

Integrative Computational Analyses Across the Central Dogma

Developing Applications for Prokaryotic Characterization in the Context of Microbial Control

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Mathias Alexander Witte Paz

aus Guayaquil/Ecuador

Tübingen

2026

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:

02.07.2026

Dekan:

Prof. Dr. Thilo Stehle

1. Berichterstatterin:

Prof. Dr. Kay Katja Nieselt

2. Berichterstatter:

Prof. Dr. Daniel Huson

3. Berichterstatter:

Prof. Dr. Andreas Dräger

Clarification on the Use of AI in the Thesis AI (here in form of ChatGPT and Writefull) was used only occasionally and very selectively for grammar and typo corrections, as well as to improve sentence flow and paragraph continuity. An example of a prompt is: "Please, improve the grammar and transitions of the following text." Even for these cases, the suggested text was always revised and adapted, or discarded, to ensure correctness as well as soundness is not lost. All content was entirely written by myself, based on my research results as well as my gained knowledge and experience in the field.

Modified template, original ["PHDUIO – Article based thesis"](#) is created by Martin Helsø and licensed under [Creative Commons CC BY 4.0](#).

Acknowledgments

“It’s called the *Patronus Charm*.” [...] Professor Lupin continued. “The *Patronus* is a kind of positive force, a projection of the very things that the Dementor feeds upon — hope, happiness, the desire to survive — but it cannot feel despair, as real humans can, so the Dementors can’t hurt it.” [...] “How do you conjure it?” “With an incantation, which will work only if you are concentrating with all your mind on a single happy memory.” ”

J.K. Rowling - Harry Potter and the Prisoner of Azkaban

The path of pursuing a PhD is rarely smooth, with lack of motivation or exhaustion as a known part of the process. Nevertheless, I had the pleasure of being surrounded by encouraging people who served as my own *Patronus charm*, providing the happy memories and positive forces to persevere.

First of all, I would like to express my deepest gratitude to my supervisor of many years, Prof. Dr. Kay Nieselt. Thank you, Kay, for your mentorship throughout my doctoral studies and also for introducing me to the world of science, already as a bachelor student. I am also grateful to the members of my TAC committee, Prof. Dr. Daniel Huson and Prof. Dr. Simon Heilbronner, for their invaluable advice and the critical questions that pushed my research further. Lastly, I would like to thank Prof. Dr. Nadine Ziemert for kindly agreeing to complement the examination committee with her expertise.

I am also grateful to the Cluster of Excellence *Controlling Microbes to Fight Infections* for funding my research and to the DAAD for the scholarship that made it possible for me to pursue my bachelor and master studies in Germany.

A huge thanks goes to the entire Integrative Transcriptomics group. First of all to my first two colleagues: Susi Zabel and Theresa Harbig. Susi, thank you for always having time for a good cup of coffee or a short walk around the Sand. Your encouragement kept me afloat, literally and figuratively, during our swimming sessions. Theresa, I am grateful for our brainstorming sessions on visualizations and for your patience with my programming questions. I also want to thank Simon Hackl, who, regardless of his *saltiness* level, was always there with a big smile, and Meret Häusler, who survived the supervision of the angry tutor in two projects and still (bravely) decided to stay in science. My thanks also to Maik Wolfram-Schauerte for encouraging my scientific path and introducing me to the world of craft beers. None of this would have been possible without the administrative work of Sabine Gebert-Rudolph and Helen Donath, who allowed the whole group to focus on science and not on bureaucracy. To complete the group, I would like to thank Findus and Théa, for bringing joy and energy to the lab every day.

I also express my gratitude to all my collaborators throughout my doctoral studies. Especially, I would like to thank Alina Bitzer and Prof. Dr. Simon Heilbronner for the many exchanges regarding siderophores, as well as Prof. Dr. Michael Krone, Nicolas Brich, and Dr. Alexander Herbig for our collaborative work on developing visualization tools.

I am also deeply grateful to all the students I had the pleasure of supervising: Camill, Valerie, Moritz, Dolores, Eileen, Meret (twice!), Thomas, Tom, Raffael and Lukas. Thank you for your hard work, for baring with me and for making my role as a supervisor so rewarding!

Acknowledgments

Within the IBMI, I would like to thank all fellow PhD students who shared this journey and provided countless happy moments: Marco, Axel, Nicolas, Selli, Caro J., Caro S., Anupam, Bitu, Sascha, and Sehmi. A special thanks goes to Dilek Tuncbilek-Dere, Alina Renz, and Martina Feierabend for making the start of my PhD so welcoming and enjoyable. I am also grateful to Martin Lang for arranging the infrastructure that made these computational analyses possible and for initiating the "Friday Beer" tradition.

I would like to thank Karla Polen-Beer and Dr. Annette Mauch for their support during times when focusing on the happy memories was challenging. My thanks also go to the entire Brochado Lab (Ana, Alex, Delaney, Joseph, Juliana, Laura, Lorine, Martin, and Martina) for your motivation and cheers during these final and intense months of writing. Looking forward to continuing to do great science with you!

To progress in science, one must also know when to take a break. I am deeply grateful to the friends who provided those essential non-scientific moments. Thanks to Vera and Max Bertsch for hosting many events in your home and garden and to Ben Röder and Sascha Schempp for the many endless game nights, even if we barely managed to actually finish a game. I am also grateful to Jaika Hoch and Felix Gehring for balancing science with the complexities of the German law system, and to Eduardo Vela, Constantin Holzapfel, and Julian Fratte for making the transition from my Master's to my doctoral studies so seamless.

Muchísimas gracias to my friends who understand what it feels like to be far from home: Meléndez, Manzi, Lucas, Nachito, Raúl, and Rafa. In particular, a heartfelt *multumesc* and a huge *sas efcharistó* go to Moni Daneliuc and Athina Gavrielidou: you were always there to listen during the most difficult moments, and your constant encouragement helped me keep my head up.

I am deeply grateful to my entire family for supporting me throughout this scientific journey. To my parents, Hans and Maria, who even from a distance, were celebrating every achievement with me. Thank you truly for teaching me to build my own path, trusting me to go abroad to discover a path that I would not have pursued in Ecuador. Thanks to my brother, Hans, and my sister-in-law, Kathi, for encouraging me to "increase the quota" of *doctors* in our family.

A huge thanks also goes to my German family: Eva, Joseph, Sylvia, Matthias, Thomas, Lena, Erik, and Laura. Although they barely understood what I was doing, they were always interested in hearing about my progress. To my grandparents, aunts, uncles, and cousins in Guayaquil: thank you for your cheers and for the energy you gave me during every visit home!

Finally, I am immensely grateful to the person who stood by me every single day: my long-time companion, the mother of my child, and my wife, Annika Witte Paz. You were my greatest support, encouraging me when I faltered and bringing peace to the darkest days. You are the main pillar that made this dissertation possible. I cannot thank you enough.

Thank you all for all the happy memories that helped cast the *Patronus* charm.

Abstract

The global rise in antibiotic resistance, coupled with the declining discovery of new compounds, has created a crisis that demands innovative strategies for microbial control. Identifying such novel strategies requires a mechanistic understanding of bacterial molecular pathways. The central dogma of molecular biology, expanded by newer findings on the flow of biological information, provides a fundamental roadmap for prokaryotic characterization based on high-throughput experiments and computational approaches. However, capturing this complexity requires integrative and reproducible computational frameworks that enable robust cross-layer analyses. This thesis addressed this challenge by developing and applying computational methods that support integrative and visual analyses across the central dogma, with focus on microbial characterization for their control.

To enhance the interactive characterization of prokaryotic genomes, the visual analytics tool **Evidente** is introduced, with the goal of bridging the gap between the exploration of single nucleotide polymorphisms (SNP) and the evolutionary context. Unlike traditional visualization tools, **Evidente** classifies SNPs based on clade-specificity and enables their visualization with metadata, as well as linking them to a functional context. By applying it to bacterial pathogens, this approach demonstrated how genome-scale variation can be interpreted in an evolutionary context to generate functional and phenotypic hypotheses.

Recognizing that genomic data alone are insufficient to explain phenotypic diversity, this thesis integrated transcriptomic data through two complementary web-applications: **TSSpredator-Web** and **TSS-Captur**. They address the characterization of the transcriptome's architecture and link the genomic with the transcriptomic layer of the central dogma. **TSSpredator-Web** extends the tool **TSSpredator** to identify and classify transcription start sites (TSS), and allows exploration of genome-wide TSS maps together with genomic data. Based on TSS maps, the Nextflow-based pipeline **TSS-Captur** characterizes transcripts starting from unclassified TSS via computational methods for sequence classification, termination site prediction, and analyses of secondary structure and promoter regions. Together, both approaches showed how transcriptomic data can be used for annotation refinement and transcript discovery, bridging the genomic and transcriptomic layer.

Lastly, the power of integrative analyses is demonstrated through two systematic studies of bacterial metallophore systems. The first project focused on the characterization of known metallophore mechanisms across the *Staphylococcal* genus using computational analyses based on genomic data, representing the first comprehensive genus-wide analysis of these systems. Expanding on this topic, the second study investigates how nasal *Corynebacterium* species exploit metallophores synthesized by *Staphylococcus aureus*, suggesting a novel way of controlling pathogens. This was achieved by identifying structural homologs of lipoproteins through the development of the Nextflow-based pipeline **PRESERVE**, and contextualizing the findings with transcriptomic data. By analyzing the putative promoter regions, we gained insight into how these mechanisms are regulated in *Corynebacteria*. These studies illustrate how the integration of multiple layers of biological data can shed light into the interactions between species, providing knowledge that can be translated into interference of colonization strategies, and be exploited for microbial control.

In summary, this thesis describes a methodological framework for integrative computational analyses across the central dogma. By combining reproducible approaches with interactive visual exploration, it illustrates how integrating data from different biological layers can help to generate insights for prokaryotic characterizations. These insights can be used to establish innovative interventions for microbial control, offering a computational pathway to address the escalating challenges of antibiotic resistance.

Deutsche Zusammenfassung

Der weltweite Anstieg der Antibiotikaresistenz in Verbindung mit der sinkenden Entdeckungsrate neuer Wirkstoffe hat zu einer internationalen Krise geführt, die innovative Strategien zur Bekämpfung von bakteriellen Pathogenen erfordert. Um neue Strategien zu identifizieren, ist es notwendig, die komplexen Mechanismen hinter den molekularen Signalwegen von Prokaryoten zu verstehen. Das zentrale Dogma der Molekularbiologie, erweitert durch neuere Erkenntnisse über den Fluss biologischer Informationen, liefert einen grundlegenden Fahrplan für die prokaryotische Charakterisierung auf der Grundlage von Hochdurchsatz-Experimenten und computergestützten Ansätzen. Um diese Komplexität zu erfassen, sind jedoch integrative Methoden erforderlich, die explorative und reproduzierbare Analysen ermöglichen, um über mehrere Ebenen Erkenntnisse zu gewinnen. Diese Dissertation hat sich dieser Herausforderung gestellt, indem sie integrative und visuelle Analysen über das zentrale Dogma hinweg entwickelt und durchführt. Dafür wurden computergestützte Methoden entwickelt und angewendet, wobei der Schwerpunkt auf der mikrobiellen Charakterisierung für deren Kontrolle liegt.

Um die interaktive Charakterisierung prokaryotischer Genome zu verbessern, wurde das *Visual Analytics Tool Evidente* eingeführt, mit dem Ziel, die Lücke zwischen der Erforschung von Einzelnukleotid-Polymorphismen (SNP, engl. *single nucleotide polymorphism*) und dem evolutionären Kontext zu schließen. Im Gegensatz zu herkömmlichen Visualisierungstools klassifiziert **Evidente** SNPs auf der Grundlage der Kladenspezifität und ermöglicht ihre Visualisierung mit Metadaten sowie ihre Verknüpfung mit funktionalen Annotationen. Durch die Anwendung auf bakterielle Pathogene zeigte dieser Ansatz, wie genomweite Variationen in einem evolutionären Kontext interpretiert werden können, um robuste funktionelle und phänotypische Hypothesen zu generieren.

Ausgehend von der Erkenntnis, dass Genomdaten allein nicht ausreichen, um die phänotypische Vielfalt zu erklären, untersuchte diese Arbeit die Integration von Transkriptomdaten durch zwei sich ergänzende Webanwendungen: **TSSpredator-Web** und **TSS-Captur**. Diese Tools befassen sich mit der Charakterisierung der Transkriptomarchitektur und verbinden die Genom- mit der Transkriptomebene des zentralen Dogmas. **TSSpredator-Web** erweitert das Tool **TSSpredator** zur Identifizierung und Klassifizierung von Transkriptionsstartstellen (TSS) und ermöglicht die Untersuchung genomweiter TSS-Karten zusammen mit Genomdaten. Basierend auf TSS-Karten, die auf Nextflow-basierte Pipeline **TSS-Captur** charakterisiert Transkripte, die von nicht klassifizierten TSS stammen, indem es Methoden für die Sequenzklassifizierung, die Vorhersage von Transkriptionsterminatoren, die Sekundärstrukturanalyse und die Promotoranalyse integriert. Zusammen zeigten beide Ansätze, wie Transkriptomdaten für die Verfeinerung von Annotationen und die Entdeckung von Transkripten verwendet werden können. Dadurch wird eine Brücke zwischen der genomischen und der transkriptomischen Ebene geschlagen.

Schließlich wird die Leistungsfähigkeit integrativer Analysen anhand von zwei systematischen Studien zu bakteriellen Metallophorsystemen demonstriert. Das erste Projekt konzentrierte sich auf die Charakterisierung bekannter Metallophormechanismen in der Gattung *Staphylococcus* mithilfe computergestützter Analysen auf der Grundlage genomischer Daten und stellte die erste umfassende gattungsweite Analyse dieser Systeme dar. In Erweiterung dieses Themas untersuchte die zweite Studie, wie nasale *Corynebacterium*-Arten die von *Staphylococcus aureus* synthetisierten Metallophore nutzen, und deutet damit einen neuen Weg zur Bekämpfung von Krankheitserregern an. Hierzu wurden strukturelle Homologe von Lipoproteinen durch die Entwicklung der Nextflow-basierten Pipeline **PRESERVE** identifiziert und die Ergebnisse mit Transkriptomdaten

in Zusammenhang gebracht. Durch die Analyse der mutmaßlichen Promotorregionen wurden Einblicke in die Regulation dieser Mechanismen in *Corynebacteria* gewonnen. Diese Studien zeigen, dass die Integration mehrerer Ebenen biologischer Daten Aufschluss über die Wechselwirkungen zwischen Arten geben kann. Diese Erkenntnisse lassen sich auf Kolonisationsstrategien übertragen und können somit für die mikrobielle Kontrolle genutzt werden.

Zusammenfassend beschreibt diese Dissertation einen methodischen Rahmen für integrative computergestützte Analysen über alle Ebenen des zentralen Dogmas hinweg. Durch die Kombination reproduzierbarer Ansätze mit interaktiver visueller Exploration wird veranschaulicht, wie die Integration von Daten aus verschiedenen biologischen Ebenen dazu beitragen kann, Erkenntnisse für die Charakterisierung von Prokaryoten zu gewinnen. Dieser computergestützte Weg ermöglicht es, Erkenntnisse zu gewinnen, mit denen innovative Maßnahmen zur mikrobiellen Kontrolle entwickelt werden können, um den zunehmenden Herausforderungen der Antibiotikaresistenz zu begegnen.

