classification”. In: *Evolution* 11.2 (1957), pp. 130–162. DOI: 10.1111/j.1558-5646.1957.tb02884.x.

Morel, Benoit et al. “GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss”. In: *Molecular biology and evolution* 37.9 (2020), pp. 2763–2774. DOI: 10.1093/molbev/msaa141.

Nguyen, Lam-Tung et al. “IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies”. In: *Molecular biology and evolution* 32.1 (2015), pp. 268–274. DOI: 10.1093/molbev/msu300.

Ohno, Susumu. *Evolution by gene duplication*. Springer Science & Business Media, 1970. DOI: 10.1007/978-3-642-86659-3.

OpenTree. “Open Tree of Life Synthetic Tree”. In: *Nature methods* 12.3 (2021). DOI: 10.5281/zenodo.3937741.

Philippe, Hervé et al. “Heterotachy and long-branch attraction in phylogenetics”. In: *BMC evolutionary biology* 5.1 (2005), pp. 1–8. DOI: 10.1186/1471-2148-5-50.

Saitou, Naruya and Masatoshi Nei. “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” In: *Molecular biology and evolution* 4.4 (1987), pp. 406–425.

Stamatakis, Alexandros. “RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies”. In: *Bioinformatics* 30.9 (2014), pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033.

Strimmer, Korbinian and Arndt Von Haeseler. “Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies”. In: *Molecular biology and evolution* 13.7 (1996), pp. 964–969. DOI: 10.1093/oxfordjournals.molbev.a025664.

Yang, Ziheng. “A space-time process model for the evolution of DNA sequences.” In: *Genetics* 139.2 (1995), pp. 993–1005. DOI: 10.1093/genetics/139.2.993.

Section 1.3

Anderson, Edward et al. *LAPACK users' guide.* SIAM, 1999. DOI: 10.1137/1.9780898719604.

Bientinesi, Paolo, Inderjit S Dhillon, and Robert A Van De Geijn. "A parallel eigensolver for dense symmetric matrices based on multiple relatively robust representations". In: *SIAM Journal on Scientific Computing* 27.1 (2005), pp. 43–66. DOI: 10.1137/030601107.

Fiedler, Miroslav. "Algebraic connectivity of graphs". In: *Czechoslovak mathematical journal* 23.2 (1973), pp. 298–305. DOI: 10.21136/CMJ.1973.101168.

Lechner, Marcus et al. "Proteinortho: detection of (co-) orthologs in large-scale analysis". In: *BMC bioinformatics* 12.1 (2011), pp. 1–9. DOI: 10.1186/1471-2105-12-124.

Miettinen, Pauli. *Chapter 6 Spectral Methods Part I: Spectral clustering.* 2017. URL: https://www.mpi-inf.mpg.de/fileadmin/inf/d5/teaching/ss17_dmm/lectures/2017-07-10-spectral_clustering.pdf.

Milgram, S. "The small world problem". In: *Psychology Today* 1 (May 1967), pp. 61–67.

Parlett, Beresford N and Inderjit S Dhillon. "Relatively robust representations of symmetric tridiagonals". In: *Linear Algebra and its applications* 309.1-3 (2000), pp. 121–151. DOI: 10.1016/S0024-3795(99)00262-1.

Plott, Sean. "Functions of the Binomial Coefficient". In: *unpublished manuscript* (2008).

Shi, Jianbo and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on pattern analysis and machine intelligence* 22.8 (2000), pp. 888–905. DOI: 10.1109/34.868688.

Section II

Cataldo, Pablo Gabriel et al. "Insights into 6S RNA in lactic acid bacteria (LAB)". In: *BMC Genomic Data* 22 (2021), pp. 1–15. DOI: 10.1186/s12863-021-00983-2.

Klemm, Paul et al. "Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family". In: *Frontiers in plant science* 13 (2022), p. 4832. DOI: 10.3389/fpls.2022.1034708.

Section III

Jumper, John et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589. DOI: s41586-021-03819-2.

Wehner, Stefanie et al. "Dissemination of 6S RNA among bacteria". In: *RNA biology* 11.11 (2014), pp. 1467–1478. DOI: 10.4161/rna.29894.

Section 3.1

Altegoer, Florian et al. "The two paralogous kiwellin proteins KWL1 and KWL1-b from maize are structurally related and have overlapping functions in plant defense". In: *Journal of Biological Chemistry* 295.23 (2020), pp. 7816–7825. DOI: 10.1074/jbc.RA119.012207.

Bi, Bo et al. "Present and future prospects of crop synthetic biology". In: *Crop Design* (2022), p. 100017. DOI: 10.1016/j.cropd.2022.100017.

Fenu, Gianni and Francesca Maridina Malloci. "Forecasting plant and crop disease: an explorative study on current algorithms". In: *Big Data and Cognitive Computing* 5.1 (2021), p. 2. DOI: 10.3390/bdcc5010002.

Han, Xiaowei et al. "A kiwellin disarms the metabolic activity of a secreted fungal virulence factor". In: *Nature* 565.7741 (2019), pp. 650–653. DOI: 10.1038/s41586-018-0857-9.

Klemm, Paul et al. "Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family". In: *Frontiers in plant science* 13 (2022), p. 4832. DOI: 10.3389/fpls.2022.1034708.

Messina, Carlos D et al. "Two decades of creating drought tolerant maize and underpinning prediction technologies in the US corn-belt: review and perspectives on the future of crop design". In: *BioRxiv* (2020), pp. 2020–10. DOI: 10.1101/2020.10.29.361337.

Scherer, Nicole M. "Pathogenesis-related proteins: phylogenetic characterization". PhD thesis. Düsseldorf, Univ., Diss., 2010, 2010.

Van Dijk, Michiel et al. "A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050". In: *Nature Food* 2.7 (2021), pp. 494–501. DOI: 10.1038/s43016-021-00322-9.

Zaidi, Syed Shan-e-Ali et al. "Engineering crops of the future: CRISPR approaches to develop climate-resilient and disease-resistant plants". In: *Genome biology* 21.1 (2020), pp. 1–19. DOI: 10.1186/s13059-020-02204-y.

Section 3.2

Beckmann, Benedikt M et al. "In vivo and in vitro analysis of 6S RNA-templated short transcripts in Bacillus subtilis". In: *RNA biology* 8.5 (2011), pp. 839–849. DOI: 10.4161/rna.8.5.16151.

Beckmann, Benedikt M et al. "A pRNA-induced structural rearrangement triggers 6S-1 RNA release from RNA polymerase in Bacillus subtilis". In: *The EMBO journal* 31.7 (2012), pp. 1727–1738. DOI: 10.1038/emboj.2012.23.

Cavanagh, Amy T and Karen M Wassarman. "6S-1 RNA function leads to a delay in sporulation in Bacillus subtilis". In: *Journal of bacteriology* 195.9 (2013), pp. 2079–2086. DOI: 10.1128/jb.00050-13.

Esberard, Marick et al. "6S RNA-Dependent susceptibility to RNA polymerase inhibitors". In: *Antimicrobial Agents and Chemotherapy* 66.5 (2022), e02435–21. DOI: 10.1128/aac.02435-21.

Gupta, Anjali, Anuradha Mishra, and Nidhi Puri. "Peptide nucleic acids: Advanced tools for biomedical applications". In: *Journal of biotechnology* 259 (2017), pp. 148–159. DOI: 10.1016/j.jbiotec.2017.07.026.

Thüring, Marietta et al. "6S-2 RNA deletion in the undomesticated B. subtilis strain NCIB 3610 causes a biofilm derepression phenotype". In: *RNA biology* 18.1 (2021), pp. 79–92. DOI: 10.1080/15476286.2020.1795408.

Section 3.3

Afgan, Enis et al. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update". In: *Nucleic Acids Research* 50.W1 (2022), W345–W351.

Berman, Helen M et al. "The protein data bank". In: *Nucleic acids research* 28.1 (2000), pp. 235–242. DOI: 10.1002/0471721204.ch9.

Bisanz, Jordan E et al. "Illuminating the microbiome's dark matter: a functional genomic toolkit for the study of human gut Actinobacteria". In: *BioRxiv* (2018), p. 304840.

Holm, Liisa et al. "DALI shines a light on remote homologs: One hundred discoveries". In: *Protein Science* 32.1 (2023), e4519. DOI: 10.1002/pro.4519.

Olsen, Gary. "The" Newick's 8: 45" tree format standard". In: *World-Wide-Web Reference. http://evolution. genetics. washington. edu/phylip/newick doc. html* (1990).

Schmitt, Thomas et al. "SeqXML and OrthoXML: standards for sequence and orthology information". In: *Briefings in bioinformatics* 12.5 (2011), pp. 485–488.

Varadi, Mihaly et al. "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models". In: *Nucleic acids research* 50.D1 (2022), pp. D439–D444. DOI: 10.1093/nar/gkab1061.

Yan, Renxiang et al. "A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction". In: *Scientific reports* 3.1 (2013), p. 2619. DOI: 10.1038/srep02619.

Appendix

This chapter covers additional challenges that did arise in the work, further details on the published results, and overarching discussions between the three articles. All programs, tools, datasets, and figures generated or used throughout this work are available via the supplementary repository[20]. Additionally the supplementary files of the three articles are compiled in this repository.