Contents

Acknowledgments	v
Contents	xi
List of Figures	xiii
List of Tables	xvii
1 Introduction	1
2 List of Papers	7
3 Contributions	9
4 Objectives	11
5 Results	13
5.1 Integrating genomic and phylogenetic data for SNP diversity exploration with Evidente	13
5.2 Prediction and Exploration of Transcription Start Sites with TSSpredator-Web	18
5.3 Characterization of Overlooked Transcriptional Units via TSS-Captur	21
5.4 Comprehensive computational analysis of metallophore synthesis and uptake potential in <i>Staphylococci</i> species	25
5.5 Characterization of Siderophore Uptake Mechanisms and their Regulation in <i>Corynebacteria</i>	30
6 Discussion and Conclusion	37
Bibliography	45
Appendices	60
I Evidente – A visual analytics tool for data enrichment in SNP-based phylogenetic trees	61
II TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration	79
III TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts	107
IV A landscape of metallophore synthesis and uptake potential of the genus <i>Staphylococcus</i>	131

V	Characterizing Siderophore Uptake and their Regulation Mechanisms in Corynebacteria based on Structural Homology and Transcriptome Profiling	157
----------	---	------------

List of Figures

1.1	Visual representation of the central dogma of molecular biology, introduced first by Crick (1958)	2
1.2	Visual representation of Anscombe's quartet (Anscombe 1973)	5
5.1	Sketch on CLASSICO's algorithm for the identification of clade-specific SNPs.	14
5.2	Visualization of a toy dataset using Evidente	15
5.3	Schematic overview of the GO-term enrichment analysis in Evidente . . .	16
5.4	Framing the data integrated in Evidente using the central dogma of molecular biology as a framework	17
5.5	Overview of the computational goals achieved by TSSpredator	19
5.6	Core components of the interactive result exploration enabled by TSSpredator-Web	20
5.7	Framing the input and computed data of TSSpredator-Web within the central dogma of molecular biology	21
5.8	Overview of the characterization of genomic regions starting at TSS via TSS-Captur	22
5.9	Framing the input and predicted data of TSS-Captur within the central dogma of molecular biology	24
5.10	Overview of data processing for the characterization of metallophore synthesis and uptake potential of staphylococcal species	26
5.11	Overview of metallophore synthesis and uptake potential in the genus <i>Staphylococcus</i>	28
5.12	Framing the data integrated in the characterization of the metallophore landscape of staphylococcal species using the central dogma of molecular biology as a framework	29
5.13	Overview of the pipeline PRESERVE for identification of structural homologs	31
5.14	Number of potential structural homologs identified by PRESERVE per sample	32
5.15	Overview of further characterization of <i>Corynebacteria</i> samples based on the identified structural homologs	32
5.16	Volcano plots summarizing the complete transcriptomic response for a control vs. treatment with the chelator EDDHA.	33
5.17	Overview of the results for motif enrichment across the analyzed species. .	35
5.18	Framing of the data integrated for the analysis of siderophore uptake in corynebacterial species.	35
I.1	Example of a SNP-table with toy data on genomes of 6 strains and their classification using CLASSICO.	63
I.2	Data visualization using Evidente.	67
I.3	Use case for Evidente with <i>Mycobacterium leprae</i> 's data presented by Schuenemann <i>et al.</i> (2018).	71
I.S1	Use case with the <i>Treponema pallidum</i> data from a study by Pla-Díaz <i>et al.</i> (2022) visualized using a cladogram.	77
II.1	Sketch of the computation of the three parameters <i>step height</i> , <i>step factor</i> and <i>enrichment factor</i> required for TSS prediction	83

II.2	Classification of TSS based on the distance to annotated genes as defined by Sharma <i>et al.</i> (2010)	84
II.3	Requirements identified for the improvement of the TSS prediction workflow	85
II.4	Sketch of different views of the <i>genome browser</i> of TSSpredator-Web.	87
II.5	Visual representation of the two modes of the genome browser to show data of multiple experiments.	88
II.6	Analysis of the overall distribution of TSS across conditions, classes and location in the genome for data collected for <i>E. coli</i> on four different conditions (one control and three treatment with antibiotics)	90
II.7	Analysis of primary TSS for <i>E. coli</i> across conditions, especially those TSS occurring only under the treatment with each antibiotic	91
II.8	Usage of the <i>genome browser</i> for the exploration and characterization of orphan TSS in <i>E. coli</i>	93
II.S1	Overview of the five predefined sets of thresholds for <i>Step height</i> , <i>Step factor</i> and <i>Enrichment factor</i> for the TSS prediction with TSSpredator.	103
II.S2	Overview of how different parameter sets influence TSS calling in TSSpredator.	104
II.S3	Screenshot of <i>drag & drop</i> process for file distribution	105
II.S4	Secondary structure of the UTR region of the gene <i>ugd</i> as predicted by RNAfold.	105
III.1	Overview of TSS-Captur	110
III.2	Evaluation of RNA classification for TSS-Captur.	114
III.3	Evaluation of TSS-Captur by treating pTSSs as oTSSs	114
III.4	Visual summary of TSS-Captur output for <i>S. coelicolor</i> 's data.	115
III.5	Distribution of start of most enriched motif in the promoter regions for <i>S. coelicolor</i> 's analyzed TSSs using MEME.	116
III.6	Runtime analysis of TSS-Captur	117
III.S1	Snapshot of the web-interface of TSS-Captur	124
III.S2	Snapshot of the interactive report of TSS-Captur visualizing <i>S. coelicolor</i> 's data	125
III.S3	Interactive report of TSS-Captur visualizing <i>S. coelicolor</i> 's transcript <i>antisense_192</i>	128
III.S4	Histogram (bin width = 50) showing the length differences between experimental results presented by Vockenhuber <i>et al.</i> (2011) and predicted transcripts with TSS-Captur for <i>S. coelicolor</i> .	128
IV.1	Schematic representation of <i>S. aureus</i> siderophores and metallophores, as well as their acquisition systems.	133
IV.2	Overview of metallophore synthesis and uptake potential in the genus <i>Staphylococcus</i> .	136
IV.3	Comparison of all incomplete BGCs related to the synthesis of staphyloferrin A and the reference BGC of <i>S. aureus</i> using the tool <i>clinker</i>	138
IV.4	Comparison of incomplete <i>cnt</i> BGCs of <i>S. durrellii</i> to the corresponding loci of <i>S. aureus</i> using <i>clinker</i>	139
IV.5	Comparison of novel BGCs identified in this study, known metallophore BGCs of <i>S. aureus</i> , and two uncharacterized BGCs from other studies using the tool <i>clinker</i> .	139
IV.6	UpSet plots for the strain analysis on four selected species: (A) <i>S. epidermidis</i> , (B) <i>S. hominis</i> , (C) <i>S. capitis</i> , and (D) <i>S. lugdunensis</i> .	143
IV.S1	Multiple sequence alignment of identified homologs of <i>HtsA</i>	154

IV.S2	Multiple sequence alignment of identified homologs of <i>SirA</i>	155
IV.S3	Multiple sequence alignment of identified homologs of <i>CntA</i>	155
IV.S4	Distribution of sequence identity of HtsA homologs in <i>S. epidermidis</i> when compared to the reference of <i>S. aureus</i> .	156
V.1	Overview of the dataset used for the computational characterization of the mechanistic siderophore uptake potential in <i>Corynebacteria</i> and <i>Staphylococcus aureus</i> ' samples	160
V.2	Overview of the pipeline PRESERVE	162
V.3	Overview of the downstream analysis used to identify possible transcription factor binding sites related to the regulation of the identified structural homologs.	164
V.4	Overview of the composition for the protein sequence clusters returned by MMseqs2.	166
V.5	Number of structural homologs identified via PRESERVE after clustering reinflation	166
V.6	Histogram on the distribution of LDDT scores for all structural homologs	168
V.7	Volcano plots summarizing the complete transcriptomic response for a control vs. treatment with the chelator EDDHA.	170
V.8	Structural alignment of gene <i>RS06480</i> from <i>C. accolens</i> and the reference lipoprotein <i>HtsA</i> from <i>S. aureus</i>	170
V.9	Heatmap showing median regularized log-transformed expression values per condition for a subset of genes grouped by orthogroups	171
V.10	Overview of the results for motif enrichment across the analyzed species.	173
V.S1	Number of lipoproteins (TAT/TATLIPO) per sample as predicted by SignalP6 (Teufel <i>et al.</i> 2022)	181
V.S2	Upset plot for shared lipoproteins as identified with the PRESERVE pipeline.	182
V.S3	Venn diagram comparing structural homologs identified by PRESERVE and lipoproteins associated with siderophore uptake based on GO terms.	182
V.S4	Overview of enriched GO terms of <i>Biological Process</i> ontology as computed using GO-Compass	187
V.S5	Tree maps summarizing the individual results of the gene ontology (GO) enrichment analysis with GO-Compass per species	188

List of Tables

I.S1	Excerpt of the SNP-table with toy data used for the visualizations of Evidente .	75
I.S2	Example of the input table for Evidente that contains the metadata of the analyzed samples	76
I.S3	Distributions of all SNPs regardless of their classification that are found within a gene within the contingency table.	76
II.1	Excerpt of the <i>MasterTable</i> showing the top 10 enriched primary TSS enriched based on step height and their associated genes	92
III.1	Evaluation of CNIT 's performance.	113
III.S1	Comparison of results of TSS-Captur with the study by Vockenhuber <i>et al.</i> (2011)	129
III.S2	Hits of <i>S. coelicolor</i> 's predicted ncRNA genes in the RFAM database	130
IV.1	Number of available strains from the GTDB and number of remaining assemblies after filtering for four selected species and quality (N50 > 50 kb, # contigs < 250).	134
IV.S1	Overview of species and assemblies used in the analysis. The classification from the NCBI and the GTDB is provided.	150
IV.S2	Overview of additional strains used to assess the variability of the antiSMASH results within selected species.	152
IV.S3	Analysis of binding site conservation of <i>htsA</i> , <i>sirA</i> and <i>cntA</i>	153
V.1	Origin of reference protein structures used for structural comparison	161
V.2	Overview of the number of reference lipoproteins present in <i>S. aureus</i> samples used for pipeline validation purposes	165
V.3	Overview of the identified structural homologs by gene and samples for <i>S. aureus</i> samples	167
V.4	Overview of the cluster representatives of <i>S. aureus</i> identified as structural homologs	167
V.5	Summary of differentially expressed genes, distributed by the magnitude of their log ₂ fold change	169
V.S1	Overview of corynebacterial and staphylococcal genome assemblies used for the identification of structural homologs as well as validation of the computational approaches	183
V.S2	Result of MMseqs2 search of reference lipoproteins within the <i>S. aureus</i> samples used for computational validation purposes	184
V.S3	Results of BLAST search against the UniProt database for all identified structural homologs of the <i>S. aureus</i> samples	185
V.S4	Genes identified by PRESERVE as structural homologs of the given references, but showing no significant expression changes under the RNA-seq experiment	186
V.S5	Genes identified as differentially up-regulated genes ($p\text{-adj.} < 0.01$ and $\log_2(\text{FC}) > 1$) and labeled by SignalP6 as lipoproteins, but not identified by PRESERVE as homologs of the reference proteins.	186

Chapter 1

Introduction

Prokaryotes are known for their ability to quickly adapt when exposed to hostile environments, such as antibiotic exposure, and develop phenotypes that allow them to survive under such hostile environments (Spratt [1994](#); Wright [2005](#)). This means that bacteria are gaining resistance against multiple known antibiotics, leading to one of the major health care issues of the new era: the antibiotic resistance crisis (Neu [1992](#)). Of the estimated 7.7 million annual deaths attributed to bacterial infections, drug-resistant pathogens are associated with about 65% and are directly linked to causing around 16% of them (Okeke *et al.* [2024](#)). To highlight pathogens of particular clinical concern, the World Health Organization (WHO) introduced the *ESKAPE* group, which includes bacterial species such as *Escherichia coli* and *Staphylococcus aureus* that are resistant to a large fraction of clinically used antibiotics.

While the rate of antibiotic resistant bacteria increases, the rate of discovery of new antibiotics has declined drastically in recent decades (Hegemann *et al.* [2022](#)). This has led researchers to pursue alternative strategies to control bacterial infections (Brown and Wright [2016](#)). Rather than relying on single antibiotics, combinatorial strategies have been explored to improve the efficacy of the compounds. These strategies include the use of antibiotic combinations, but also approaches that pair antibiotics with nutrients or specialized bacterial metabolites to promote drug uptake (Bognár *et al.* [2024](#); Yamano [2019](#)). In addition, therapies based on viruses that infect bacteria, called bacteriophages, have regained attention as an alternative way to control bacteria (Gordillo Altamirano and Barr [2019](#)). Lastly, researchers are also interested in understanding how bacterial interactions can be exploited for microbial controls by studying the composition of communities of microorganisms (Maier *et al.* [2024](#)). Despite their conceptual diversity, the successful development of these strategies largely depends on a detailed understanding of the molecular mechanisms underlying antibiotic resistance, as well as the general molecular mechanisms behind prokaryotic organisms. However, these processes do not operate in isolation but arise from interconnected activities across multiple molecular layers.

To provide a conceptual framework for this interplay, Francis Crick introduced the central dogma of molecular biology, which describes the unidirectional flow of biological information within a cell (Crick [1958](#); Crick [1970](#)). By linearly linking DNA, RNA and proteins, it was described how genetic information gives rise to a molecular function in a unidirectional manner (Fig. [1.1](#)). Genomic nucleotide sequences are transcribed into messenger RNAs (mRNAs), which are then translated into amino acid sequences that fold into functional proteins.

Although Crick already acknowledged that there may be exceptions to this information flow (Crick [1970](#)), further research has gone beyond his expected exceptions and expanded the original formulation of the central dogma (Miller *et al.* [2023](#); Shapiro [2009](#)). For example, Crick's proposed transfer of information from RNA to DNA or between RNA molecules was later confirmed by the discovery of RNA bacteriophages (Thieffry and Sarkar [1998](#)). In addition, the identification of various non-coding RNA (ncRNA) transcripts has demonstrated that RNA molecules play active regulatory roles, mediating interactions between RNA and DNA and shaping gene expression beyond protein-coding functions (Singh *et al.* [2018](#)). Recent studies have also shown that the central dogma does not fully capture the information flow initiated by proteins themselves (Fig. [1.1](#)).

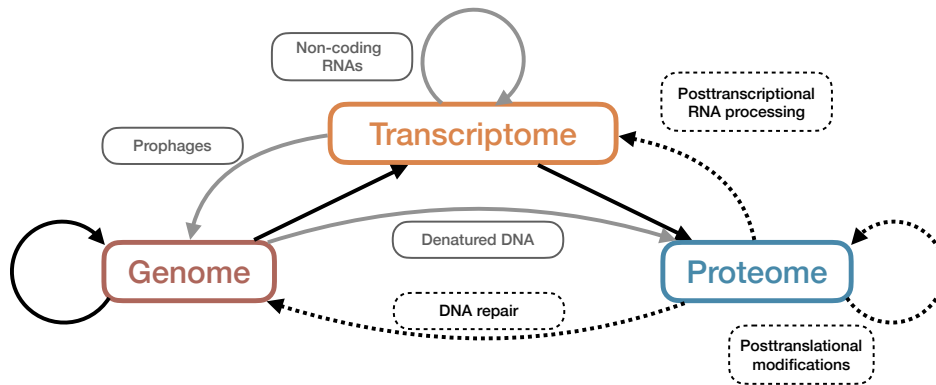


Figure 1.1: Visual representation of the central dogma of molecular biology, introduced first by Crick (1958). Black arrows represent information flows with clear evidence at time of publication of the central dogma, while gray arrows represent hypothesis that were described as possible by Crick. Dotted arrows represent extensions of the flow of information that have been verified in more recent studies (Miller *et al.* 2023)

Proteins play an important role in connectivity across layers by participating in processes such as DNA repair, post-transcriptional modifications of RNA, and post-translational modifications of other proteins, thus directly shaping molecular function beyond linear information transfer (Drazic *et al.* 2016; Nachtergaele and He 2017). Together, these findings show that molecular mechanisms depend on highly interconnected layers, particularly in prokaryotic systems.

This expanded view of the information flow in prokaryotic organisms has been driven by the increasing availability of high-throughput technologies, which enable the investigation of biological processes at unprecedented scale and resolution (Goodwin *et al.* 2016; Hasin *et al.* 2017). The large volume and complexity of the data generated by these technologies have made computational approaches an integral component of their processing and analysis (Gauthier *et al.* 2019). This methodological necessity often increases the number of players needed for a general analysis, as domain experts design and run high-throughput experiments, while often computation experts process the data into human-readable formats for domain experts to generate insights. Together, experimental advances and computational methods have led to distinct *omics* layers, each with the aim of systematically capturing molecular information by focusing on a specific aspect of cellular function (W. Zhang *et al.* 2010). While numerous *omics* layers have been described, such as epigenomics, metabolomics, and metagenomics, three align directly with the central dogma of molecular biology: genomics, transcriptomics, and proteomics. Advances within these layers have refined our understanding of prokaryotic systems, from genomic sequence information to gene expression and protein function.

At the genomic level, next-generation sequencing (NGS) technologies have enabled the reconstruction of complete genomes by generating large numbers of genomic fragments, commonly known as *reads* (Forde and O'Toole 2013). Subsequently, these reads are assembled into contiguous genomic sequences using computational approaches, either *de novo* approaches or guided by a reference genome (Flicek and Birney 2009). Together, these experimental and computational advances have made it possible to reconstruct complete genomes for a wide range of prokaryotic organisms, going beyond laboratory strains to include isolates from clinical environments (Besser *et al.* 2018).

The transcriptomic layer has similarly benefited from NGS-based approaches, as all

RNA transcripts can be sequenced and mapped to a corresponding reference genome (Sorek and Cossart [2010](#)). This enables genome-wide quantification of gene expression under multiple conditions and is not limited to previously annotated genes, in contrast to earlier high-throughput technologies such as microarrays (Croucher and Thomson [2010](#)). Additionally, further experimental protocols have been developed that allow a deeper understanding of the transcriptome, such as focusing on the ends of transcribed mRNA transcripts by depletion or enrichment of certain mRNA (Dar *et al.* [2016](#); Sharma *et al.* [2010](#)), or simultaneous analysis of the transcriptome of two species (Westermann *et al.* [2012](#)).

In addition, the proteomic level has profited through the development of high-throughput approaches, such as mass spectrometry-based technologies (Messner *et al.* [2023](#); Z. Zhang *et al.* [2014](#)). These approaches allow the quantification of known proteins and the characterization of post-translational modifications (PTMs) (Aebersold and Mann [2016](#)). Beyond protein abundance, further improvements in low-throughput techniques, such as X-ray crystallography and cryo-electron microscopy (cryo-EM), have enabled detailed analysis of protein structures, providing insight into the structurome of prokaryotic organisms (Wells and Marsh [2018](#)).

Although these advances have substantially improved the characterization of individual molecular layers, they also form the foundation for downstream computational analyses that extract even more biological information, especially as data is often shared in large repositories. This availability of large-scale biological datasets has allowed the development and refinement of computational methods. At the genomic level, these methods allow the identification of open reading frames (ORFs) based on known gene characteristics, thereby defining the protein-coding potential of an organism (Koonin [2001](#)). Beyond individual genes, higher-order analyses can resolve structural organizations such as polycistronic operons or biosynthetic gene clusters (BGCs), providing insight into co-regulation and functional relatedness (Medema *et al.* [2014](#); Rogozin [2004](#)). On the proteomic layers, a more recent breakthrough has been the training of machine learning models from experimentally confirmed protein structures to allow the accurate prediction directly from amino acid sequences (Jumper *et al.* [2021](#)). In addition, the easier access to experimental data, either via own generation or via databases, has allowed comparative approaches to rise. These analyses enable the investigation of variation between strains or species by identifying single-nucleotide polymorphisms (SNPs), which constitute a major driver of genomic diversity across bacterial populations, as well as reconstructing evolutionary relationships using phylogenetic methods (Bentley and Parkhill [2004](#); Ford *et al.* [2011](#)).

However, the full explanatory power of omics data unfolds when information is integrated between layers (W. Zhang *et al.* [2010](#)). For example, based on identified ORFs, mapped transcriptomic data can be used to quantify the expression of annotated genes, but also to refine the annotation through the identification of untranslated regions (UTRs) or even overlooked genes (Sorek and Cossart [2010](#)). It can even go further and combine processed data from multiple layers to go beyond individual insight generations. For example, isolated genomic analyses to identify polycistronic gene arrangements can imply co-expression and shared regulation potential. However, only overlaying transcriptomic data allows the identification that the co-expression of these structures is condition-dependent (Güell *et al.* [2009](#)). In a similar manner, transcriptomic and proteomic profiles represent snapshots of specific conditions, but rarely provide deeper insights into the underlying regulation layer. Hence, by including the genomic context, differences that alter regulatory behavior between strains can be identified, such as the occurrence of specific SNPs (Dugar *et al.* [2013](#)). By integrating such layers, researchers can better capture the complex molecular

networks underlying prokaryotic processes, as anticipated by the central dogma and its subsequent extensions.

These examples illustrate that in order to generate insights from multiple input data, many computational methods are required to first analyze the data individually and then integrate them. This complex processing requires continuous and clear communication between domain experts and computational scientists, as each iterative step often reveals new perspectives that need specialized knowledge for their interpretation. To simplify the communication process, new computational tools should move beyond the development of algorithms and provide more accessible ways for domain experts to perform integrative analyses independently.

A significant barrier for domain experts is the usability gap of scientific software (Mangul *et al.* 2019). While integrative methods are algorithmically robust, limited resources often prevent developers from focusing on user experience or accessibility. Hence, many state-of-the-art tools are released as command-line interfaces (CLIs), effectively restricting their use to researchers with significant technical expertise. Although some developers have introduced graphical user interfaces (GUIs) to their applications to lower this barrier (Joppich and Zimmer 2019), these often still require the installation of dependencies. More recently, web-based processing platforms have emerged, offering ready-to-go environments that allow domain experts to analyze data directly and accelerate hypothesis testing (Jalili *et al.* 2020).

However, these platforms tend to focus on narrow, hypothesis-driven tasks, such as gene expression quantification or sequence assembly, as they answer specific research question, hence they lack the flexibility required for the exploratory tasks essential for integrative analyses. Explorative tools should do more than identify correlations within the data, as they should facilitate the generation of new hypotheses via data exploration, for example, by increasing the transparency of the results. By developing interfaces that provide access to intermediate processing steps, researchers can better understand why specific results emerge. Such transparency is particularly critical for multi-omics integration, where the underlying biological mechanisms are difficult to grasp because of the complexity of the data.

Beyond understanding the algorithm, researchers must be able to explore both raw and processed data from multiple perspectives. Visual exploration is essential in this regard, as it takes advantage of human perception to detect patterns and anomalies that numerical summaries often fail to capture. A classic example is Anscombe’s Quartet (Anscombe 1973), which demonstrates how datasets with identical descriptive statistics can possess radically different structures when visualized (Fig. 1.2). This principle has been widely adopted in omics data analysis (Munzner 2014; Revell *et al.* 2018), leading to a variety of visualization tools that use data abstraction techniques to make complex information interpretable. Some tools focus on visualizing data from one omic layer, such as phylogenetic trees (Letunic and Bork 2019; Subramanian *et al.* 2019), or protein structures (DeLano *et al.* 2002), while others integrate multiple omic layers within one view (Thorvaldsdóttir *et al.* 2013). Nevertheless, these tools often only allow visualization of data that the user has previously processed, disconnecting the visual data from the analysis process.

To expand the scope of visual analyses, the field of bioinformatics has seen the rise of visual analytics tools, which combine computational power with interactive visualization to transform raw data into actionable biological insights (Cui 2019). By integrating automated analysis with interactive visual representation, visual analytics creates a human-in-the-loop system. This approach enforces the idea that while computational methods are necessary for processing high-dimensional data, humans remain superior at identifying biological

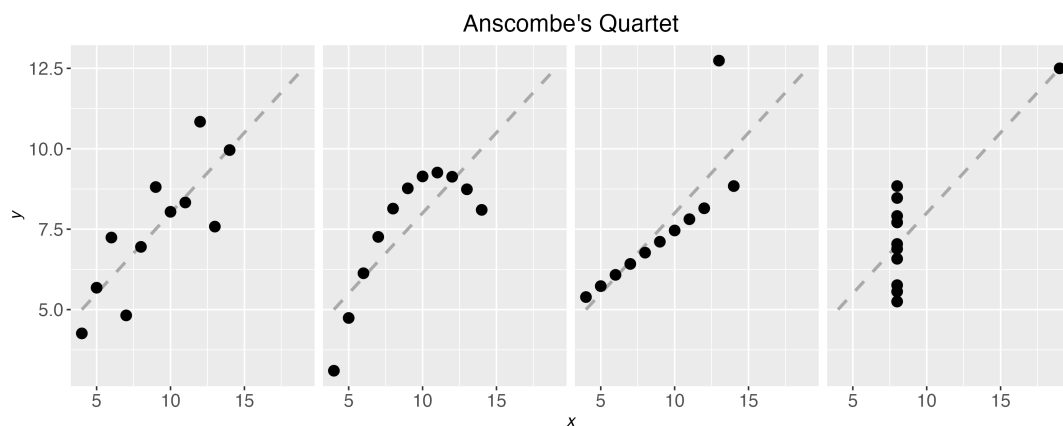


Figure 1.2: Visual representation of Anscombe’s quartet (Anscombe [1973](#)). All four data sets consist of eleven observed paired data points (x and y). When summarized using common descriptive statistics and a linear regression model, all datasets return identical values ($\bar{x} = 9.0$, $\bar{y} = 7.5$, $R^2 = 0.667$). However, the visualization of the data reveals how the distributions differ substantially.

contexts and unexpected relationships between the data. In the context of multi-omics, this approach is particularly potent as it allows researchers to create multiple perspectives of the data, for example by an interactive data selection or parameter adjustment, and immediately observe how these changes shift the resulting biological landscape.

While accessibility and visual exploration are essential to reduce barriers and improve insight generation, they must be complemented by reproducibility. For insights to be validated, shared, and extended by the scientific community, integrative workflows must be transparent and repeatable (Mangul *et al.* [2019](#)). This requires the use of modern workflow management systems that formalize data processing steps and enable scalable execution across computational infrastructures for the consistent integration of the omics layers. In bioinformatics, Nextflow has emerged as a state-of-the-art framework for implementing such workflows (Di Tommaso *et al.* [2017](#)). Beyond ensuring consistent integration, Nextflow supports containerization technologies such as Docker and **conda**, thus enhancing portability and long-term reproducibility, regardless of the computational environment.

These methodological requisites should be regarded essential for addressing the antibiotic resistance crisis, as they allow researchers to move beyond the analysis of isolated omic layers toward an integrated perspective of prokaryotic systems. Consequently, the development of transparent, reproducible, and exploratory computational environments represents a suitable approach to capture the interplay between genomic variation, transcriptional regulation, and protein structure and function. In this context, the central dogma of molecular biology provides both a conceptual anchor and a practical roadmap for the systematic characterization of bacterial systems, forming a foundation for the rational development of alternative microbial control strategies that extend beyond the use of classical antibiotics. In the following chapters, the specific methodological and conceptual gaps of the current state of computational methods are described, followed by the developed solutions on how they have been addressed to improve research on microbial control.

Chapter 2

List of Papers

Paper I

Witte Paz, M., Harbig, T. A., and Nieselt, K. (2022). “Evidente—a Visual Analytics Tool for Data Enrichment in SNP-based Phylogenetic Trees”. In: *Bioinformatics Advances* vol. 2, no. 1, vbac075. DOI: [10.1093/bioadv/vbac075](https://doi.org/10.1093/bioadv/vbac075)

Paper II

Witte Paz, M., Herbig, A., and Nieselt, K. (2026). “TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration”. In: *PLOS ONE* vol. 21, no. 3, e0326299. DOI: [10.1371/journal.pone.0326299](https://doi.org/10.1371/journal.pone.0326299)

Paper III

Witte Paz, M., Vogel, T., and Nieselt, K. (2024). “TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA Transcripts”. In: *NAR Genomics and Bioinformatics* vol. 6, no. 4, lqae168. DOI: [10.1093/nargab/lqae168](https://doi.org/10.1093/nargab/lqae168)

Paper IV

Witte Paz, M., Bitzer, A., Nieselt, K., and Heilbronner, S. (2025). “A Landscape of Metallophore Synthesis and Uptake Potential of the Genus *Staphylococcus*”. In: *NAR Genomics and Bioinformatics* vol. 7, no. 4, lqaf183. DOI: [10.1093/nargab/lqaf183](https://doi.org/10.1093/nargab/lqaf183)

Paper V

Witte Paz, M., Bitzer, A., Heilbronner, S. and Nieselt, K. “[Characterizing Siderophore Uptake and their Regulation Mechanisms in Corynebacteria based on Structural Homology and Transcriptome Profiling](#)”.

Manuscript in progress

2. List of Papers

Publications not included in this thesis

Harbig, T. A., **Witte Paz, M.**, and Nieselt, K. (2023). “GO-Compass: Visual Navigation of Multiple Lists of GO Terms”. In: *Computer Graphics Forum* vol. 42, no. 3, pp. 271–281. DOI: [10.1111/cgf.14829](https://doi.org/10.1111/cgf.14829)

Brich, N., Harbig, T. A., **Witte Paz, M.**, Nieselt, K., and Krone, M. (2024). “ProtE-GOnist: Visual Analysis of Interactions in Small World Networks Using Ego-graphs”. In: *Computer Graphics Forum* vol. 43, no. 3, e15078. DOI: [10.1111/cgf.15078](https://doi.org/10.1111/cgf.15078)

Hackl, S., Jachmann, C., **Witte Paz, M.**, Harbig, T. A., Martens, L., and Nieselt, K. (2025). “PTMVision: An Interactive Visualization Webserver for Post-translational Modifications of Proteins”. In: *Journal of Proteome Research* vol. 24, no. 2, pp. 919–928. DOI: [10.1021/acs.jproteome.4c00679](https://doi.org/10.1021/acs.jproteome.4c00679)

Rahmdel, S., Türkoglu, T., Nikjoo, N., Babaali, E., Moradi Mirhesari, D., Nega, M., Brüggemann, H., Huang, L., **Witte Paz, M.**, Nieselt, K., and Götz, F. (2025). “Beyond Commensalism: Genomic Insights into Micrococcin P1-producing *Staphylococcus Chromogenes*”. In: *mSphere*, e00733–25. DOI: [10.1128/msphere.00733-25](https://doi.org/10.1128/msphere.00733-25)

Chapter 3

Contributions

Paper I

Witte Paz, M., Harbig, T. A., and Nieselt, K. (2022). “Evidente—a Visual Analytics Tool for Data Enrichment in SNP-based Phylogenetic Trees”. In: *Bioinformatics Advances* vol. 2, no. 1, vbac075. DOI: [10.1093/bioadv/vbac075](https://doi.org/10.1093/bioadv/vbac075)

Mathias Witte Paz and Kay Nieselt conceived the concept of **Evidente**. **Mathias Witte Paz** developed the first prototype. **Mathias Witte Paz** and Theresa Harbig contributed both to the development of further features. **Mathias Witte Paz** conducted the use cases and created the figures. All authors wrote and reviewed the manuscript.

Paper II

Witte Paz, M., Herbig, A., Nieselt, K. “**TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration**”.

Alexander Herbig and Kay Nieselt conceived the idea for the TSS prediction **TSSpredator**. **Mathias Witte Paz** and Kay Nieselt conceived the idea for the web-application **TSSpredator-Web**. **Mathias Witte Paz** worked on the implementation and deployment of **TSSpredator-Web**, conducted the use cases, created the figures and wrote the original draft of the manuscript. All authors commented on the draft and reviewed the final version of the manuscript.

Paper III

Witte Paz, M., Vogel, T., and Nieselt, K. (2024). “TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA Transcripts”. In: *NAR Genomics and Bioinformatics* vol. 6, no. 4, lqae168. DOI: [10.1093/nargab/lqae168](https://doi.org/10.1093/nargab/lqae168)

Mathias Witte Paz and Kay Nieselt conceived the idea for **TSS-Captur**. **Mathias Witte Paz** implemented the first pipeline version and conducted the use cases. Based on the first version and under the supervision of **Mathias Witte Paz**, Thomas Vogel implemented the **Nextflow DSL2** version of the pipeline and the web interface. **Mathias Witte Paz** was responsible for the deployment of the web-application, wrote the original draft of the manuscript and created the figures. All authors commented on the draft and reviewed the final version of the manuscript.

Paper IV

Witte Paz, M., Bitzer, A., Nieselt, K., and Heilbronner, S. (2025). “A Landscape of Metallophore Synthesis and Uptake Potential of the Genus *Staphylococcus*”. In: *NAR Genomics and Bioinformatics* vol. 7, no. 4, lqaf183. DOI: [10.1093/nargab/lqaf183](https://doi.org/10.1093/nargab/lqaf183)

All authors conceived the idea. **Mathias Witte Paz** conducted the computational analyses and created the figures for the results. **Mathias Witte Paz** and Simon Heilbronner wrote the first draft of the manuscript. All authors reviewed the final manuscript.

3. Contributions

Paper V

Witte Paz, M., Bitzer, A., Heilbronner, S. and Nieselt, K. “Characterizing Siderophore Uptake and their Regulation Mechanisms in Corynebacteria based on Structural Homology and Transcriptome Profiling”.

Simon Heilbronner and Kay Nieselt conceived the idea to characterize the siderophore uptake mechanisms of *Corynebacteria* samples. **Mathias Witte Paz** and Simon Hackl jointly developed the **PRESERVE** pipeline applied to the data set of this study. The work was distributed as follows: Simon Hack conceptually developed **PRESERVE** to include the clustering of multiple sequences and the selection of a representative sequence for its structural prediction. **Mathias Witte Paz** contributed to the development of **PRESERVE** by including the option of predicting lipoproteins, further downstream processes, and the comparison of the predicted structure to a set of structural references. Moreover, **Mathias Witte Paz** developed the first version of the Nextflow pipeline of **PRESERVE**. For this study, **Mathias Witte Paz** conducted all analyzes using the **PRESERVE** pipeline, as well as subsequent analyzes, such as the computational evaluation of the identified homologs, analysis of RNA-seq data, and motif enrichment analyses. Alina Bitzer conducted part of the experiments used for the creation of metadata of the dataset. All authors discussed and evaluated applied methodologies. **Mathias Witte Paz** conceptualized the manuscript, wrote the original draft, and prepared all the figures. **Mathias Witte Paz** and Kay Nieselt reviewed and edited the original draft.