A1 Kiwellins in Embryophyta

A1.1 Improving Classification Accuracy of Kiwellins

The start point of the analysis was the published set of Kiwellins (Han, 2019), which were manually validated and grouped using `AlphaFold2` and `PyMOL`. Next, HMM models were generated and queried against the `UniProt` dataset. An adaptive E-value cutoff was applied based on false positives, which was determined for each model by considering the best result of unrelated proteins encountered along the way. A naive approach of selecting the best scoring model that among all significant ones (Kiwellin, Kissper, BL) resulted in plenty of false positive hits, as demonstrated in Tab. 1A. For Kiwellins a precision of 86.4% was reached, with a notable number of misclassified BLs. Kissper-Kiwellins suffered from contaminations with fusion variants and Kiwellins without a kissper domain, resulting in a precision of 80.3%. This shows that Kiwellins are challenging to classify based solely on sequence level information with a HMM approach.

An ensemble of descriptors was employed to improve the classification, including primary and structural information (more details can be found in the supplementary material of the article 2.1). With this approach, the classification could be improved to approximately 98% precision.

Tab. 1B summarizes the improvements made to the classification. During the analysis, for one of the 49 initial Kissper-Kwls a better fitting isoform[21] was identified, which lacks an additional N-terminus extension compared to the initial candidate (see $*$ in Tab. 1). The false positive BL hit contains a long N-terminal extension with multiple α-helices[22]. While it lacked the β-hairpin, it contained the elongated loop within the DPBB that is characteristic for Kiwellins. Notably, the filtering process did not achieve perfect precision, but it significantly reduced the number of false positives, facilitating a more straightforward verification process.

Structural metrics were utilized in our efforts to enhance the classification of the Kiwellin protein family. The two key figures for structural similarities are `RMSD` (rooted mean squared differences of atomic positions in Å) and `MA` (number of matching atoms). While the `RMSD` is a common metric of choice, relying solely on it can be misleading. High similarities on low

[20] https://gitlab.uni-marburg.de/synmikro/ag-lechner/paul-klemm-dissertation-supplement

[21] A0A7I4DT69 → A0A2K1KL29

[22] B8BHD9

Table 1: Confusion matrix of Kiwellin prediction. Columns: true class described by defining features (DPBB, kissper domain, β-hairpin), rows: predicted class from different HMM models. Naive approach, choose the best scoring model (Kiwellin, Kissper, BL) **(A)**. Sophisticated model including primary information (length, signal peptide, cysteine count) and structure information (RMSD, matching atoms, RMSDPMAS) **(B)**. $*$: A better isoform was identified for one protein of 49 Kissper-Kwls that replaces the hit. New: additional hits in the Uniprot database 2022_01.

A)

	defining feature										
β-hairpin	✓	✓	x	x	x	x	✓	✓	x		
DPBB	✓	✓	✓	✓	✓	x	x	x	x		
kissper	x	✓	x	✓	x	x	x	x	x		
	true class										
query HMM ↓	Kwl	Kissper-Kwl	BL	Kissper-BL	BL-fusion	BL-lite	Kwl-lite	Kwl-lite-fusion	unrelated	precision	new
Kwl	184		18		1	2	7		1	86.4	762
Kissper-Kwl	2	49	3	2	1		1	2	1	80.3	145
total number	186	49	117	2	7	9	10	2	28		
sensitivity (%)	98.9	100									

B)

	true class										
query HMM ↓	Kwl	Kissper-Kwl	BL	Kissper-BL	BL-fusion	BL-lite	Kwl-lite	Kwl-lite-fusion	unrelated	precision	new
Kwl	186		1							99.5	588
Kissper-Kwl		48+1$*$			1					98.0	94-1$*$
total number	186	49	117	2	7	9	10	2	28		
sensitivity (%)	100	100									

matching atoms can produce good RMSD, values as demonstrated in Fig. 13. Similarly, a high number of matching atoms can be obtained by poor alignment.

To address this issue, the two key figures were combined to create the RMSDPMAS (RMSD per MA squared):

$$\mathtt{RMSDPMAS} := \frac{\mathtt{RMSD}}{\mathtt{MA}^2}$$

A lower value of this metric indicates a higher number of matching atoms or a lower RMSD. Consequently, a low RMSD is penalized by a high MA, and vice versa.

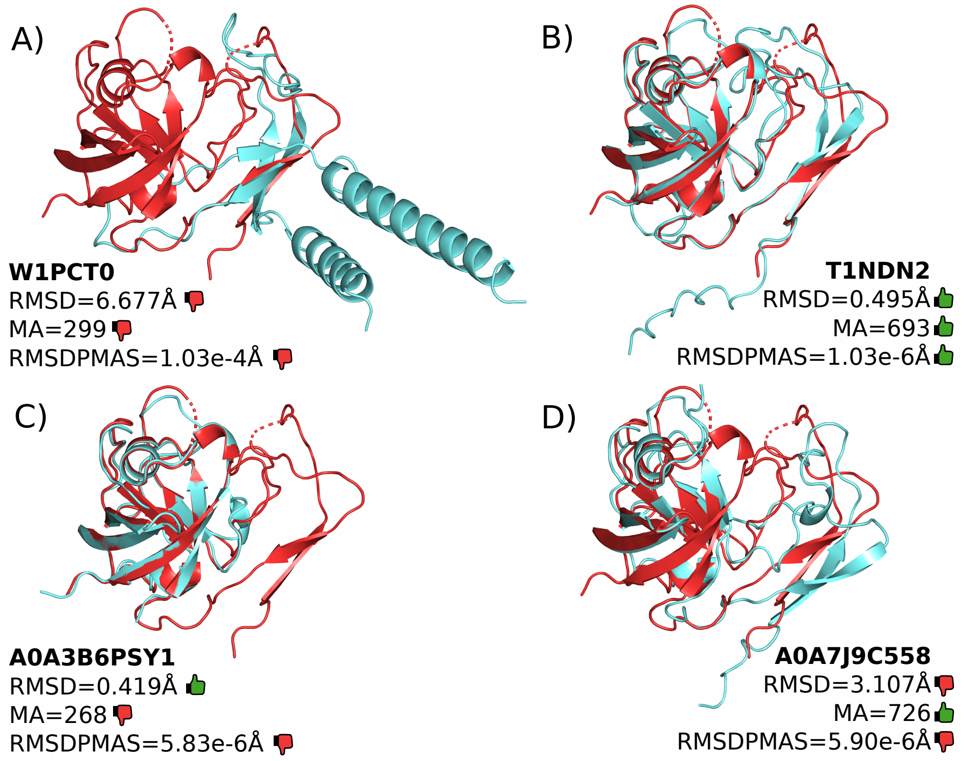


Figure 13: Superimposition of Kwl3-1b (A0A1D6GNR3, crystal structure (Han, 2019)) in red and blue the proteins A) W1PCT0, B) Kwl2-2c (T1NDN2), C) A0A3B6PSY1 and D) A0A7J9C558. A) shows an alignment with a high RMSD and low MA values resulting in a high RMSDPMAS. In contrast, the similar structures of B) result in a low RMSDPMAS. In comparison, the non-optimal alignments of C (short overlap with high similarity) and D (long overlap with low similarity) with similar RMSD or MA values, respectively, both result in a higher RMSDPMAS. Adapted from 2.1 (Klemm, 2022)

To evaluate the performance of this metric, the hits of the hidden Markov model were tested against the reference structure Kwl3-1b. Once again, the manually validated set of Kiwellins served as the positive set. When sorting the HMM hits in ascending order of RMSD, the top 500 hits included 144 true Kiwellins. Using RMSDPMAS, this improved to 293 out of 500 hits. Combined with other criteria, this improvement was sufficient for analyzing the Kiwellin family, as shown in Tab. 1.

The RMSDPMAS can be generalized as follows:

$$\texttt{RMSDPMA}_n := \frac{\texttt{RMSD}}{\texttt{MA}^n}$$

Using the same testing strategy, further improvements were evaluated by introducing hand-curated reference structures (β-hairpin, kissper domain, DPBB, and again the crystal structure of Kwl3-1b[23]). For instance, the reference β-hairpin structure should identify both Kiwellins and Kissper-Kiwellins, while the kissper domain structure should identify Kissper-Kiwellins exclusively.

For each reference structure, the top 3 performing metrics by precision were extracted. In order from best to worst they are $\texttt{RMSDPMA}_1$ (14) and $\texttt{RMSDPMA}_2$ (14), MA (12), RMSD (6), $\texttt{RMSDPMA}_4$

[23] A0A1D6GNR3

(5). Suprisingly, `MA` outperformed `RMSD` in multiple instances. The overall best-performing metrics were $\texttt{RMSDPMA}_n$ with $n = 1, 2$. Although different structures and search spaces may yield different rankings, these results demonstrate the importance of considering more than the plain `RMSD` value in analogous analyses. Further details can be found in the supplementary repository.