Chapter 4

Objectives

The central dogma of molecular biology describes the flow of information from the genome to the transcriptome and ultimately to functional molecules, providing a conceptual framework for understanding cellular processes. With the increase in the generation of omics data, the integration of data across the layers of the central dogma of molecular biology has been shown to be essential to fully understand the mechanisms and behavior of prokaryotic organisms. Hence, the main goal of this thesis is to develop applications that support the integration of biological data originating from different layers of the central dogma, with a particular focus on genomics and transcriptomics, to facilitate insight generation in the context of microbial control. Such integrative approaches can support the investigation of processes such as antibiotic resistance, transcriptional regulation, and interaction mechanisms that shape the cohabitation of different organisms. In addition to methodological integration, this thesis emphasizes the development of tools that are accessible to end users, particularly researchers with limited computational expertise, by providing intuitive and user-friendly interfaces for data exploration. Based on these goals, more specific objectives were defined to address key challenges in the analysis and integration of data across the different layers of the central dogma, as outlined in the following sections.

As described by the central dogma of molecular biology, the genomic layer provides the foundation for the downstream layers. Accordingly, the characterization of prokaryotes can reasonably begin at the genomic level, for example through the analysis of one of the most frequently observed forms of variation: single nucleotide polymorphisms (SNPs). However, due to the fast evolution of bacterial genomes, the interpretation of SNPs requires an explicit evolutionary context. This motivated the development of **Evidente**, a visual analytics tool designed for the integrative exploration of SNP and phylogenetic data. **Evidente** combines these data types by computing the distribution of mutations across a phylogenetic tree and classifying them according to their clade-specificity. In addition, the tool supports the integration of metadata, including sample-specific descriptors, as well as functional annotations such as Gene Ontology terms, enabling an integrative analysis with genomic variation. Within this project, an objective of **Evidente** was to support systematic exploration of genotype–phenotype associations, such as those related to antibiotic resistance, to facilitate the characterization of prokaryotic organisms.

Although the main focus of **Evidente** was on the integration of genomic and phylogenetic information, a clear need to move to the second layer and integrate transcriptomic data became evident. While many computational approaches emphasize the analysis of gene expression levels, there is a gap in the exploration of the transcriptome architecture. Here, this project focused on the prediction of transcription start sites (TSS) in prokaryotes. Although current tools, such as **TSSpredator** (Dugar *et al.* 2013), manage to integrate experimental genome-wide transcriptomic data for the prediction of TSS, they are often limited with respect to their user-friendliness and integrative exploration of results. To tackle this issue, we developed **TSSpredator-Web**, a web-application for the prediction and exploration of TSS data based on the tool **TSSpredator**. Providing such a web-application facilitates access to the tool to multiple users and adds the possibility of exploring the data, for example, by the integrated visualization of transcriptomic and the underlying genomic data. To align with the goal of prokaryote characterization, an

4. Objectives

objective of this project was to apply **TSSpredator-Web** to a dataset characterizing the transcriptome of the *ESKAPE* organism *Escherichia coli* under antibiotic treatment and allow the exploration of condition-specific transcriptional landscapes.

Genome-wide TSS studies based on tools such as **TSSpredator-Web** have revealed multiple genomic regions with strong transcriptional signals that lack corresponding annotation, suggesting the presence of overlooked genes. This observation motivated the integration of experimentally determined TSS positions with their underlying genomic sequences in an automated and high-throughput manner. To address this gap, the **TSS-Captur** pipeline was developed. The objective of **TSS-Captur** is to apply a combination of bioinformatic methods to allow a systematic characterization of these putative functional regions. Beyond characterization, the process behind **TSS-Captur** should be accessible and allow the exploration of results, for example, via a web application, targeting primarily users with limited technical expertise.

While the objectives so far were focused on data-integration techniques and user-friendly tool development, the next objectives addressed a central biological question: how can bacteria uptake metallophores. As metallophores aid bacteria to obtain essential metals for proliferation, characterizing the mechanistic process behind their synthesis and uptake provides insight for the development of novel microbial control strategies. This mechanism was characterized from two different perspectives. The first perspective aimed to assess the potential for metallophore synthesis and uptake within the genus *Staphylococcus*, which includes both human and animal colonizers, and especially the *ESKAPE* pathogen *S. aureus*. Using genomic data, this objective focused on the integration of multiple computational analyses to characterize the diversity of metallophore-related systems, with an additional evolutionary context of the genus.

A characteristic of metallophores is that, once secreted, they can be utilized by other species present in the same environment. For example, members of the evolutionary distant genus *Corynebacterium* have been shown to inhibit the growth of *Staphylococcus aureus* through the exploitation of its siderophores, a specialized class of metallophores involved in iron acquisition. As the molecular mechanisms underlying this interaction remain largely unresolved, one objective of this project was to characterize the potential of distantly related species to exploit xenosiderophores by identifying lipoproteins known to be involved in the uptake process. As we observed that sequence homology is insufficient for inter-genera comparisons, the goal was to derive structural protein information from genomic sequences to enable structure-based homology evaluation using the pipeline **PRESERVE**. To contextualize these findings, transcriptomic data were incorporated to support the exploration of species-specific transcriptional responses associated with siderophore uptake. In addition, putative promoter regions of the identified structural homologs were analyzed to investigate their regulatory characteristics of the transcriptome. Together, these objectives aim to establish a reproducible framework for the evaluation of cross-genera homology and to nominate candidate genes for subsequent experimental validation, by accounting for genomic, genome-wide transcriptomic and protein structural data.

Chapter 5

Results

The main goal of this thesis is to support research on the control of microbes by providing methods for the characterization of prokaryotic organisms. This is achieved by analyzing data from single or multiple layers around the central dogma of molecular biology, either by developing new applications or extending established approaches, applying them to provided data, and, in most cases, providing new ways of exploring the results. Five projects are presented within this thesis, while a complete list of publications throughout my doctoral studies is available on page [7](#). The following sections will introduce the main results of the individual projects and discuss to what extent they helped achieve the determined objectives of the thesis.

5.1 Integrating genomic and phylogenetic data for SNP diversity exploration with **Evidente**

This section presents the article **Witte Paz, M.**, Harbig, T. A., and Nieselt, K. (2022). “Evidente—a Visual Analytics Tool for Data Enrichment in SNP-based Phylogenetic Trees”. In: *Bioinformatics Advances* vol. 2, no. 1, vbac075. DOI: [10.1093/bioadv/vbac075](https://doi.org/10.1093/bioadv/vbac075). For the full publication, refer to [Paper I](#).

Prokaryotic genomes evolve rapidly, resulting in a high level of genetic diversity, even between individuals of the same species (Arber [2000](#)). While the evolution process might result in large-scale genomic changes, such as insertions, deletions, and recombination events, small mutations like *single-nucleotide polymorphisms* or *variants* (SNPs/SNVs, hereafter called SNPs) play a major role in the rise of significant phenotypic variation (Ko and Choi [2021](#); Vignal *et al.* [2002](#)). The increased availability of whole-genome sequences now allows researchers to identify such mutations in a genome-wide manner, and even infer the evolutionary relationships based on SNP-based phylogenetic trees (Faison *et al.* [2014](#)). These trees provide a powerful tool for tracing lineage divergence that can be used to characterize a set of prokaryotic strains from one species. However, the analysis of the evolutionary history often focuses only on the computed tree, while ignoring the underlying genomic variations, despite its integration being crucial for further insights. For example, one could identify specific SNPs contributing to evolution (e.g. via uniqueness in a specific clade), or identify correlations between bacterial strains with specific characteristics, such as geographic origin, sampling date, or other phenotypic traits. For such cases, it is also important to know whether the SNP is located within a gene, as this might influence the phenotype. Hence, further genomic data, such as the position of coding genes, can be included for a more complete analysis of the set of bacteria. By exploring these different data sources within one common interface, hypotheses can be generated about genotype–phenotype associations or clade-specific adaptations. However, existing tools rarely support the visual exploration of such integrative datasets.

To allow such exploration, **Evidente** focuses on the integration of multiple data originating mainly from the first layer of the central dogma: the genome. **Evidente** allows joint visual exploration of SNP-based phylogenetic trees, genome-wide SNP profiles, together with strain-specific characterization via metadata files, and gene information via Gene Ontology (GO) terms. The main difference between **Evidente** and other phylogenetic tree

5. Results

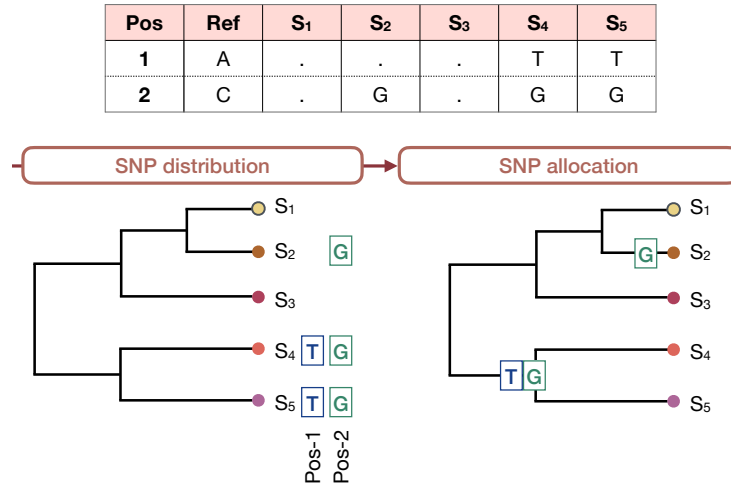


Figure 5.1: Sketch on **CLASSICO**’s algorithm for the identification of clade-specific SNPs. The tabular SNP data (upper table) represents the input for **CLASSICO** and indicates the presence of SNPs by showing the observed allele at each genomic position and sample. A dot represents the presence of the reference allele at that position, i.e. no SNP is found. Based on this data, SNPs are mapped onto the leaves of a phylogenetic tree based on their presence in each sample (bottom left). **CLASSICO** propagates each SNP from the leaves towards the root only if all child nodes of an internal node share the same mutation (right). If a SNP is found only in a single internal node, it is considered clade-specific or supporting (Position 1). If a SNP is found in multiple internal nodes, it is labeled non-supporting (e.g., Position 2). This figure has been adapted from the original publication (see Fig. [I.1](#)).

visualizers, such as **Evolview** (Subramanian *et al.* [2019](#)) or **iTOL** (Letunic and Bork [2019](#)), is the underlying integration and categorization of SNP data into the visual interface. This is computed via **CLASSICO** (CLAdE-Specific Snp IdentifiCatOr). **CLASSICO** is an efficient algorithm that classifies each SNP from a genome-wide analysis into either supporting or non-supporting for a given phylogenetic tree (Fig. [5.1](#)). *Supporting* or *clade-specific* SNPs are those positions where the algorithm terminates as only one node contains the corresponding mutation. This means that the specific SNP occurs only within members of one clade (see position 1 in Fig. [5.1](#), bottom-left). Otherwise, *non-supporting* SNPs are those positions where, after the propagation has ended, the SNP is found in more than one node (see position 2 in Fig. [5.1](#), bottom-right).

While the processing of the SNP data via **CLASSICO** achieves the first integration within **Evidente**, it is difficult to connect the SNP data to the underlying tree in an exploratory manner. First, studies focusing on the genomic diversity of bacteria typically involve numerous strains with extensive SNP variation (Arora *et al.* [2016](#); Schuenemann *et al.* [2018](#)), producing long lists as the output of **CLASSICO**. Moreover, the size and complexity of both the SNP-clade allocations and the phylogenetic tree make these output files difficult to interpret manually. Hence, we solve this problem by implementing a visualization dashboard for **Evidente**, consisting of four panels (Fig. [5.2](#)). To integrate the views of the phylogenetic tree and the SNP data, an interactive dendrogram visualization (Fig. [5.2I](#)) is aligned to a heatmap (Fig. [5.2II](#)) that visualizes the presence of SNPs that have been selected by the users. Color is used to encode the allele of a SNP, while the clade specificity is visualized by using four categorical colors to represent the four possible nucleotides. For supporting SNPs, the respective heatmap cell is filled, while for non-supporting SNPs only the borders of the heatmap cell are colored. A third heatmap is aligned to visualize any metadata provided by the user into the visualization (Fig. [5.2III](#)), and hence integrated in

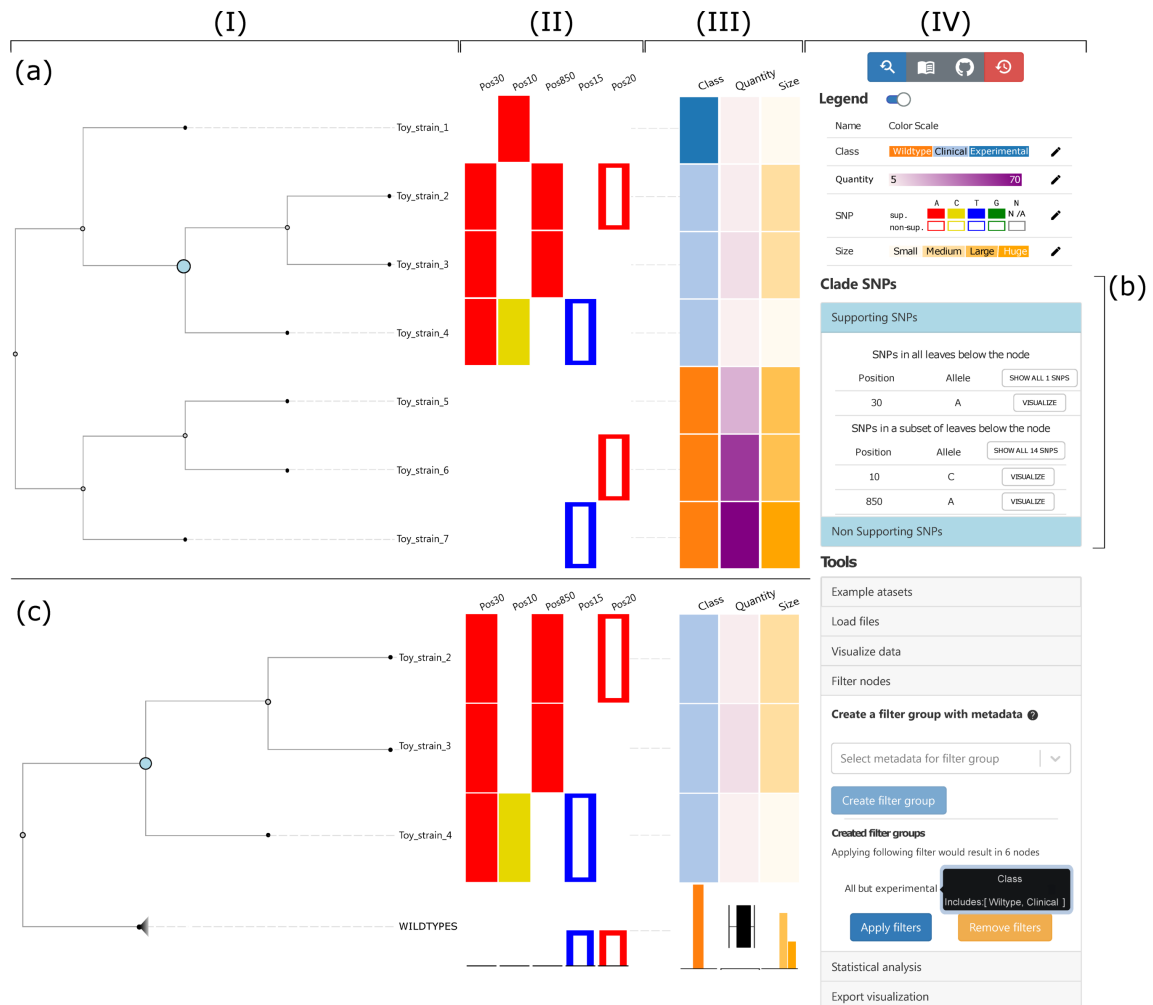


Figure 5.2: Visualization of a toy dataset using **Evidente**. The interface is divided into four components: the main visualization panel (I–III) and the sidebar (IV). Panel (I) displays the phylogenetic tree, while panels (II) and (III) show heat maps for selected SNPs and metadata, respectively. The tool’s sidebar (IV) visualizes the legend for the color scales used and further tool menus for interaction. (a) The main visualization shows some SNPs and metadata of a toy data set. A clade node (blue) has been selected. The first SNP (position 30) is a supporting SNP found in all leaves of the clade. The next two SNPs (positions 10 and 850) are supporting but appear only in a subset of clade members. The final two SNPs (positions 15 and 20) are non-supporting. (b) The sidebar shows SNPs for the selected clade, which can be added to the heat map. (c) Demonstration of the aggregation feature: samples labeled as *Wild Type* (orange rectangles in the first column of the metadata heat map) have been collapsed into a single node, renamed *WILDTYPES*. The end node of a collapsed node is emphasized by a specific glyph (a shaded triangle). Aggregated metadata is shown using box plots (for numerical values) or bar plots (for categorical and SNP data). Moreover, the filter group “All but experimental” was created as seen in the corresponding menu. It excludes all samples not classified as *Wild Type* or *Clinical*, hiding strain *Toy_strain_1* from the view. This figure and caption have been adapted from the original publication. The original version can be found in Fig. [1.2](#)

5. Results

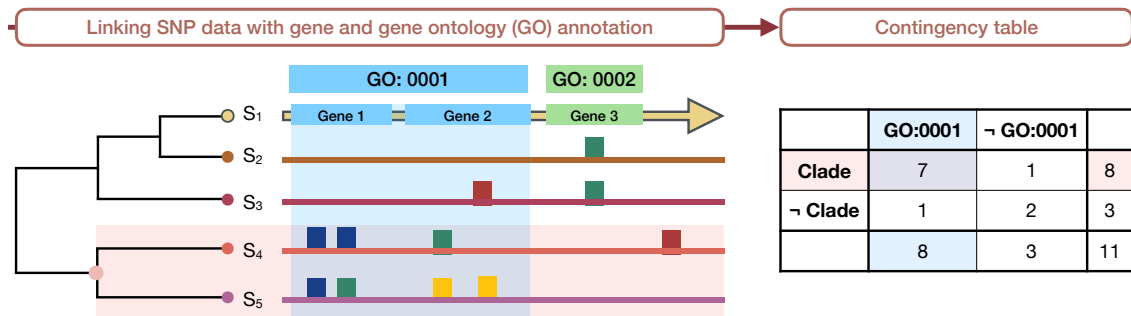


Figure 5.3: Schematic overview of the GO-term enrichment analysis in **Evidente**. It illustrates how SNPs are linked to annotated gene regions and their associated Gene Ontology (GO) terms (left), enabling a functional interpretation of genomic variation. Each SNP located within a gene is annotated with the corresponding GO terms (light-blue highlight). Also the SNPs within the selected clade are counted (red internal node and corresponding highlight) For each GO term, a contingency table is constructed (right) comparing the number of SNPs associated to the term found in the selected clade, versus those outside the clade.

the exploration process. Following the attribute type definitions by Munzner (2014), any sample-related data classified as categorical, ordinal, or numerical can be visualized using an appropriate color map. These three panels represent the core of **Evidente**'s visual integrative approach, enabling users to visualize SNP data, phylogenetic relationships, and sample metadata and explore possible correlations through visual alignment (Fig. 5.2a).

The strength of **Evidente** lies not only in the visualization of the data, but also in enabling its interactive exploration, allowing users to iteratively refine their analyses. Users can select clades in the dendrogram to inspect **CLASSICO**-allocated SNPs and add SNPs to the visualization (Fig. 5.2b). Standard interactions such as filtering and clade aggregation are supported, and the aligned heatmap adapts accordingly by summarizing the data into type-specific plots (Fig. 5.2c). Additional visualization, filtering, and export options are provided in the sidebar (Fig. 5.2IV).

The last data-integration approach in **Evidente** is to assess whether SNPs within a specific clade are found more often within genes associated with particular functions. To perform this analysis, users provide a genome annotation file for the reference organism along with gene-associated GO terms. These inputs are used to perform a GO enrichment analysis with respect to a selected phylogenetic clade, as follows (see Fig. 5.3 for an overview). With this functionality, **Evidente** allows users to inspect whether specific SNPs might be affecting biological processes and how this could be linked with possible phenotypes of the strains.

The data integration approaches developed within **Evidente** were evaluated using two prokaryotic datasets. In the first case, the analysis of ancient and modern samples of *Mycobacterium leprae* (Schuenemann *et al.* 2018) showed how **Evidente** can combine phylogenetic and SNP data to identify SNPs uniquely associated with specific clades. By additionally integrating metadata on sampling date and geographical origin, such clade-specific SNPs become more informative. For example, **Evidente** facilitated the identification of clades related to ancient samples. Moreover, the tool was used to identify SNPs that enforced a hypothesis of the original publication, stating that the diversification of *M. leprae* should be considered differently. Hence, using **Evidente** could allow researchers to trace pathogen evolution across time and location, and to pinpoint mutations that may have emerged under these contexts. The second use case focused on the analysis of multiple strains of the pathogen *Treponema pallidum* (Arora *et al.* 2016). Unlike the

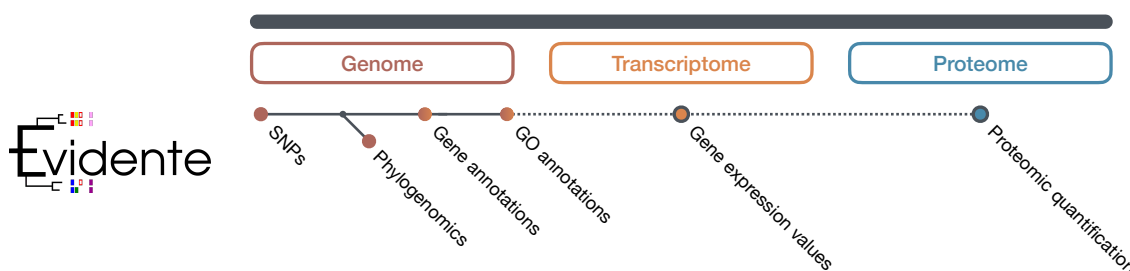


Figure 5.4: Framing the data integrated in **Evidente** using the central dogma of molecular biology as a framework. The data types (colored circles) are distributed onto the genomic, transcriptomic, and proteomic layers, depending on their origin. Circles with a solid border represent potential inputs with metadata inclusion that might expand the outreach of **Evidente** (dashed lines). Multicolored circles indicate data can link multiple layers.

previous case, **Evidente** facilitated the analysis of correlations between non-supporting SNPs and antibiotic resistance phenotypes by facilitating a side-by-side visualization of SNPs and metadata. Together, these use cases show how **Evidente** facilitates the process of integrating multiple genomic data together with metadata in order to characterize bacteria, enabling the translation from evolutionary analyses to microbial control strategies.

In order for other researchers to benefit from it, **Evidente** has been deployed as a web-application as part of the Tübingen Visualization Server *TueVis*¹.

Framing the inputs and processed data within the central dogma (Fig. 5.4) clearly shows that SNPs and phylogenetic trees constitute the foundation of its approach. Although phylogenomic data was not explicitly considered in the original formulation of the central dogma, it can still be conceptually assigned to the genomic layer, since phylogenetic reconstructions are derived from individual genomes. Beyond SNP and phylogenetic data, **Evidente** also integrates additional data, such as gene annotations and Gene Ontology terms. Though these two aspects are based on the genomic sequence, they establish a first connection to the transcriptomic layer, since they describe the functional potential of the organism. Furthermore, due to the general structure of the metadata input, it can be used to extend the scope of **Evidente** toward higher layers of the central dogma. Although this functionality has not yet been demonstrated, users could provide transcriptomic or proteomic expression data from the analyzed strains and visualize it in the interface. However, the feasibility of such large-scale integration across these layers remains challenging, mainly due to data availability and experimental costs.

In summary, **Evidente** shows how integrative analyses of genomic data and metadata can facilitate the gain of knowledge about microbes by focusing on an evolutionary perspective. Its strength lies in combining phylogenetic trees with SNP-level resolution, allowing the identification of clade-specific mutations, functional enrichments, and correlations with other data, such as phenotypic data.

However, while such evolutionary analyses of genomic variation are essential for understanding the origin of genomic differences, genomic information alone cannot fully explain the phenotypic differences observed between closely related strains. In the context of the central dogma, this highlights the need to complement evolutionary genomics with information from other molecular layers. Hence, **Evidente** also supports the integration of processed data from higher layers, such as quantification of the transcriptomic or proteomic layer. This observation led to the development of the next methods that integrate genome data with higher layers of the central dogma, in this case, the transcriptome.

¹ Accessible via <https://evidente-tuevis.cs.uni-tuebingen.de/>

5. Results

5.2 Prediction and Exploration of Transcription Start Sites with TSSpredator-Web

This section presents the article **Witte Paz, M.**, Herbig, A., and Nieselt, K. (2026). “TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration”. In: *PLOS ONE* vol. 21, no. 3, e0326299. DOI: [10.1371/journal.pone.0326299](https://doi.org/10.1371/journal.pone.0326299). For the full manuscript, refer to [Paper II](#).

As previously stated, gene annotation is a first option to bridge the genomic and the more functional layers of the central dogma: the transcriptome and the proteome (Stein [2001](#)). In the case of the transcriptome, genome annotation enables the interpretation of genomic sequences with respect to their functional contribution to a specific environment via RNA-seq experiments and bioinformatic downstream analyses (Lowe *et al.* [2017](#)). However, neither standard approaches for genome annotation nor standard RNA-seq experiments allow researchers to identify a key position for further insights into the transcriptomic layer: the exact position of the transcription start site (TSS, Backofen *et al.* [2014](#); Sorek and Cossart [2010](#)). While gene annotation methods often miss the TSS by focusing on the identification of coding sequences, expression profiles based on typical RNA-seq (i.e. a mixture of different processing stages) are too noisy for the identification of base-specific TSS. However, identifying TSS can help to understand the important interplay between the genomic and transcriptomic layers. For example, it can allow the definition of genomic elements, such as promoter regions and 5' untranslated regions (UTRs), that play a crucial role in transcriptome regulation (Sorek and Cossart [2010](#)).

To tackle the gap in the identification of TSS, experimental protocols for prokaryotes were developed, such as differential RNA-seq (dRNA-seq, Sharma *et al.* [2010](#)), Cappable-seq (Ettwiller *et al.* [2016](#)), or tagRNA-seq (Innocenti *et al.* [2015](#)). These protocols derived from the standard RNA-seq approach allow the sequencing of primary transcripts (i.e., those not yet processed during transcription), and hence the computation of profiles that are skewed towards the TSS (Fig. [5.5](#), left). With these, researchers can compare standard and enriched expression profiles to identify positions that show a prominent signal for a clear TSS. While the annotation can be pursued manually, multiple tools have been developed to process the data (Amman *et al.* [2014](#); Čuklina *et al.* [2016](#); Jorjani and Zavolan [2013](#)), with TSSpredator as one representative of these algorithms (Dugar *et al.* [2013](#)). TSSpredator detects TSS by comparing enriched and non-enriched RNA-seq profiles across genomic positions and computing different parameters based on this comparison (Fig. [5.5](#), left). If at a specific position, the computed parameters surpass a set of predefined thresholds, a TSS signal is called and it will be classified with respect to nearby genes as primary, secondary, internal, antisense, or orphan (Fig. [5.5](#), right).

But even though the characteristics of TSSpredator' have helped scientists describe the transcriptome of several prokaryotic organisms (Cheah *et al.* [2024](#); Ryan, Bornet, *et al.* [2024](#); Wolański *et al.* [2025](#); Wolfram-Schauerte *et al.* [2024](#)), it has limitations. Originally designed as a command-line tool with a graphical user interface, TSSpredator necessitates some technical knowledge, such as fixing installation dependencies. Hence, the audience that could potentially benefit from this tool is reduced as domain experts, the intended end users, should not be expected to have such an experience. Furthermore, TSS prediction requires multiple input files to be manually allocated into standard and enriched expression profiles, but also into the genomic strand of origin. Lastly, users cannot explore TSS data within the same tool and integrate other aspects, such as genomic sequences and annotation. By removing these hurdles, the tool becomes accessible to a wider range of researchers, but also facilitates the scaling to align with the rapid growth of RNA-seq data as well as with

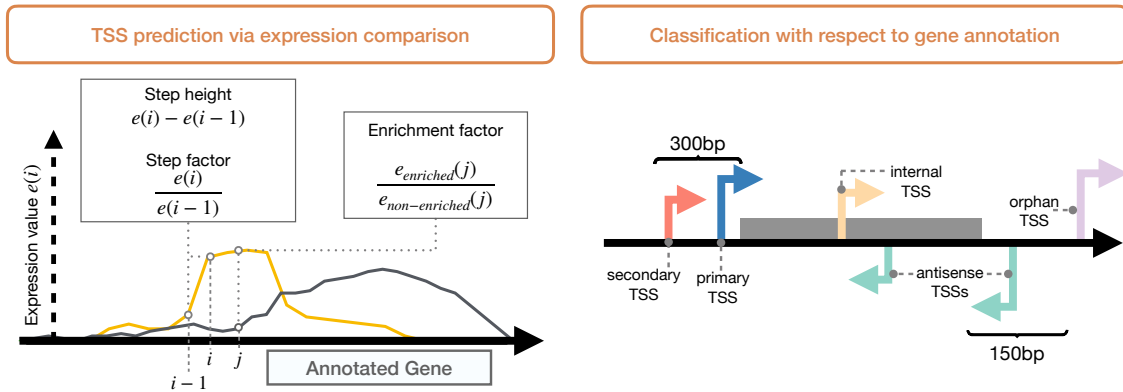


Figure 5.5: Overview of the computational goals achieved by TSSpredator. (left) TSSpredator predicts transcription start sites (TSS) by comparing two expression profiles: a standard RNA-seq (gray line) and an 5'-enriched profile (yellow line). Based on these, it computes three parameters across the genome: step height, step factor and the enrichment factor. After prediction, TSS are classified with respect to nearby genes. Primary and secondary TSS are located (within 300 bp) upstream of annotated genes, where secondary TSS show a lower enriched expression signal compared to the respective primary TSS. Internal TSS are located within the genes themselves, while antisense TSS are located on the antisense strand close to a gene (within 150 bp) or within the gene itself. Lastly, all other identified TSS are called orphan TSS. These figures have been adapted from the original publication and can be found in Figs. II.1 and II.2.

the demand for fast insight generation. This highlights the need for more user-friendly tools that furthermore enable the integration of additional data to characterize the prokaryotic transcriptome. To address this, we have developed TSSpredator-Web, a web application that not only performs TSS prediction via TSSpredator but also facilitates interactive exploration of the results. By providing an interface for uploading large datasets and by combining interactive genome-integrated exploration with automated TSS prediction, TSSpredator-Web bridges the gap between biological interpretation and data creation, primarily for domain experts.

The main add-on of TSSpredator-Web is the implementation of an exploration interface with multiple features (Fig. 5.6). The tabular file called *MasterTable*, which details all projected TSS, is the primary output of TSSpredator and is therefore essential for the explorative analysis. For this reason, we include an interactive view of the *MasterTable*, supporting common exploration features (Fig. 5.6, center). Moreover, to improve the exploration process, we included additional visualizations that mirror the exploration and visualization techniques used in multiple genome-wide TSS studies (Čuklina *et al.* 2016; Hocq *et al.* 2022; Ryan, Jenniches, *et al.* 2020). These features can be put into the context of the information-seeking mantra introduced by Shneiderman (1996) “Overview first, zoom and filter, details on demand”. To provide a broad overview of the results, an UpSet plot (Lex *et al.* 2014) summarizes the distribution of TSS (Fig. 5.6, left).

While the UpSet plot provides an overview on experiments and classes, TSSpredator-Web provides a genomic overview of all TSS positions (Fig. 5.6, right) via a genome explorer implemented with the toolkit Gosling (LYi *et al.* 2022). Based also on Shneiderman’s mantra, the genome explorer shows for each experiment all predicted TSS aggregated by their genomic position and class to show a genome-wide distribution, as well as including regions with annotated genes. Via semantic zooming, more information is added to the visualization, as TSS positions are no longer aggregated, but shown individually using glyphs. To provide more context on the genomic layer, the

5. Results

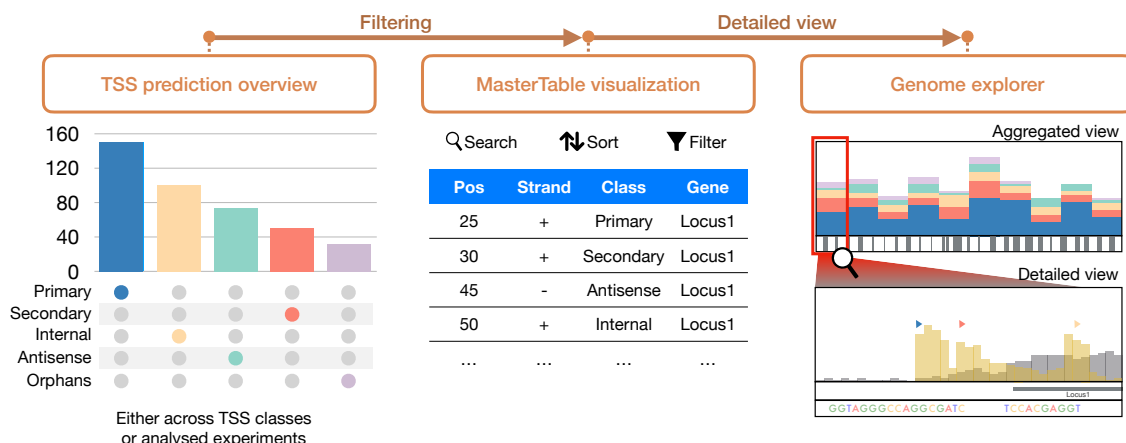


Figure 5.6: Core components of the interactive result exploration enabled by TSSpredator-Web. The UpSet plot (left) provides an overview of TSS distribution across categories (e.g., primary, secondary, internal, antisense, orphan), allowing users to identify overlaps and combinations across classes, but also across experiments. To explore the MasterTable (center), TSSpredator-Web provides interactive search, sort, and filtering functionalities to explore all predicted TSS and their attributes. The genome viewer (right) integrates transcriptomic and genomic data, providing both an aggregated view (showing TSS counts binned by a genomic region) and a detailed view (showing individual TSS, expression profiles, and upstream sequences). Both views also integrate the genome annotations. Arrows between the components indicate how the visualizations are linked via interaction, facilitating the exploration workflow in the tool. Some parts of this figure have been adapted from the original figure found in Fig. II.4

genomic sequence upstream of the TSS is shown in the genome explorer. This allows users to search for putative binding sites of known regulators (e.g., the Pribnow box in prokaryotes (Sharma *et al.* 2010)). Additionally, the detailed view integrates the expression data for each experiment for both enriched and control datasets, allowing the evaluation for the expression at each predicted TSS position. With these three visualizations, users can get an overview (e.g. with the UpSet plot and the genome viewer), zoom and filter, and get a detailed view (e.g. on the MasterTable or the detailed view of the genome viewer), mirroring all three steps of Sheneiderman’s mantra.