A1.2 The New `core` Function of `Proteinortho`

To establish the Kiwellin nomenclature, it is necessary to have both a gene tree and a corresponding species tree, along with a reconciliation between the two. The reconciliation allows us to investigate events near the root, which contain valuable information to build a phylogenetic nomenclature. Speciation events, which lead to disjoint sets of species in the underlying proteins, are less suitable for classification. On the other hand, deep duplication events often accompany neofunctionalization and effectively divide the group into different classes. Therefore it is key to find a species tree that captures the evolutionary differences between the species of the analysis. The species tree was generated from published trees, enriched with a core proteome generated with `Proteinortho`. This section is aimed to highlight the improvement of `Proteinortho` contributing to this type of analysis.

The species tree was based on `Open Tree of Life: Synthetic Tree` v13.4 (OpenTree, 2021) re-rooted and pruned to the relevant species of the analysis. The major challenge here is that this tree does not include distances (as it is a taxonomic tree). With the help of `Proteinortho` (`core` modus), a set of conserved proteins occurring in all species was generated. This set was used to compile a supermatrix, and the distances of the species tree were estimated using `IQ-Tree`. Details can be found in the Methods section of the article 2.1.

In versions before version 6 of `Proteinortho`, users were required to manually test different clustering parameters to find the clustering that results in the most core-groups, groups that span all species, while minimizing the number of proteins per group. E.g., the default clustering threshold of $\alpha = 0.1$ is too strict for this dataset as it results in no core-group.

The new feature `core` of `Proteinortho` improved this analysis. In the `core` modus, a group is only split if it would result in at least one subgroup with the same number of species as initially present. Details can be found in the third article. This results in the most optimal number of core-groups while minimizing the number of proteins per group, as shown in Fig. 14.

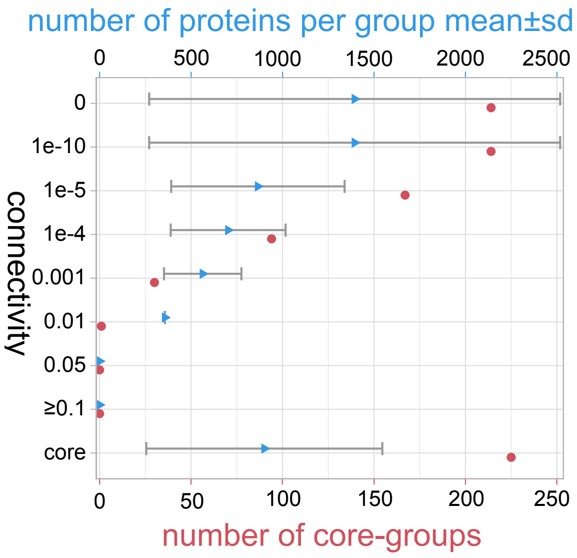


Figure 14: Number of core-groups in red and average number of proteins per core-group in blue for different connectivity thresholds (`conn`) and the `core` modus of `Proteinortho`.

A1.3 Unexpected Results

Two unexpected results catch the attention when evaluating the E-value distribution of `HMMer` search using a Kiwellin-specific model.

In a single instance, a putative Kiwellin was discovered in the saprophyte fungal species *Blyttiomyces helicus*[24], which grows on pollen. The respective other sequence-based metrics that were used to filter Kiwellins (sequence length, signal peptide, number of cysteine residues) as well as structure metrics of the candidate are on par with plant Kiwellins. See the supplementary material of the article for more information.

On the sequence level, 9 out of the 10 best `BLAST` hits based on E-value belong to the Kwl3 class, albeit with a maximum percent identity of only 60%. Structurally the candidate also exhibits similarities to the Kwl3 class, characterized by an extended loop region within the β-hairpin (Fig. 15B). The lowest `RMSDPMAS` was also observed for the Kwl3 consensus fold compared to the other classes. Despite this, in the phylogenetic tree, the fungi hit is placed among the oldest species, such as the taxonomic groups of Bryophyta and Lycopodiopsida, where Kwl1 is dominant. While it is unlikely that a horizontal gene transfer occurred from a plant in this study, it should be noted that not all embryophyte plants have been fully sequenced. For instance, *Allium cepa* (onion) was missing at the time of research, and especially these taxonomic groups are studied poorly.

Furthermore, among all 784 fungal species investigated in the analysis, no other hit was found that bears any resemblance to this case. The dataset (SAMN05443170) from which the Kiwellin was inferred originated from a pond in the USA (Ahrendt, 2018), which introduces the possibility of contamination. Moreover, the fungus *Blyttiomyces helicus* has not been successfully cultured so far, making direct verification challenging.

[24] A0A4P9WPM3

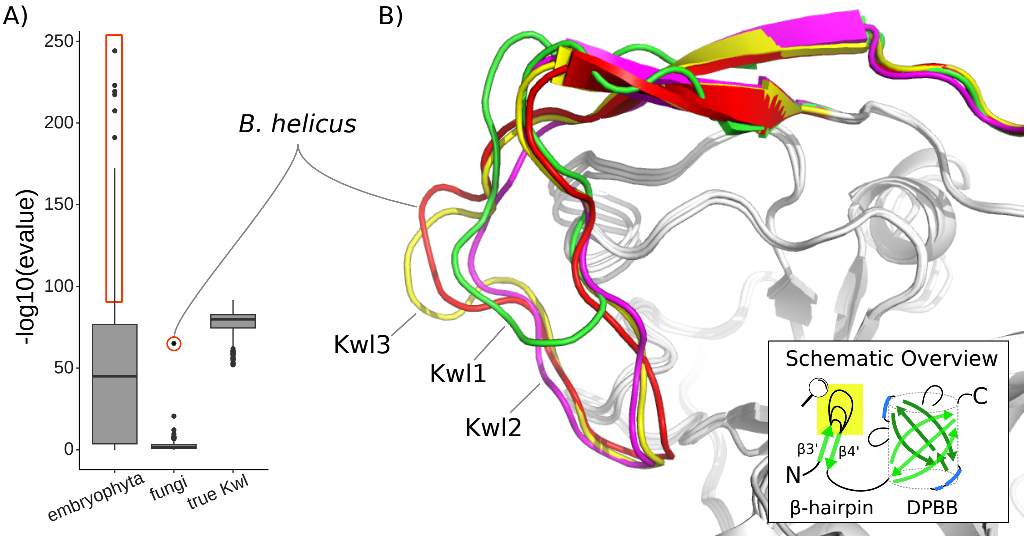


Figure 15: Boxplot of `HMMer` search E-values of Kiwellin models **(A)**. Red circle: *B. helicus* hit. Red box: fusion proteins. The consensus structure of the loop in the β-hairpin **(B)**. The Kiwellin-groups are highlighted by different colors: red: fungi candidate, green: Kwl1, magenta: Kwl2, yellow: Kwl3. The bottom right: a schematic overview of Kiwellins. Green arrow: β-sheet, blue rectangle: α-helix, yellow box: zoomed region.

The boxed hits with unexpectedly low E-values in Fig. 15A represent proteins that contain Kiwellins as domains. This includes proteins that harbor two or even three Kiwellins. The potential role of these fusion proteins can only be speculated, like a precursor form of the Kiwellins, and further research needs to be done. More details can be found in the supplementary material of the article.

A1.4 Phylogenetic Biases Evaluation

This section will assess the stability and other phylogenetic biases of the article's phylogenetic analysis. The assessment can serve as a template for conducting similar analyses in the future. The core concept revolves around introducing random noise and making changes to parameters or programs to test the resilience of the phylogenetic model. By quantifying the resulting effects using various metrics, we can gain valuable insights into the reliability and limitations of the analysis. Moreover, this framework allows to explore additional hypotheses, such as determining the domain that exerts the most influence on the phylogenetic tree.

To quantify the resulting change in the topology, the classical RF-distance for unrooted unweighted trees ("RF" in the figure) as well as for the model, the log-likelihood ("ML" in the figure) values of the tree inference were used. The higher the RF distance, the more differences there are between the tree topologies, and the higher the likelihood value, the better the tree fits the model. Three corrupted versions are generated with one, ten, and 100 randomly sampled sequences to give points of reference. Analogously a version with ten corrupted columns (random positions in each protein sequence) is generated. In general, topology changes up to 100 corrupted sequences are considered non-significant. Fig. 16 gives an overview of the impact of various parameters, like different substitution models, alignment

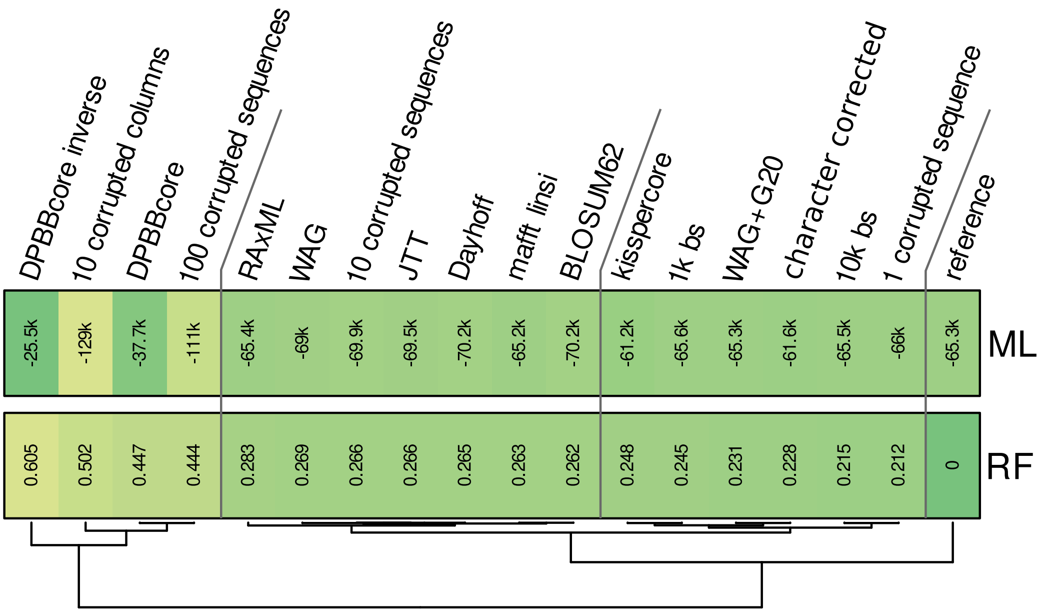


Figure 16: Impact on the phylogenetic inference for the Kiwellin protein family. Each row specifies a change of parameters compared to the reference configuration. Reference: `IQ-Tree` `WAG+G4` with 100k bs from `muscle` alignment, RF: Robinson-Foulds distance to the reference, ML: Maximum log-likelihood value of the prediction, bs: bootstraps

tools or tree inference tools. The tree of the article did serve as the reference using `IQ-Tree` with the `WAG+G4` substitution model with 100k bootstrap iterations based on a `muscle` alignment.

Domain Composition

To investigate the impact of the different domains on phylogenetic inference, different variations of the domain composition were generated. In the first variation, the kissper domain was removed from the analysis ("kisspercore"), which did not significantly alter the tree topology (RF: 0.248). The Kissper-Kiwellins were found close to the Kwl1 class, similar to the reference tree (manual inspection, see supplementary repository). This indicates that the kissper domain is not the major discriminating factor, and the evolutionary position of the Kissper-Kiwellins is not due to their increased length. Two further versions, where either the DPBB is removed ("DPBBcore") or the inverse case ("DPBBcore inverse"), led to substantial changes in the tree topology. This highlights the importance of the combination of both domains for the classification (β-hairpin and DPBB).

Alternative Programs and Substitution Models

The trees produced based on a `MAFFT` (linsi) or `muscle` alignment are similar (RF: 0.263) and result in similar likelihoods. Similarly, `RAxML` and `IQ-Tree` produced similar likelihoods and resulted in differences that are below the 100 corrupted sequences fix-point (RF: 0.283). In summary, all alignment and tree inference tools produce very similar likelihood values and agree on the overall topology of the dataset, whereby `RAxML` produces the most deviation.

The change in substitution model to BLOSUM62, Dayhoff, WAG, and JTT had only minor effects on the resulting tree, negatively impacting the likelihood value. The increase in rate heterogeneity classes (WAG+G20) did not affect the likelihood value. Finally, the change of bootstrap iterations from 100k to 10k was among the smallest measured effects.

Overall the change in parameters showed only minor effects on the tree, showing that this phylogenetic tree is robust and the program choice is largely irrelevant.

Character Composition Heterogeneity

The heterogeneity of character compositions was assessed using the χ^2-squared homogeneity test provided by `IQ-Tree`. In total 31+1[25] sequences failed the test to an alpha of 5% of which most are Kwl3 (Kwl1: 8, Kwl2: 6, Kwl3: 17), not kissper domain containing (Kissper: 8, Kwl: 23) and of the taxonomic group MON (MON: 19, NSP: 5, ROS: 4, AST: 3). The Fisher's Exact test revealed a significant result (p-value: 1.961e-05) for the taxonomic group, indicating an over-representation of MON than expected by pure chance. Despite this, the tree with and without those entries (character corrected) was inspected, and no significant difference was found (RF: 0.228, less than the effect of 10 corrupted sequences).

A1.5 Enlarged Figures

[25] the fungi Kiwellin candidate, A0A4P9WPM3

Figure 1

Structure-models of Barwin-like (A), Kiwellin (B), and Kissper-Kiwellin (C) proteins based on consensus sequences of all proteins identified for each group (signal peptide removed). Elements visible in the 3D structure on the top are indicated in a planar visualization on the bottom with identical coloring (green: β-sheets, blue: α-helices, red: loop regions). Highlighted in yellow is the β-hairpin and in red is the kissper domain. Numbered, yellow circles indicate the respective disulfide-boundforming cysteine residues. A loop region with variable length is indicated by *.

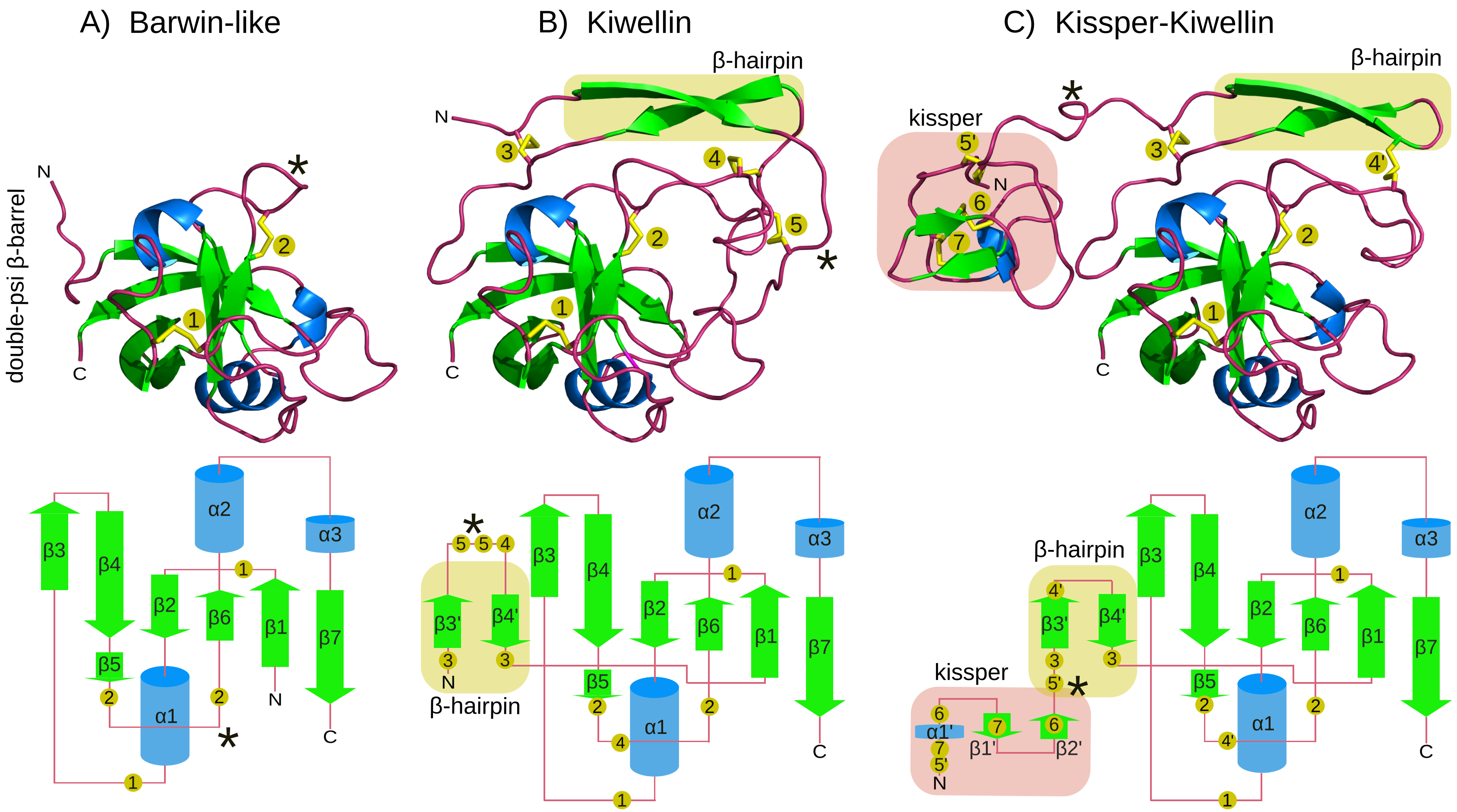


Figure 3

Aligned consensus sequences (without signal peptide) with secondary structure information of the Kiwellin subfamilies and a set of 391 BL proteins for reference. The conservation score of the consensus alignment for all Kiwellins is indicated above. The family-specific conservation is shown below the respective sequence: (mostly gaps), 0…9, + (property conservation, ascending), (perfect conservation). Amino acids are colored according to their physicochemical properties. Positions and secondary structure elements were drawn corresponding to Kissper-Kwl1. Green represents β-sheets, blue α-helices. Numbered, yellow circles indicate the cysteine residues forming disulfide bounds.