The improvement in TSS prediction and novel exploration in TSSpredator-Web was evaluated using a Cappable-seq dataset from the *ESKAPE* bacterium *Escherichia coli* (Balkin *et al.* 2025), comprising 48 coverage profiles (four conditions, three replicates, two strands, and two library types: enriched and non-enriched) that were easily uploaded to the new interface via *drag & drop*. The prediction returned a total of 7,194 TSS positions that were further explored using the new exploration platform. Via the UpSet plot we were able to identify condition-specific primary TSS linked to antibiotic stress-response genes, such as *ugd* and *arnB*. These genes were explored using the genome explorer to consider genomic and transcriptomic data, such as their location in the genome, their upstream sequences, and expression profiles. With the novel exploration platform, we could also explore orphan TSS in more detail. We showed that these TSS positions did not only have a clear TSS signal but also a clear promoter motif, such as the Pribnow box. Hence, rather than hinting at false positives, this analysis showed that these regions might contain overlooked transcriptional units. In general, this use case demonstrates how TSSpredator-Web not only facilitates the process of TSS prediction, but also the insight generation process of genome-wide TSS maps, such as the identification of regulatory mechanisms of condition-specific genes, or the discovery overlooked transcriptional units.

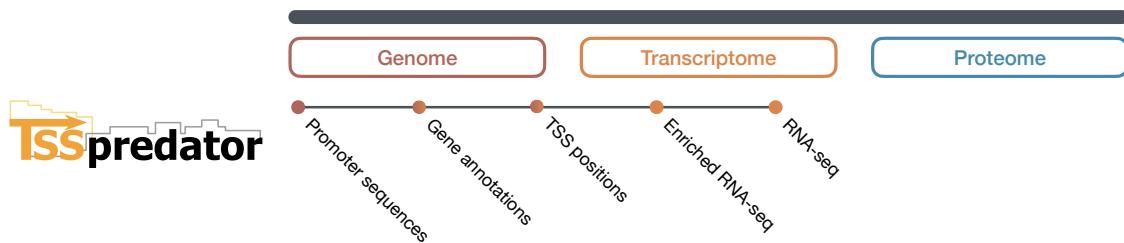


Figure 5.7: Framing the input and computed data of TSSpredator-Web within the central dogma of molecular biology. These are distributed (colored circles) onto the genomic and transcriptomic layer. Multicolored circles indicate data that connect different layers.

The implementation of TSSpredator-Web handles multiple data types originating from the genomic and transcriptomic layers of the central dogma of molecular biology, providing a bridge between them (Fig. 5.7). Starting from enriched and standard RNA-seq coverage profiles, the platform derives a new layer of information: genome-wide TSS maps. To gain even more knowledge on the data, TSSpredator integrates these maps with the genome sequence, their annotations but also the underlying expression data, in a single and interactive interface. This integration enables the easy generation of biological insight starting from expression data, addressing a key gap in existing TSS prediction workflows. TSSpredator-Web is available as a containerized tool and, similarly to Evidente, has been deployed to the TueVis visualization server in order for researchers to explore their own data². Moreover, this allows for future instances of TSSpredator-Web to be linked with other downstream analyses. One possibility would be to analyze the data with the TSS-Captur pipeline, which will be introduced in the following section.

5.3 Characterization of Overlooked Transcriptional Units via TSS-Captur

This section presents the article **Witte Paz, M., Vogel, T., and Nieselt, K. (2024).** “TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA Transcripts”. In: *NAR Genomics and Bioinformatics* vol. 6, no. 4, lqae168. DOI: [10.1093/nargab/lqae168](https://doi.org/10.1093/nargab/lqae168). For the full publication, refer to [Paper III](#).

The integration of genomic and transcriptomic data into genome-wide TSS maps has immensely helped researchers to characterize transcriptional elements within prokaryotic genomes, but has also led to the discovery of overlooked transcriptional units. More specifically, multiple studies analyzing TSS data, including the one introduced in the previous section, have shown a number of predicted TSS that cannot be associated with annotated genes (Ettwiller *et al.* 2016; Sharma *et al.* 2010; Wolfram-Schauerte *et al.* 2024). These include orphan TSS (oTSS), which are not located in the vicinity of any annotated gene, but also antisense TSS (aTSS) located on the antisense strand of an annotated gene, but with no gene located downstream. Due to the strong experimental signal, these sites provide an interesting start for identifying overlooked transcriptional units, such as small protein-coding genes or regulatory non-coding RNAs (ncRNAs), which are often overlooked by standard annotation methods (Barquist *et al.* 2016; Stazic and Voß 2016).

The diversity of these TSS signals across multiple studies motivated the development of TSS-Captur, a computational pipeline based on Nextflow (Di Tommaso *et al.* 2017). This pipeline addresses the automatic characterization of transcripts starting at oTSS

²Accessible via <https://tsspredator-tuevis.cs.uni-tuebingen.de/>

5. Results

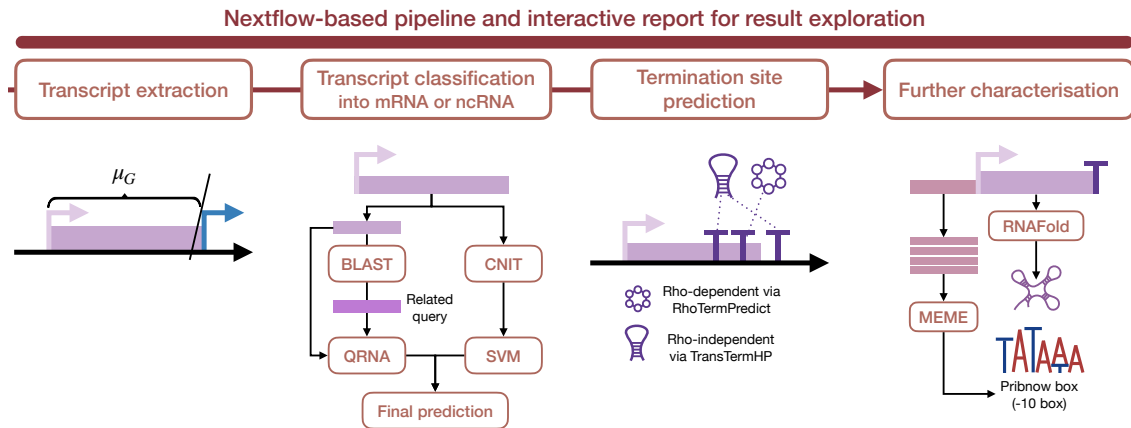


Figure 5.8: Overview of the characterization of genomic regions starting at TSS via **TSS-Captur**. The pipeline extracts the downstream regions of orphan and antisense TSS and classifies them as mRNA or ncRNA using **CNIT** and **QRNA**. Transcription termination sites (TTS) are identified genome-wide via **TransTermHP** (intrinsic) and **RhoTermPredict** (Rho-dependent). RNA secondary structure is predicted using **RNAFold**. Upstream promoter regions are analyzed with **MEME** to detect enriched motifs such as the Pribnow box. All results are integrated into an interactive report for exploration.

and aTSS by integrating genomic data with transcriptomic signals to define four different aspects: length, RNA class (messenger or non-coding), putative regulation mechanism, and secondary structure. To achieve this, **TSS-Captur** takes as input the *MasterTable* with the predictions of **TSSpredator** or **TSSpredator-Web**, as well as the genome sequence and its corresponding annotation file. Various methods are combined in the following manner (Fig. 5.8). First, **TSS-Captur** extracts a putative transcript for each TSS position, taking into account the mean gene length (μ_G) based on all annotated genes and other genomic features, such as TSS or annotated genes (Fig. 5.8 left-most). This transcript is then used as a base for the following modules to run the characterization.

To assign a functional class (mRNA or ncRNA) to each transcript, **TSS-Captur** integrates two complementary prediction strategies (Fig. 5.8, second from the left). On one hand, **CNIT** (Guo *et al.* 2019) is a support vector machine (SVM)-based method that classifies the extracted sequence as either non-coding RNA (ncRNA) or messenger RNA (mRNA) based on intrinsic sequence composition. On the other hand, **QRNA** (Rivas and Eddy 2001) employs a comparative genomics approach to detect evolutionarily conserved patterns indicative of coding or non-coding potential. This method requires a homologous sequence from a related organism, which is retrieved via **BLAST** (Altschul *et al.* 1990). The performance of both tools was evaluated based on known annotated transcripts to identify their strengths and limitations. This analysis showed that the accuracy of **QRNA** highly depends on an optimal homologous sequence and that **CNIT** overpredicts non-coding RNAs. Based on these observations, **TSS-Captur** combines both predictions based on a set of rules. If the predictions disagree, the class mRNA is chosen because of the previous limitations of each individual tool. This combined classification method showed a more robust classification accuracy than each tool individually.

Besides providing an RNA class for each transcript, **CNIT** and **QRNA** also provide the coordinates of the most probable functional subsequence. For example, in the case of an mRNA, the translated sequence. Based on these coordinates, **TSS-Captur** predicts the most-probable transcription termination site (TTS) to identify the coordinates of the full transcriptional unit (Fig. 5.8, second from the right). This is based on the

genome-wide prediction of TTS with the tools **TransTermHP** (Kingsford *et al.* [2007]) and **RhoTermPredict** (Di Salvo *et al.* [2019]). These two tools complement each other, as the former one focuses on the identification of intrinsic (or Rho-independent) terminators, while the latter one predicts those termination regions where the Rho protein is involved. Based on the positions of the TTS, their confidence score and the coordinates returned by **CNIT** and **QRNA**, **TSS-Captur** assigns a TTS for each transcriptional unit.

TSS-Captur characterizes two additional aspects for the analyzed transcriptional units to provide further functional and regulatory context (Fig. 5.8 right). **RNAFold** (Lorenz *et al.* [2011]) is applied to the full transcript unit (i.e., from TSS to predicted TTS) to compute a secondary structure based on the minimal free energy of the structure. Besides providing a hint on the stability of the RNA molecule, this step might help identify a putative function for those units classified as ncRNAs. The second characterization involves the analysis of the upstream regions of all oTSS and aTSS with **MEME** (Bailey *et al.* [2009]) to identify enriched sequence motifs. This allows researchers to identify any overrepresented sites that might correspond to a transcription factor binding site, and hence gain insight on how these overlooked transcripts are regulated.

To facilitate accessibility and interpretability of the results, an interactive interface is provided. Within the interface, users can not only access the final predictions for each transcript but also inspect intermediate results, allowing them to trace the processes that led to the final characterization. Moreover, it facilitates the visualization of the computed secondary structures as well as the integration of the results of the enrichment analysis returned by **MEME**. To further increase user accessibility, **TSS-Captur** is not only provided as a Nextflow pipeline, but has been extended to serve as a web application within the TueVis visualization server³. Similar to **TSSpredator-Web**, the deployment of **TSS-Captur** as a web application removes the installation hurdles and provides the user with a way to access the computational resources needed for complete characterization. This accessibility improvement lowers the barrier for domain experts for genome-wide TSS characterization.

TSS-Captur was applied to a genome-wide TSS dataset from *S. coelicolor* (Jeong *et al.* [2016]), analyzing 1,542 antisense and orphan TSS in a single high-throughput run of around 8 hours and 20 minutes. For most TSS, the pipeline predicted the transcript's functional class (925 as ncRNAs, 617 as mRNA), allocated a terminator site ($\approx 80\%$ Rho-dependent sites), and hence estimated its length (median length of 933 nt for mRNAs and 501 nt for ncRNAs). Moreover, it identified enriched motifs in the putative promoter region of the TSS. One highly enriched G/C-rich motif was detected in over 560 promoter regions, but had no match in any regulatory motif database, suggesting that a yet unannotated regulatory mechanism might be involved in the regulation of these overlooked transcripts. Lastly, to assess the accuracy of the characterization of the *S. coelicolor* dataset, the results were compared with an experimentally validated set of 63 uncharacterized transcripts reported by Vockenhuber *et al.* ([2011]). While the overlap between the two datasets was low (23 TSS), we identified that **TSS-Captur** correctly classified 18 as ncRNAs, with an overestimated length. The results of this use case show how **TSS-Captur** can characterize thousands of unannotated transcripts in a reproducible and accessible manner and hence complement existing annotation pipelines to expand the scope of prokaryotic characterization. By starting with TSS derived from experimental data and then providing predictions on transcript class, length, promoter motifs, and secondary structures, **TSS-Captur** provides an overview of hypothesis generation prior to any experimental validation. For example, researchers could analyze their genome-wide TSS maps to identify the most promising transcriptional units that might lead to putative novel regulatory ncRNAs. In such a case,

³<https://tsscatur-tuevis.cs.uni-tuebingen.de/>

5. Results

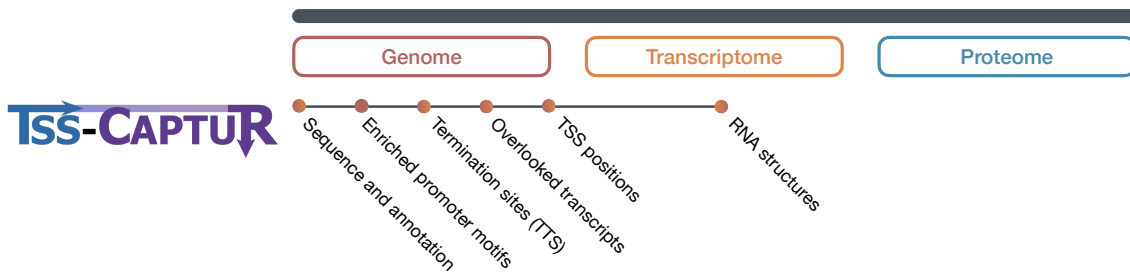


Figure 5.9: Framing input and outputs of TSS-Captur within the central dogma of molecular biology. Starting with experimentally predicted transcription start sites (TSS), the genome and the current annotation, TSS-Captur derives further information from the genomic and transcriptomic layer (colored circles). Multicolored circles indicate data that connect different layers.

the predicted data through TSS-Captur reduces both time and cost by already providing an overview of transcriptional units with high predicted functional potential.

As with the other methods presented in this thesis, TSS-Captur integrates and computes data from multiple layers of the central dogma (Fig. 5.9). While TSS are typically derived from experimental data, as in the case of TSSpredator-Web, TSS-Captur is not restricted to such inputs and would work with any set of genomic positions, as long as the data is provided in the correct format. Based on this set of genomic positions, the corresponding genome sequence and its annotation, TSS-Captur generates transcription termination sites as well as general annotations for overlooked transcriptional units, hence focused mainly on the computation of data for the genomic layer of the central dogma. A similar case occurs for the analysis of the enriched motifs in putative promoter regions. However, TSS-Captur also characterizes an aspect of the transcriptomic layer, as it allows researchers to take a first glimpse of the *structurome* of all predicted RNA transcripts.

Altogether, TSS-Captur not only bridges the gap between TSS data and functional annotation but also provides an integrative, reproducible, and accessible framework for the characterization of the prokaryotic transcriptome. While the presented examples do not directly contribute to the overall context of microbial control, one can speculate that characterizing the current *dark transcriptome* can have future relevance in this field. For example, novel regulatory mechanisms such as ncRNAs might help first to understand and later to control multiple aspects of prokaryotic pathogenesis (Chauvier and Walter 2024; Gottesman 2005). With this, TSS-Captur not only complements existing genome annotation pipelines but also provides a foundation for generating hypotheses that may reveal novel regulatory elements or genes, which in turn could open new avenues for experimental validation and lead to novel strategies for microbial control.