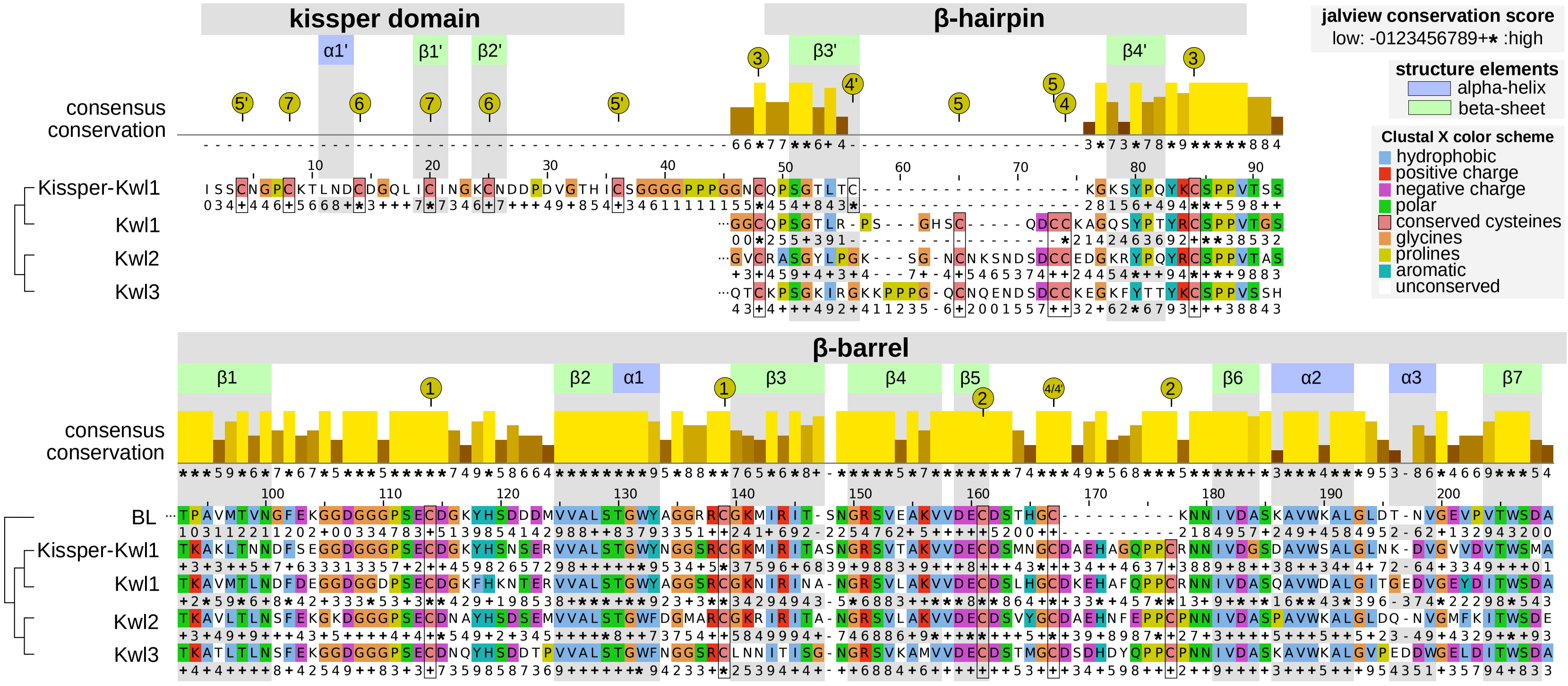


Figure 5

Species tree cladogram. The inner circle encodes taxonomic groups. The outer circle indicates if a species is found in the Kiwellin group Kwl1, Kwl2 or Kwl3. K$^+$: Kwl1 contains Kissper-Kiwellins, K: only Kissper-Kwl1. *: contains subspecies/cultivars. Ⓛ: putative loss event

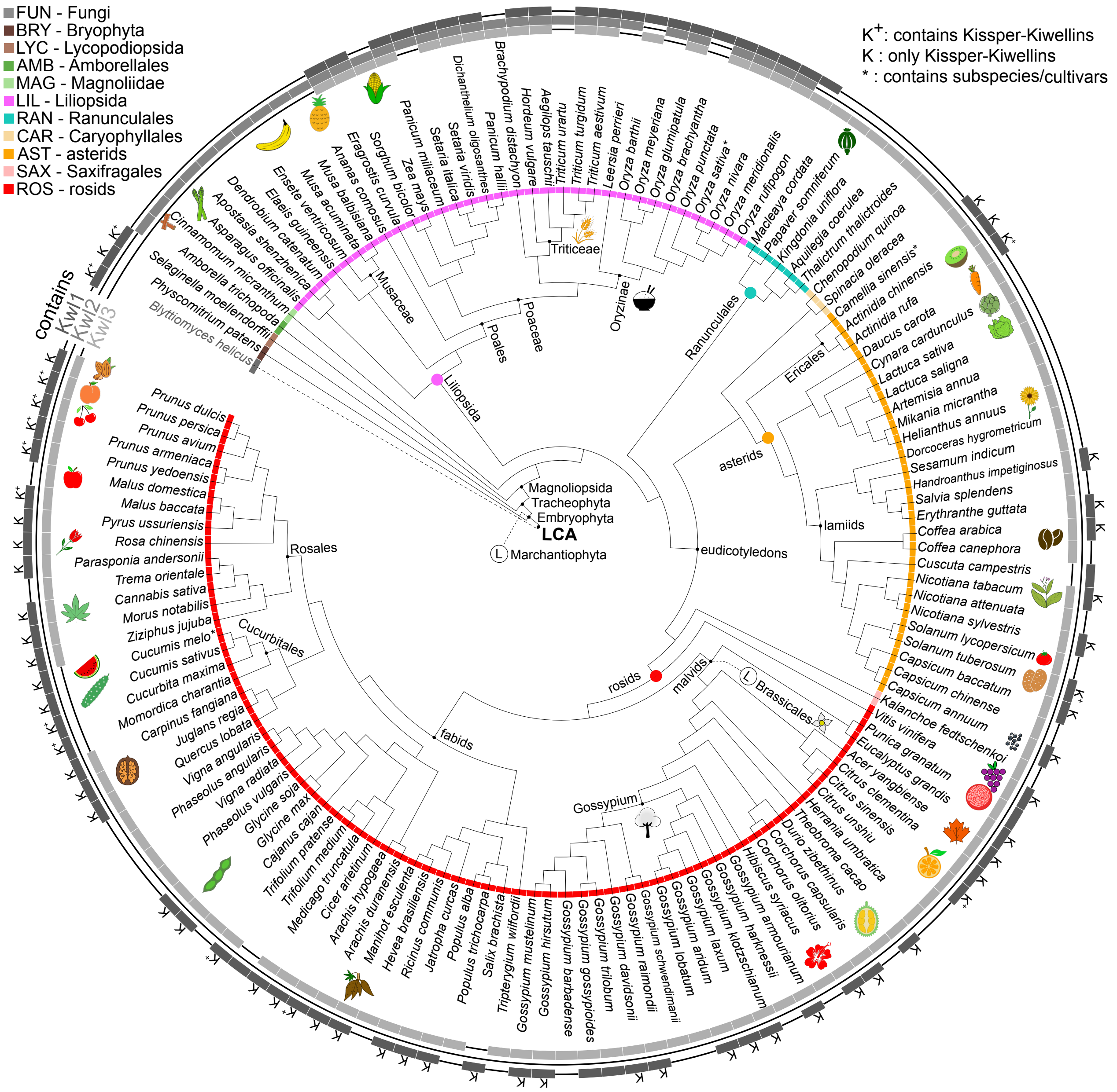
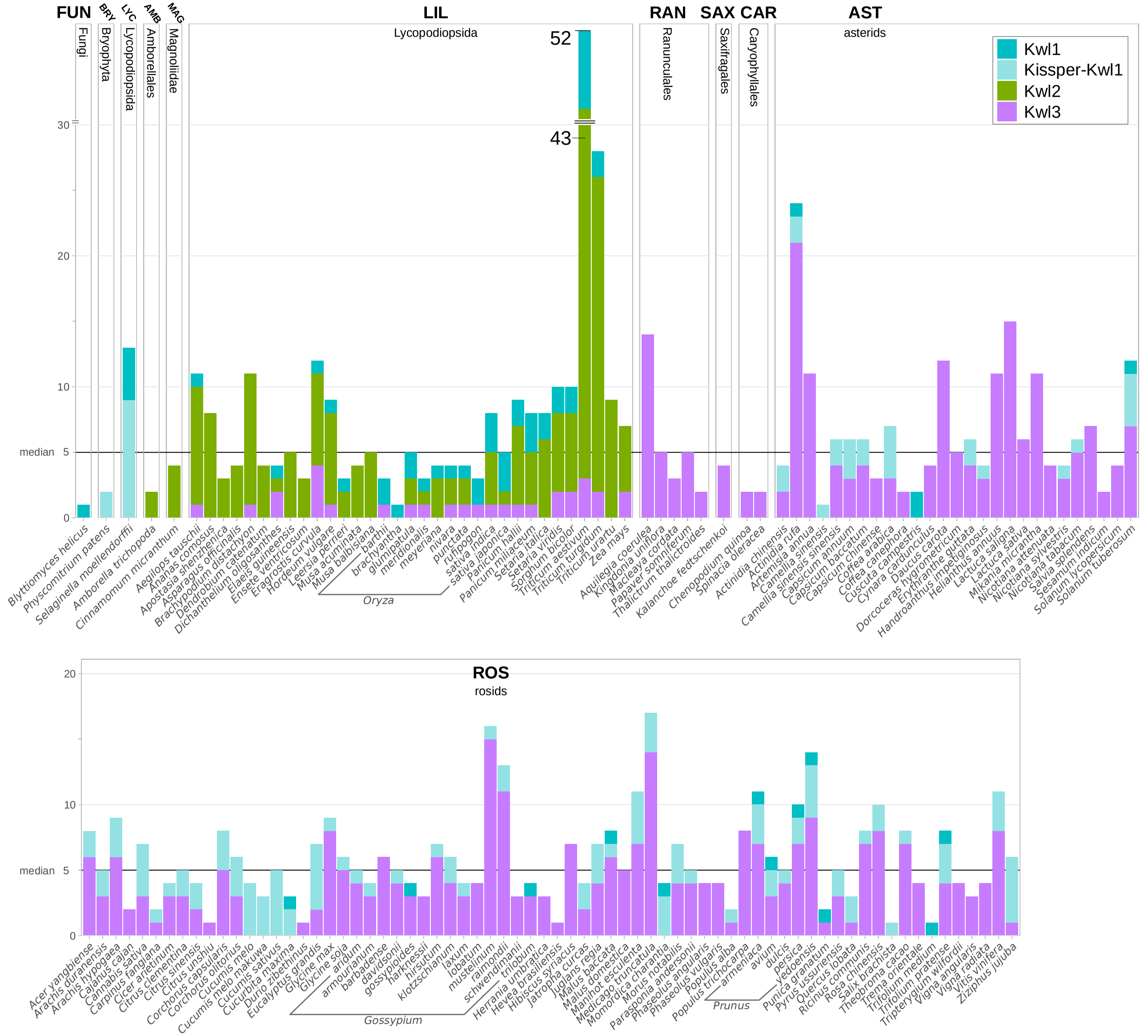