5.4 Comprehensive computational analysis of metallophore synthesis and uptake potential in *Staphylococci* species

This section presents the article **Witte Paz, M.**, Bitzer, A., Nieselt, K., and Heilbronner, S. (2025). “A Landscape of Metallophore Synthesis and Uptake Potential of the Genus *Staphylococcus*”. In: *NAR Genomics and Bioinformatics* vol. 7, no. 4, lqaf183. DOI: [10.1093/nargab/lqaf183](https://doi.org/10.1093/nargab/lqaf183). For the full publication, refer to [Paper IV](#).

While the previous projects were primarily driven by a data-centric motivation that aims to develop novel computational methods for prokaryotic characterization, the following two projects focus on the uptake of metallophores in bacteria. Specifically, in this project, we investigated the genomic potential of *Staphylococcus* species to synthesize and acquire these metabolites.

Members of the *Staphylococcus* genus are colonizers of human and animal microbiomes. Prominent examples include *S. epidermidis* found on human skin (Bomar *et al.* [2018](#)), *S. lugdunensis* commonly found on the skin and in the anterior nares (Heilbronner [2021](#)), and *S. aureus*, a member of the *ESKAPE* pathogens and therefore a significant burden for healthcare systems (Howden *et al.* [2023](#)). Although they are found in multiple host environments, successful colonization is challenging, as essential nutrients, such as metals, are limited (Weinberg [2000](#); Zackular *et al.* [2015](#)). To overcome this, bacteria have evolved metallophores: secondary metabolites with a high affinity for metal ions (Neumann *et al.* [2017](#)). A well-studied subclass of metallophores are siderophores, metabolites specialized in iron retrieval (Kraemer *et al.* [2015](#)). The ability to synthesize such metabolites is gained via biosynthetic gene clusters (BGCs), sets of genes that translate the enzymatic machinery required for the synthesis, modification and export of metallophores (Reitz and Medema [2022](#)). After secretion and metal chelation, for Gram-positive bacteria, metallophore-metal complexes are re-imported into the cell through lipoprotein receptors and ATP-binding cassette (ABC) transporters (Kraemer *et al.* [2015](#)). However, since metallophores function outside the cell, the chelated complex is available not only to the organism that synthesizes it but also to neighboring microbes that express compatible receptors (Kramer *et al.* [2020](#)), creating an ecological interaction between bacteria (Cordero *et al.* [2012](#)).

One well-studied organism with respect to the synthesis and uptake of metallophores is *S. aureus*. This species synthesizes a repertoire of metal-chelating compounds that includes two siderophores and one metallophore (Ghssein and Ezzeddine [2022](#)). *S. aureus* synthesizes the carboxylate-type siderophores staphyloferrin A (SF-A) and staphyloferrin B (SF-B) via two distinct BGCs (Beasley, Vinés, *et al.* [2009](#); Dale, Doherty-Kirby, *et al.* [2004](#)). Chelated SF-A and SF-B molecules are re-imported into the cell via dedicated and highly specific uptake systems: HtsABC and SirABC, respectively. In both cases, substrate specificity is mediated by an extracellular lipoprotein (HtsA or SirA), which recognizes the chelated complex and delivers it to associated ABC transporters (HtsBC or SirBC) for transport across the membrane (Cotton *et al.* [2009](#); Dale, Sebulska, *et al.* [2004](#)). Beyond iron acquisition, *S. aureus* also produces the metallophore staphylopin (STP), which enables the uptake of manganese and zinc (Ghssein, Brutescio, *et al.* [2016](#)). For the uptake of the metal-loaded metallophore, the CntABCDF system is required, with CntA being the lipoprotein receptor. Besides synthesizing its own metallophores, *S. aureus* is capable of exploiting metallophores produced by other species, a phenomenon known as xenosiderophore usage. *S. aureus* has been shown to uptake hydroxamate-type siderophores through the FhuBCD1D2G transporters and catecholate-type siderophores through the SstABCD transporter (Beasley, Marolda, *et al.* [2011](#); Sebulska, Hohnstein,

5. Results

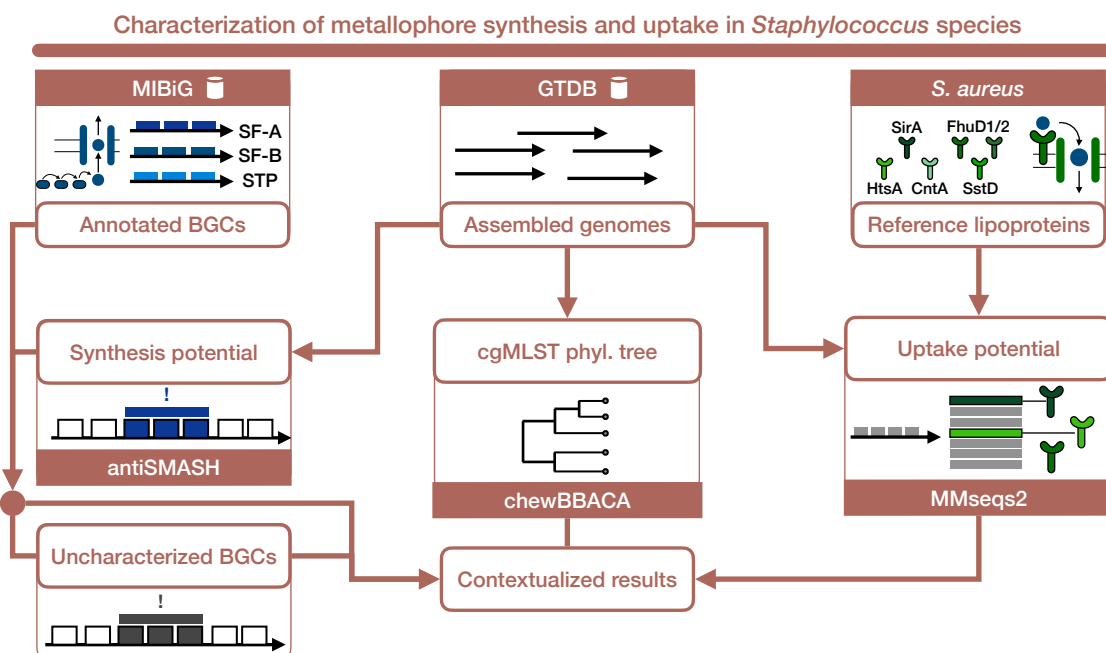


Figure 5.10: Overview of data processing for the characterization of metallophore synthesis and uptake potential of staphylococcal species.

et al. [2000]; Sebulsky, Speziali, *et al.* [2004]), where FhuD1, FhuD2, and SstD are lipoproteins for recognition of metallophores (Beasley, Marolda, *et al.* [2011]).

Although the synthesis and uptake potential of the metallophores of *S. aureus* has been studied in detail, the extent to which these potentials are conserved or diversified across the *Staphylococcus* genus remains poorly understood. To address this gap, this study performed a large-scale computational analysis of *Staphylococcus* genomes to systematically characterize the diversity of metallophore synthesis and uptake potential. Based on genomic sequences of more than 1,900 *Staphylococcus* strains available through the Genome Taxonomy Database (GTDB) (Parks *et al.* [2022]), we assessed the presence, absence, and variability of metallophore-related biosynthetic and transport systems, providing a comprehensive overview of the metallophore landscape within the genus.

The analysis was performed at both the species and strain levels to identify genus-wide trends as well as intra-species variability. At the species level, one representative genome per species, as provided by the GTDB, was analyzed, resulting in a total of 74 *Staphylococcus* species. As not all representative genomes correspond to complete assemblies, assembly metadata was retrieved and considered to assess potential assembly-related biases in downstream analyses.

For the strain-level analysis, four species associated with human microbiome were investigated in detail: *S. epidermidis* (skin), *S. capitis* (scalp), *S. hominis* (skin), and *S. lugdunensis* (skin). Although they are described as commensal species, they have also been reported to cause serious infections under some circumstances (Becker *et al.* [2014]). All available strains for these species were downloaded from the GTDB and underwent a quality control screening to reduce the impact of highly fragmented genome assemblies, as they could compromise the reliable identification of BGCs and proteins involved in metallophore uptake.

Both levels of analysis followed the same computational workflow (Fig. 5.10). To assess the genomic potential for the synthesis of metallophores, each genome sequence

was analyzed using **antiSMASH** (Blin *et al.* 2023) to identify BGCs, but considering only those associated with metallophore synthesis. As **antiSMASH** also compares predicted BGCs with curated clusters from the **MIBiG** database, it allowed the identification of SF-A, SF-B, and STP biosynthesis genes as well as the evaluation of their completeness. Hence, for each species or strain, the BGCs were classified as complete, incomplete, or absent. If **antiSMASH** identified any BGCs as putative metallophore clusters but without a counterpart in the **MIBiG**, they were analyzed against known metallophore BGCs using the tool **clinker** (Gilchrist and Chooi 2021), allowing evaluation of structural conservation and potential functional relatedness. For the analysis of the strain level of *S. lugdunensis*, a previously described BGC with potential for metallophore synthesis (Heilbronner *et al.* 2011) was searched in all genomes based on **BiG-SCAPE** (Navarro-Muñoz *et al.* 2020).

As **antiSMASH** only allows the characterization of the synthesis potential, the metallophore uptake potential was analyzed using a sequence homology-based approach using **MMseqs2** (Hauser *et al.* 2016). As lipoproteins are essential for uptake of metallophores, strains capable of importing a metallophore are expected to encode homologs of the corresponding uptake lipoproteins. Therefore, the amino acid sequences of the six lipoproteins known to be involved in *Staphylococcus aureus*' metallophore uptake were used as reference to identify homologs. If more than one gene within a strain exceeded the predefined similarity threshold for a given reference, the strain was considered to harbor multiple homologs of the corresponding uptake lipoprotein.

The results of the metallophore synthesis and uptake analyses were contextualized with a phylogenetic tree (Fig. 5.11) based on the core-genome multilocus sequence typing (cgMLST) approach of the tool **chewBBACA** (Silva *et al.* 2018). Among the 74 staphylococcal species, the landscape of metallophore biosynthesis varies (Fig. 5.11, leftmost heatmap column). All 74 species harbored the SF-A biosynthetic locus, 10 species exhibited truncations due to missing essential genes, suggesting an inability to synthesize SF-A. In contrast, staphyloferrin B (SF-B) biosynthesis was detected in only 17 species distributed across multiple phylogenetic lineages. Similarly, staphylopine (STP) biosynthesis was identified in 29 diverse staphylococcal species, with truncation observed in a single case.

Beyond these well-characterized metallophores, **antiSMASH** identified an additional BGC predicted to encode a metallophore that is not represented in the **MIBiG** database (Fig. 5.11, gray column) but showed a high similarity to a siderophore biosynthetic cluster described in *Mammaliococcus sciuri* when compared via **clinker**. Interestingly, this putative metallophore BGC was restricted to a specific phylogenetic clade that lacks the potential for synthesis of SF-A, suggesting a potential functional replacement of SF-A in these species.

The observed diversity goes beyond metallophore synthesis and is also reflected in the uptake systems (Fig. 5.11, green columns). In most species, the presence of the *htsA* gene closely mirrors the capacity for SF-A production, such that species with truncated SF-A biosynthetic loci generally also lack an *htsA* homolog. However, exceptions to this pattern were identified. *S. lugdunensis* has retained *htsA* despite carrying a non-functional SF-A biosynthetic cluster, suggesting the ability to exploit SF-A via xenosiderophore uptake, aligning with previous experimental observations (Brozyna *et al.* 2014). A similar pattern was also observed for SF-B and STP. In most cases, the truncation of a BGC coincided with the absence of the respective uptake lipoproteins (SirA for SF-B and CntA for STP), with a few species deviating from this pattern, such as *S. lugdunensis*.

The diversity among lipoproteins associated with the usage of xenosiderophores (FhuD1, FhuD2, and SstD) was also analyzed. In addition to being present in multiple *Staphylococcus* species, several strains encoded multiple copies of these lipoproteins (Fig. 5.11, red columns),

5. Results

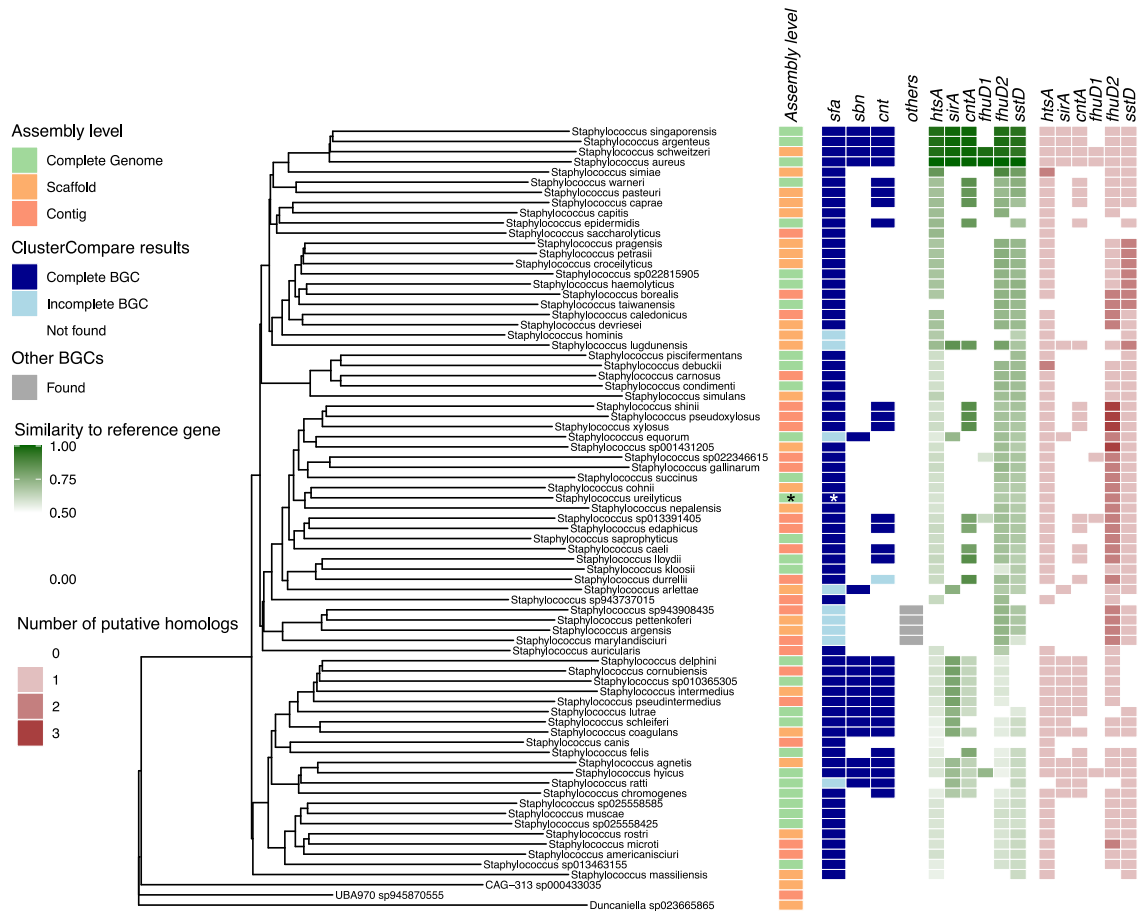


Figure 5.11: Overview of metallophore synthesis and uptake potential in the genus *Staphylococcus*. The dendrogram shows the phylogenetic relationships of staphylococcal species, alongside the results summarizing their metallophore-related biosynthetic and uptake capacities. The assembly level of each used reference is visualized in the first heatmap. The synthesis potential was assessed with *antiSMASH*. Results annotated as the three known *S. aureus* metallophores obtained with *ClusterCompare* are shown in the second to the left heatmap column. Blue rectangles indicate biosynthetic gene clusters (BGCs), with dark blue representing complete BGCs (all synthesis genes present) and light blue indicating incomplete BGCs. Gray rectangles mark additional BGCs predicted as potential metallophore synthesizers but without specific annotation. The last two columns represent the uptake potential: green shading indicates the sequence similarity of the proteome to *S. aureus* metallophore uptake lipoproteins, while red shading shows the number of putative homologs identified. Strains that were reassessed and that have shown different results have been marked with an asterisk (*). This figure and caption have been adapted from the original publication. The original version can be found in Fig. [IV.2](#)

suggesting a potential broadening of uptake capacity for hydroxamate- and catecholate-type siderophores.