Figure 6

Kiwellins detected among specified species. The black horizontal line indicates the overall median of five Kiwellins per proteome. Colors indicate Kiwellin groups. Dark-turquoise: Kwl1, light-turquoise: Kissper-Kwl1, green: Kwl2, magenta: Kwl3.



A2 6S in LAB

A2.1 Sequence versus Secondary Structure

The general experimental procedure employed in this section closely follows the methodology outlined in the previous section A1.4. The corrupted sequences were generated analogously and processed identically to the published reference tree using `RNAclust`. However, it is essential to note that the proportion of corrupted sequences in the dataset analyzed is relatively higher compared to the Kiwellins dataset. Specifically, only 175 sequences were utilized for this particular analysis, as opposed to more than 900 Kiwellins sequences.

`RNAclust` leverages the alignment tool `mlocarna` as its underlying method. `mlocarna` incorporates primary and secondary sequence information to generate alignments. It is worth mentioning that the `RNAclust` tool does not provide a straightforward means to modify the underlying alignments that contribute to the creation of the hierarchical cluster tree. Furthermore, `RNAclust` is a hierarchical clustering method and does not represent a classic phylogenetic inference tool. Thus it does neither implement substitution models nor bootstrap iterations.

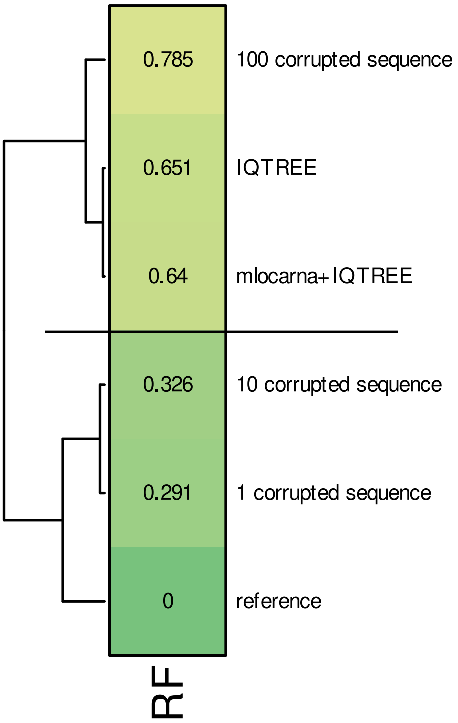


Figure 17: Impact on the phylogenetic inference for 6S RNA. Reference: `RNAclust` result of the article, RF: Robinson-Foulds distance to the reference, ML: Maximum log-likelihood value of the prediction, `IQ-Tree`: primary phylogeny based on muscle alignment and `IQ-Tree`, `mlocarna`+`IQ-Tree`: `mlocarna` alignment in combination with `IQ-Tree`

In order to facilitate comparisons, alternative versions of the tree were generated using different combinations of tools utilizing primary or secondary information. Firstly, a version was produced of the tree solely based on primary sequence information alone, utilizing the muscle alignment tool in conjunction with `IQ-Tree`. A mixed version was also generated based on the `mlocarna` alignment and `IQ-Tree`. These alternative approaches yielded markedly different results when compared to the `RNAclust` approach.

A3 Proteinortho

A3.1 Parameter Exploration

Proteinortho offers a comprehensive set of parameters to control the speed and the precision of co-ortholog detection. In this section, the impact of a selected set of parameters on similarity to the default, precision measured by QfO, and performance in terms of execution time and memory usage will be examined. The results build on top of the results of the manuscript and are aimed to give the reader insights into the effect of these parameters on the results.

The analysis will be conducted using the 2020_04 dataset of 78 species and performed on a 64-core AMD_EPYC processor analogously to the attached article. Furthermore, all results are generated using diamond v2.0.15 (Buchfink, 2015), and if not further specified, default parameters are used for Proteinortho v6.3.0[26]. Technical terms, like ARI or mean improvement, are defined in the manuscript, and further details like auxiliary scripts or precise execution times can be found in the supplementary repository.

It's important to note that results may vary when using different datasets or taxonomic compositions. Additionally, it should be noted that further parameters significantly affect the output. Furthermore, the combined effects of parameters may not directly correlate with the isolated findings of single parameters presented here.

The E-value and Similarity Threshold

Two critical parameters in the Proteinortho framework are the similarity threshold (sim) and the E-value cutoff (e), which control the all-versus-all BLAST step. The parameter e describes the E-value cutoff for the BLAST results, with a default value of 10^{-5}. sim controls the fraction of sub-optimal hits gathered below the best reciprocal hit, which is also described as the similarity threshold f (Lechner, 2011). A value of $f = 1$ corresponds to the classic reciprocal best hit algorithm, and the default for Proteinortho is $f = 95\%$. Lowering this value will increase the number of edges in the RBH graph.

Tab. 2 summarizes the impact of these two parameters on the RBH graph. Unsurprisingly, neither f nor the E-value cutoff significantly affects the runtime or memory consumption, as all observed changes are within 0.5 log2 fold changes from the default parameterization. Parameter f has the most significant effect on the size of the RBH graph, with a value of 0.1 resulting in approximately four times more edges than in default. e did show only minor effects on the size of the RBH graph.

Using Proteinortho clustering with default parameters, Tab. 3 presents the effects of these parameters on the clustering step, and Tab. 4 further displays the subsequent performance in the QfO benchmark. Parameter f did show the most impact on the size of the RBH graph and, in turn, has the most impact on the runtime of the clustering step (almost a 10-fold increase).

[26] using diamond with -sensitive and the -normal mode, the classic reciprocal best alignment heuristic

Moreover, decreasing f results in fewer clusters, but larger cluster sizes are observed (up to approximately 4-fold maximal cluster size). Despite the increase in information, a significant drop in precision in the QfO benchmark was observed. The `improvement` ranged between -0.05 and -0.97, indicating worse performance than the default setting. Furthermore, none of the metrics ranked among the top 50% for any f threshold below the default value. These results rank among the worst scores for any parameterizations besides the negative control (see manuscript for more details).

The E-value cutoff `e` did show similar clustering results as the RBH graph was only slightly affected (ARI > 0.88 for cutoffs between 10 and 10^{-50}). A stricter E-value cutoff lowers the number of connected components, similarly to a low f value, but without increasing the maximal cluster sizes. Generally, stricter E-value cutoff increases precision in the QfO benchmarks but only slightly, with up to 0.07 `improvement`. On the flip side, the best performing E-value cutoff of 10^{-50} significantly reduces the number of proteins in the graph (approximately 20%).

The impact of the two discussed parameters suggests that while they can significantly influence the results, no clear improvement without significant drawbacks could be achieved. Therefore, the default configuration[27] of the similarity threshold and the E-value cutoff was evaluated as sufficient.

Table 2: Sensitivity, precision, and resource consumption of sequence comparisons in the context of the adaptive reciprocal best hit strategy employed by `Proteinortho` with different parameters applied to the QfO benchmark dataset 2020_{04}. All results are calculated using the classic reciprocal best hit algorithm with `diamond sensitive`. Sensitivity and precision are relative to the first row. Wall time: total processing time, memory: peak memory usage, l2FC: log2 fold change relative to the first row.

algorithm	edges	sensitivity	precision	wall time		memory	
		%	%	l2FC	h	l2FC	GB
`default`	5,366k	100	100	0	.93	0	6.17
f = .1	22,615k	100	23.7	.31	.8	-.3	7.6
f = .25	20,886k	100	25.7	.31	.8	-.3	7.3
f = .5	15,264k	100	35.2	.3	.7	-.2	6.9
f = .8	8,639k	100	62.1	.1	.9	-.1	6.5
f = .99	4,390k	81.8	100	.4	.7	-.09	6.6
f = 1	4,112k	76.6	100	.3	.8	-.005	6.2
e = 10	5,406k	100	99.2	.3	.8	-.06	6.4
e = 1	5,405k	100	99.3	.3	.7	-.04	6.3
$e = 10^{-3}$	5,390k	100	99.5	.1	.9	-.05	6.4
$e = 10^{-10}$	5,258k	98.0	100	.3	.7	-.07	6.5
$e = 10^{-20}$	4,918k	91.7	100	.4	.7	.06	5.9
$e = 10^{-50}$	3,739k	69.7	100	.4	.7	.4	4.7

The Algebraic Connectivity Threshold

The algebraic connectivity threshold (`conn` or α) serves as the primary stopping condition for the spectral clustering algorithm in `Proteinortho`. Algebraic connectivity describes the

[27] f=0.95, `e`=10^{-5}

degree of connectivity within a group. Consequently, adjusting the cutoff directly influences the size and the number of connected components (CC). Increasing the threshold leads to more and smaller CCs, as shown by the shift towards smaller sizes in Table 3. By using smaller values of α, the algorithm requires less effort to satisfy the threshold, resulting in decreased runtime.

Regarding the evaluation of precision using the QfO benchmark, increasing the α generally improves overall precision. To provide a reference point, the improvement between no clustering and the default setting (0.013) was set to one unit of reference. In this notation, the improvement achieved with $\alpha = 0.75$ corresponds to almost five reference units. However, this increase in precision changes the size distribution of the connected components. The largest bin, covering at least 75% of the species, reduces by approximately 10-fold. Nonetheless, when aiming to optimize the results, the α cutoff appears to be a preferable parameter choice compared to the similarity threshold f.

Weights and Floating Point Precision

The RBH graph of Proteinortho is determined with the adaptive reciprocal best alignment heuristic based on sequence similarity calculations using tools like diamond. The resulting graph contains pairwise bit-scores as a quality measure that can be used as edge weights in the graph using, for example, the average of both values. Proteinortho6 optionally extends the original spectral clustering approach to account for these weights that guide the clustering. At the same time, floating-point precision can be adjusted as the underlying data structure in Proteinortho6. In both cases, double precision, as well as for weighted graphs, calculations offer higher precision at the expense of increased execution time and memory usage. This effect is particularly noticeable in the Lapack eigenvalue decomposition, as the quadratic nature of the input files significantly amplifies the memory consumption. By default, Proteinortho6 employs a single-precision unweighted approach.

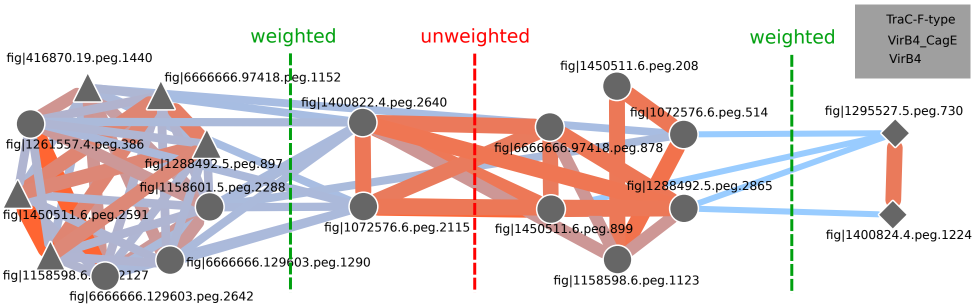


Figure 18: Weights can improve the clustering. Exemplary component from the Type IV secretion system in a real-world dataset of 29 food-related and probiotic *Enterococcus* strains (Bonacina, 2017). The connected component is split once using the unweighted algorithm. Two TraC-F enzymes are removed from a cluster of highly similar proteins while two separable VirB4 proteins remain connected. Weights (size and color-coded from blue to orange) enable more fine-grained splits which better resemble the protein annotation.

Fig. 18 illustrates an example between the weighted and unweighted clustering for a connected component of the dataset of 29 food-related and probiotic *Enterococcus* strains (Bonacina, 2017). Additional details can be found in the supplementary repository. The proteins depicted in the figure belong to the secretion IV system and can be further subdivided into `VirB4_CagE`, `VirB4` and `TraC-F` types[28].

When using an unweighted clustering approach, the connected component is split into two clusters (marked by a red dotted line). In total, the unweighted clustering removes fewer edges compared to the weighted variation, which removes a smaller sum of weights.

The unweighted clustering does not effectively capture the annotated protein types, while the weighted clustering produces more plausible results. To quantify the quality of the two clusterings, we measure clustering purity as the sum of the sizes of the most frequent class in each cluster divided by the total number of elements. The weighted algorithm increases the clustering purity by 15.79%, extracting a cleaner `TraC-F` type cluster.

Besides this specific finding, it is noteworthy that, in general, the clustering results between weighted and unweighted approaches were highly similar, with ARI values above 0.99. The distribution of sizes of connected components was similar, as indicated by a non-significant χ^2-squared test (p = 0.22) comparing size bins (0-25%, 25-50%, 50-75%, 75-100%) between the weighted and unweighted cases (float precision). Additionally, there was no observable difference in clustering results between float and double precision for the unweighted case (ARI: 1). A minor difference was observed for the weighted variation (ARI: 0.98). Therefore, it is unsurprising that no combination of the two variations significantly affected the average log2 change in the QfO precision evaluation between the approaches, with differences of less than 0.01 or less than one reference unit. The execution time followed the expected pattern, with the slowest combination being weighted & double precision and the fastest being unweighted & single precision, which was nearly five times faster than the former.

In summary, due to the lack of significant differences in results (ARI) and similar precision scores, the fastest approach (unweighted & single precision) was evaluated as the most optimal configuration.

References

Section A1

Ahrendt, Steven R et al. “Leveraging single-cell genomics to expand the fungal tree of life”. In: *Nature microbiology* 3.12 (2018), pp. 1417–1428. DOI: 10.1038/s41564-018-0261-0.

Han, Xiaowei et al. “A kiwellin disarms the metabolic activity of a secreted fungal virulence factor”. In: *Nature* 565.7741 (2019), pp. 650–653. DOI: 10.1038/s41586-018-0857-9.

Klemm, Paul et al. “Evolutionary reconstruction, nomenclature and functional meta-analysis of the Kiwellin protein family”. In: *Frontiers in plant science* 13 (2022), p. 4832. DOI: 10.3389/fpls.2022.1034708.

OpenTree. “Open Tree of Life Synthetic Tree”. In: *Nature methods* 12.3 (2021). DOI: 10.5281/zenodo.3937741.

Section A3

Bonacina, Julieta et al. “A genomic view of food-related and probiotic Enterococcus strains.” In: *DNA research* 24.1 (Feb. 2017), pp. 11–24. ISSN: 1756-1663. DOI: 10.1093/dnares/dsw043.

Buchfink, Benjamin, Chao Xie, and Daniel H Huson. “Fast and sensitive protein alignment using DIAMOND”. In: *Nature methods* 12.1 (2015), pp. 59–60. DOI: 10.1038/nmeth.3176.

Lechner, Marcus et al. “Proteinortho: detection of (co-) orthologs in large-scale analysis”. In: *BMC bioinformatics* 12.1 (2011), pp. 1–9. DOI: 10.1186/1471-2105-12-124.

[28] NCBI conserved domains annotation

Table 3: Influence of the threshold on the clustering of the QfO benchmark dataset 2020_{04} (78 species): reciprocal best hit similarity f, algebraic connectivity α, and E-value `e`. l2FC: log2 fold change relative to the default, ARI: Adjusted Rand Index compared to the `Proteinortho6` with default parameters, dashed line: relative position of the default parameter, max(proteins/group): the maximal number of proteins per group, f: similarity threshold (`sim` parameter, default: 0.95), `e`: E-value cutoff (default: 10^{-5}), α: algebraic connectivity threshold (`conn` parameter, default: 0.1).

	ortho-groups							wall time		memory		
	proteins	max(proteins/group)	groups	0-25% species	25-50% species	50-75% species	75-100% species	l2FC	seconds	l2FC	GB	ARI (`default`)
`Proteinortho6` with diamond `sensitive`:												
default	523k	280	71k	67k	2921	1179	106	0	125	0	2	1
$f = .1$	551k	1208	51k	48k	2626	995	131	3	1005	.53	3	.543
$f = .25$	547k	1119	51k	47k	2666	1018	134	2.7	807	.54	3	.583
$f = .5$	554k	494	54k	50k	2845	1134	124	2	512	.35	3	.675
$f = .8$	557k	304	63k	59k	3054	1196	114	1.5	364	.078	2	.82
$f = .99$	510k	125	75k	70k	2944	1141	104	-.15	113	-.54	1	.888
$f = 1$	502k	92	75k	71k	2969	1114	105	-.16	112	-.66	1	.874
$e = 10$	516k	280	71k	67k	2863	1133	107	1.2	287	-.33	2	.961
$e = 1$	515k	280	71k	67k	2861	1135	106	1.4	324	-.42	1	.961
$e = 10^{-3}$	525k	280	72k	67k	2947	1167	113	.25	149	-.32	2	.964
$e = 10^{-10}$	515k	280	69k	65k	2896	1139	101	-.014	124	-.55	1	.95
$e = 10^{-20}$	492k	280	65k	62k	2723	1056	89	-.45	91	-.61	1	.93
$e = 10^{-50}$	427k	269	57k	54k	2093	731	56	-1.7	38	-1.7	1	.89
$\alpha = 10^{-6}$	569k	9626	35k	32k	2029	1103	350	-1.1	60	-1.1	1	.115
$\alpha = 10^{-5}$	569k	9626	36k	32k	2108	1183	377	-1.1	57	-.84	1	.179
$\alpha = 10^{-3}$	566k	9626	42k	37k	2807	1476	319	-1.1	58	-.32	2	.527
$\alpha = .005$	560k	9626	48k	43k	3004	1494	262	-.29	102	-.22	2	.701
$\alpha = .01$	549k	296	53k	48k	3058	1454	231	.5	177	-.15	2	.772
$\alpha = .05$	537k	284	65k	61k	3044	1292	147	.33	157	-.19	2	.932
$\alpha = .09$	532k	280	71k	67k	2956	1200	116	.52	179	-.12	2	.985
$\alpha = .11$	530k	279	73k	69k	2930	1159	102	.29	152	-.24	2	.986
$\alpha = .15$	527k	272	76k	72k	2884	1072	84	.21	144	-.12	2	.959
$\alpha = .2$	523k	267	79k	75k	2826	994	60	.61	190	-.15	2	.93
$\alpha = .3$	517k	257	83k	79k	2722	831	40	.41	166	-.18	2	.884
$\alpha = .5$	503k	99	90k	87k	2425	484	11	.59	188	-.16	2	.787
$\alpha = .75$	492k	69	99k	97k	1922	186	10	.18	142	-.15	2	.703

Table 4: Quantifying Orthology Inference Precision: Assessing `Proteinortho` parameters using precision metrics of QfO benchmark dataset 2020_{04}. A full description of all reference tools and the detailed benchmark results can be found in the supplementary table of the third article. Negative control: grouped upload of $\alpha = 10^{-5}$ (see manuscript for more details), dashed line: relative position of the default parameter, `improvement`: average log2 fold change of all benchmark scores relative to default, 🟩: top 25%, 🟨: top 50%.

type	functional		phlyogeny							reference				
metric	avg. Schlicker		avg. Robinson-Foulds							PPV				
participant id	EC	GO	G STD2 Eukaryota	G STD2 Fungi	G STD2 Luca	G STD2 Vertebrata	STD Bacteria	STD Eukaryota	STD Fungi	SwissTrees	TreeFam-A	VGNC	top 25%	`improvement`
`Proteinortho6` with `diamond` `sensitive`:														
default	.969	.488	.222	.251	.244	.231	.538	.042	.297	.952	.947	.981	2	0
no clustering	.956	.483	.226	.25	.242	.234	.566	.041	.304	.949	.942	.98	0	-0.013
negative control	.673	.391	.703	.619	.756	.737	.796	.583	.623	.647	.659	.059	0	-1.473
$f = .1$	.92	.447	.537	.454	.529	.657	.669	.439	.478	.668	.689	.402	0	-0.966
$f = .25$	.923	.45	.483	.419	.465	.617	.657	.376	.446	.692	.714	.476	0	-0.862
$f = .5$	.936	.457	.337	.312	.338	.457	.607	.21	.357	.755	.766	.661	0	-0.054
$f = .8$	.957	.474	.244	.256	.262	.269	.56	.064	.307	.921	.894	.934	0	-0.122
$f = .99$	.972	.495	.216	.253	.23	.222	.535	.038	.3	.942	.956	.997	8	0.003
$f = 1$	.972	.498	.217	.245	.262	.216	.534	.038	.301	.942	.958	1	7	0.021
$e = 10$	.97	.488	.223	.245	.251	.234	.533	.041	.296	.952	.948	.981	2	0.001
$e = 1$	.969	.488	.23	.243	.252	.231	.535	.044	.296	.952	.948	.98	2	-0.009
$e = 10^{-3}$	.969	.488	.226	.244	.249	.232	.538	.042	.3	.952	.948	.98	1	-0.003
$e = 10^{-10}$	.971	.489	.231	.25	.232	.232	.535	.042	.298	.95	.948	.98	1	0.001
$e = 10^{-20}$	.972	.488	.22	.243	.226	.237	.533	.042	.289	.95	.947	.98	5	0.017
$e = 10^{-50}$	.977	.485	.211	.217	.211	.224	.484	.039	.273	.92	.947	.98	8	0.072
$\alpha = 10^{-6}$	.956	.483	.227	.252	.258	.235	.567	.041	.303	.949	.942	.98	0	-0.022
$\alpha = 10^{-5}$	.956	.483	.22	.257	.25	.234	.567	.042	.304	.949	.941	.98	0	-0.002
$\alpha = 10^{-3}$	.957	.484	.222	.245	.254	.231	.564	.04	.301	.95	.942	.98	1	-0.007
$\alpha = .005$	.959	.484	.219	.252	.248	.233	.562	.041	.305	.95	.943	.98	0	-0.011
$\alpha = .01$	.96	.485	.215	.247	.248	.227	.559	.042	.303	.95	.943	.98	1	-0.004
$\alpha = .05$	.965	.486	.221	.248	.238	.227	.546	.044	.301	.951	.945	.98	2	-0.002
$\alpha = .09$	.969	.488	.223	.244	.239	.231	.54	.043	.298	.952	.948	.98	2	0.002
$\alpha = .11$	.97	.488	.22	.247	.239	.231	.538	.041	.296	.952	.947	.98	2	0.007
$\alpha = .15$	.971	.489	.226	.25	.254	.227	.539	.042	.3	.95	.949	.98	2	-0.005
$\alpha = .2$	.974	.49	.216	.235	.253	.226	.531	.041	.289	.949	.95	.98	8	0.019
$\alpha = .3$	.978	.493	.219	.242	.236	.226	.525	.039	.287	.948	.952	.982	11	0.003
$\alpha = .5$	.982	.497	.212	.229	.263	.218	.51	.037	.281	.939	.956	.995	10	0.047
$\alpha = .75$	.982	.5	.209	.237	.292	.207	.515	.029	.285	.924	.959	.999	10	0.063

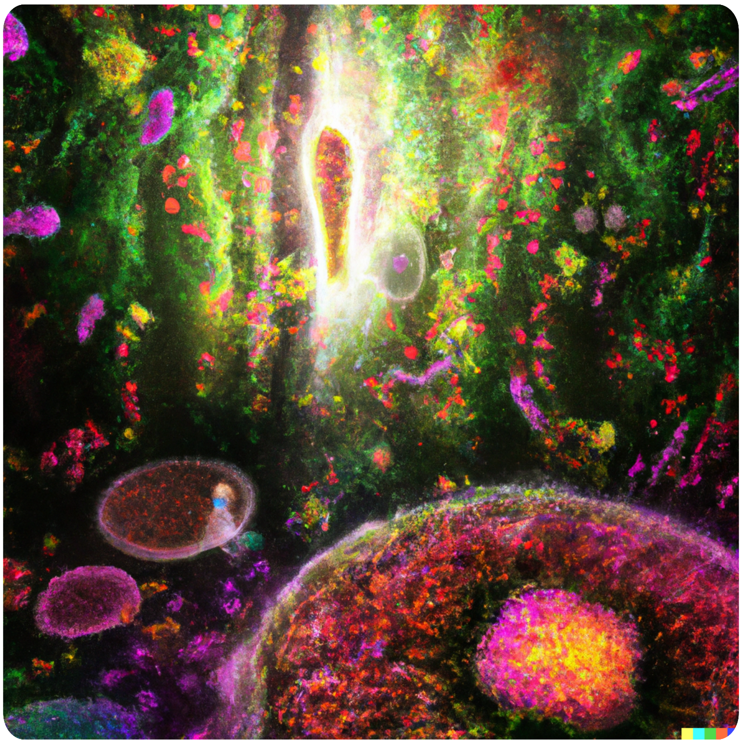
In the realm of genes, a world unknown,
Comparative genomics, a path we're shown.
Across species, the code's secrets untold,
Phylogenetic whispers and stories unfold.

Kiwellin, a protein in plants' embrace,
Against pathogens, it guards with grace.
With nomenclature defined, a systematic space,
Meta-analysis reveals responses in every place.

6S RNA, in LABs we find,
A stress-coping guide, to unwind.
Alignments, structures, a treasure to bind,
Fermentation's art, its secrets aligned.

Proteinortho, with power anew,
Orthology's magic, brought to view.
Updates and advancements, its power takes flight,
Genomic knowledge, ever shining bright.

With comparative genomics, the journey goes on,
A dance of genes, from dusk till dawn.
In nature's code, a symphony is drawn,
A harmonious ode, forever to spawn.



Poem drafted with ChatGPT (May 24 version) based on the abstract. Image generated with the assistance of DALL·E 2