Due to the high metabolic cost of metallophore synthesis, selective pressure is expected to develop mechanisms that restrict the exploitation of these metabolites to producing organisms. Previous studies have shown that this pressure drives diversification of both metallophores and their cognate lipoprotein receptors, restricting access by non-producers (Gu *et al.* [2025](#)). To investigate the presence of these mechanisms within *Staphylococcus* species, we analyzed the conservation of the known binding-site residues of the three lipoproteins involved in metallophore uptake (HtsA, SirA, and CntA). For each lipoprotein, at least two distinct binding-site conservation patterns were identified. These

5.4 Comprehensive computational analysis of metallophore synthesis and uptake potential in *Staphylococci* species

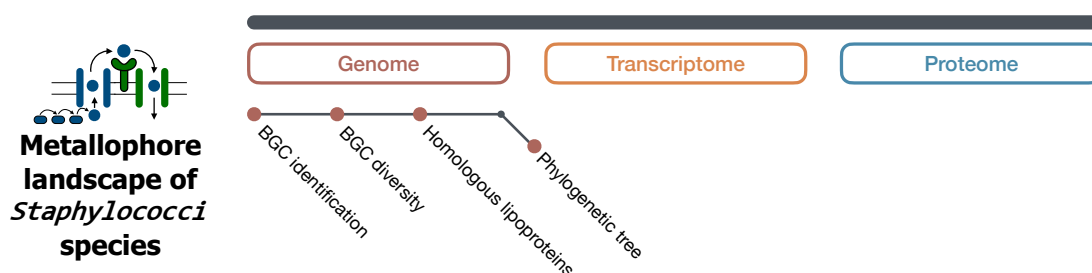


Figure 5.12: Framing the data integrated in the characterization of the metallophore landscape of staphylococcal species using the central dogma of molecular biology as a framework. The data types (colored circles) are distributed onto the genomic, as they all originate from this data layer.

patterns could reflect evolutionary strategies to restrict siderophore usage, that should be experimentally confirmed in the future.

Lastly, strain-level analysis revealed species-specific diversity patterns. *S. hominis* and *S. epidermidis* strains showed a high degree of conservation in both metallophore synthesis and uptake potential, since most strains closely mirror the observed traits in the respective representative genomes. In contrast, only 16% of the *S. capitis* strains reflected the potential inferred from the representative genome, indicating that the representative strain does not fully represent the metallophore-related traits of this species. *S. lugdunensis* also displayed intra-species diversity, as 48% of the strains shared the same trait profile as the reference, while a second distinct pattern was observed in 40% of the strains. Notably, despite this diversity in *S. lugdunensis*, the yet uncharacterized BGC with metallophore synthesis potential was consistently identified in all *S. lugdunensis* strains.

Altogether, the results of this project establish the first systematic, genus-wide characterization of metallophore synthesis and uptake potential in staphylococcal species. This work continues the overall strategy the preceding projects by integrating diverse types of genomic data to extend biological understanding of a biological question. Moreover, it provides the foundation for the development of a new application that addresses this aspect in prokaryotic characterization, as multiple bioinformatic tools were applied to answer the research question. Here, similarly to **Evidente**, the approach focused on integrating multiple aspects of genomic data (Fig. 5.12). By doing so, we demonstrate how information embedded in the genome can be leveraged to derive insights into functional layers, such as metabolic capabilities, that are not apparent from the primary sequence alone. This highlights once more how integrative bioinformatic analysis can bridge the different layers of the central dogma, advancing our understanding of bacterial metabolism. This integrative approach was also used in the following project to investigate the role of metallophores, more specifically siderophores, in bacterial interactions between *S. aureus* and multiple species from the genus *Corynebacteria*.

5.5 Characterization of Siderophore Uptake Mechanisms and their Regulation in *Corynebacteria*

This section presents the ongoing manuscript **Witte Paz, M.**, Bitzer, A., Heilbronner, S. and Nieselt, K. “Characterizing Siderophore Uptake and their Regulation Mechanisms in *Corynebacteria* based on Structural Homology and Transcriptome Profiling”. For the full manuscript, refer to **Paper V**

Given that *Staphylococcus aureus* poses a substantial threat to healthcare systems worldwide due to its high mortality rates and broad spectrum of antibiotic resistance (Howden *et al.* [2023]), understanding how metallophore-mediated interactions with other organisms can be used for antimicrobial strategies has gained importance. A recent study by Zhao *et al.* [2024] demonstrated that several *Corynebacteria* species are able to inhibit the growth of *S. aureus* by exploiting its siderophores, highlighting the piracy of xenosiderophores as a potential mechanism of microbial interference within the nasal environment. While the molecular basis of this interaction was elucidated only for one species (*C. hesseae*), other *Corynebacterium* species exhibiting a similar inhibitory phenotype remained largely uncharacterized. Hence, the main goal of this project was to systematically characterize the mechanisms by which multiple *Corynebacteria* exploit xenosiderophores. This integrative analyses integrated three complementary perspectives: the identification of uptake systems via structural homology, their transcriptomic response under iron-limited conditions, and the analysis of the underlying regulatory mechanisms.

Based on sequenced genomic data, the first goal was to identify the mechanisms that allow *Corynebacteria* to profit from these siderophores. For this, our dataset consisted of 52 *Corynebacterium* sequenced genomes, which strains covered six different species (an overview can be found in Figure V.1). Here, it is important to mention that the dataset also included the *C. hesseae* species analyzed by Zhao *et al.* [2024]. All strains have been experimentally assessed to identify their capability to synthesize siderophores via chrome azurol S (CAS), as well as their utilization of siderophores via co-cultures with *S. aureus* and derivated mutants. Moreover, the dataset included eight *S. aureus* sequenced strains for reference and validation of the computational approaches.

One of the mechanisms identified by the authors was the presence of a distant homolog of *S. aureus*’ HtsA to make siderophore uptake possible. Hence, similarly as in the previous project, we hypothesize that the uptake mechanisms in *Corynebacteria* should involve homologous genes to any lipoprotein involved in siderophore uptake of *S. aureus* (HtsA, SirA, FhuD1, FhuD2 or SstD). First approaches with sequence homology approaches did not align with the results of the co-culture assays, suggesting a low sensitivity due to the evolutionary distance between the genera. For this reason, the pipeline PRESERVE was applied for a structural homology analysis.

PRESERVE is a computational framework commonly developed with Simon Hack⁴ that clusters conserved proteins based on sequence homology and predicts their structure based on a representative sequence. For a reproducible identification of homologous lipoproteins across species based on structural similarity, PRESERVE has been developed as a Nextflow pipeline (Fig. 5.13). The pipeline consists in identifying lipoproteins from proteome sequences using SignalP6 (Teufel *et al.* [2022]), followed by a clustering of genes based on MMseqs2 (Hauser *et al.* [2016]) to reduce the redundancy of the set. Based on the reduced set, the tertiary structure of the cluster representative is predicted via ColabFold (Mirdita *et al.* [2022]) to represent all lipoproteins in the cluster. Going beyond structural prediction,

⁴Research Group Integrative Transcriptomics, University of Tübingen

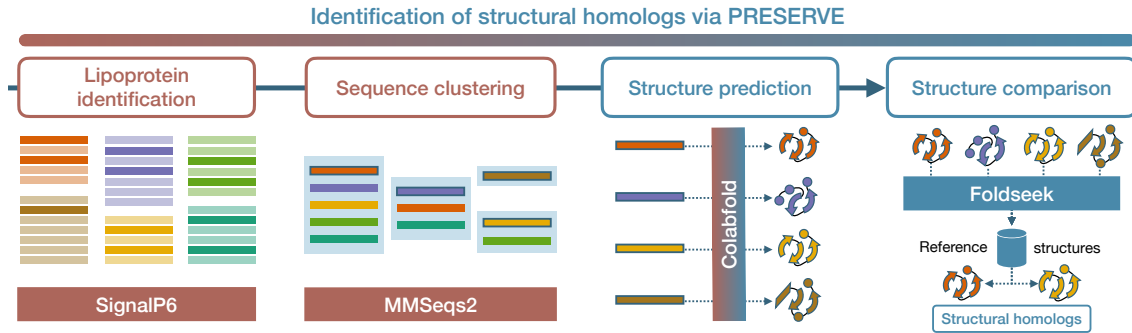


Figure 5.13: Overview of the pipeline **PRESERVE** for identification of structural homologs based on genomic (red rectangles) and proteomic data (blue rectangles). Starting with a set of multiple genomes, the pipeline reduces the proteome to get only lipoproteins using **SignalP6** (Teufel *et al.* 2022). By clustering the amino acid sequences with **MMSeqs** (Hauser *et al.* 2016), one representative per cluster can be selected (highlighted in blue) for structural prediction using **Colabfold** (Mirdita *et al.* 2022). Lastly, the structures are compared to the reference set via **Foldseek** (van Kempen *et al.* 2024) to identify those with structural homology.

PRESERVE performs a structural comparison of predicted lipoprotein structures through **Foldseek** (van Kempen *et al.* 2024) to a set of references, such as the five lipoproteins of *S. aureus* for this study. In the end, the clusters of identified homologs are reinflated to obtain a set of homologs for each provided strain.

We applied **PRESERVE** to the dataset collected for this study, which includes 8 strains of *S. aureus* that were used as a validation. By defining a minimal coverage of 0.7 and a minimal LDDT value of 0.5 for structural alignments to identify a structural homolog, we identified a total of 444 lipoproteins in all 60 samples, with 381 found in the different *Corynebacteria* species (Fig. 5.14). The number of potential homologs was species-specific, but showed a stable profile across strains of the same species. Here, we could note that for the *C. hesseae* strain, our approach was able to capture the homolog identified in the study by Zhao *et al.* (2024). Moreover, to validate our approach, we analyzed the results of *S. aureus* strains in more detail. The results showed that **PRESERVE** identified all five references as structural homologs. Nevertheless, other lipoproteins that are not known to be involved in SID uptake were also identified. This suggested that while our objective was to improve the sensitivity of the SID-uptake mechanisms, the specificity of our approach dropped, and the set of structural homologs might include false positives. Nevertheless, since there are no references for corynebacterial species, we kept all 381 as potential hits for further characterization of the species.

A further goal of this project was to characterize the transcriptomic response of the corynebacterial species in an iron-deprived environment. To achieve this, RNA-seq experiments were conducted for four of the analyzed species (Fig. 5.15, upper part). First, the overall results of the RNA-seq experiments were evaluated using a standard RNA-seq analysis pipeline (Harbig 2023) and **DESeq2** (Love *et al.* 2014). This returned over 980 differentially expressed genes for each species, even though a conservative significance threshold ($p - \text{adj.} < 0.01$) was selected. Moreover, a GO-enrichment analysis with the tool **GO-Compass** (Harbig *et al.* 2023) returned multiple GO terms involved in siderophore uptake, with many of them linked to up-regulated genes. This confirmed the overall usage of xenosiderophores by these species.

This general transcriptomic characterization of the corynebacterial species was used to contextualize the identified structural homologs (Fig. 5.16, red triangles). 86.5% genes (32 out of 37) identified with **PRESERVE** were strongly and significantly up-regulated,

5. Results

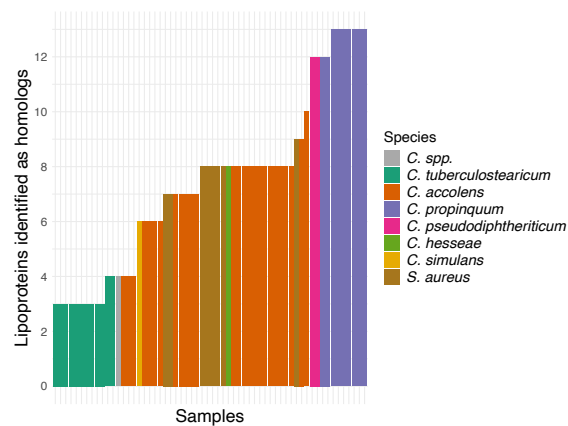


Figure 5.14: Number of potential structural homologs identified by **PRESERVE** per sample, colored by species.

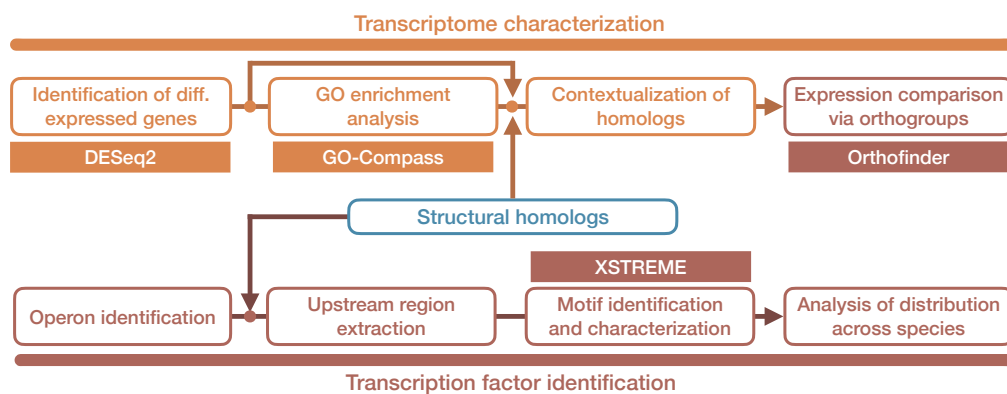


Figure 5.15: Overview of further characterization of *Corynebacteria* samples based on the identified structural homologs. The upper approach describes how the transcriptomic data was analyzed and integrated in the analysis. The first steps describe the general analysis of the data, while the latter approaches describe the integration of the structural homologs in the analysis. The bottom pipeline describes how the promoter regions of the identified homologs were analyzed in order to identify putative regulatory mechanisms.

indicating a clear need for these genes in the exposed environment. The five remaining genes showed a low log fold-change and also a p -value above the significance threshold. However, the annotation files described these genes as related to iron transport, suggesting they might not be false positives, but rather inactive or constantly expressed across both conditions. Moreover, the previous GO-enrichment analysis provided an overview of the annotation of the identified homologs. In total, the GO annotation identified 16 lipoproteins related to siderophore transport. Among these 16 genes, only nine overlapped with the 37 structural homologs of our previous analysis, resulting in 28 lipoproteins uniquely identified by **PRESERVE**, and seven unique for the GO enrichment. When accounting for the expression of the genes, only one gene exclusively found by the GO-enrichment showed an up-regulation in the RNA-seq experiment, whereas most of the candidates identified via **PRESERVE** were strongly up-regulated, as previously mentioned. Hence, this highlights that the genes identified by **PRESERVE** might have a stronger linkage with iron-deprived environments than the ones annotated by the GO-enrichment analysis. This also shows the necessity of other methods for gene annotation, as GO annotation tools seem to overlook potential genes related to this molecular function.

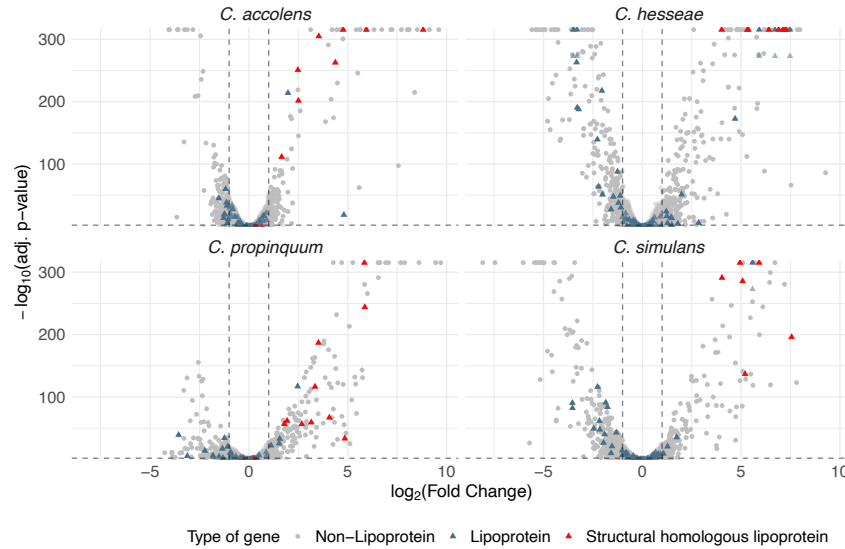


Figure 5.16: Volcano plots summarizing the complete transcriptomic response for a control vs. treatment with the chelator EDDHA (Fold change: treated – control). The vertical lines highlight where the absolute $\log_2$ fold-change ($|\log_2(FC)|$) is equal to one. The volcano plot differentiates between lipoproteins identified only by **SignalP6** (blue triangles), lipoproteins identified as structural homologs via **Foldseek** (red triangles) and other genes not identified as lipoproteins (gray circles). The original figure can be found in Fig. [V.7](#).

Besides analyzing the expression of the structural homologs, all up-regulated lipoproteins (Fig. [5.16](#), blue triangles with $\log$ fold change > 1) were assessed in detail. Here, the results of the structural alignments with these genes were revised to assess whether **PRESERVE** had missed any relevant lipoproteins. Only four proteins were identified as putative false positives, as they returned a high coverage of the references ($> 94\%$), but LDDT values slightly below the defined thresholds (≈ 0.4). A more detailed inspection of the structural alignment showed that the low LDDT value was caused due to changes in unstructured regions rather than in the core of the protein. This indicates that the predicted structures should be preprocessed, for example, using **AlphaCutter** (Tam and Iwasaki [2023](#)), to improve the structural comparison to clearly defined regions. The rest of the up-regulated lipoproteins was identified to be other types of lipoproteins also needed for the siderophore uptake, such as ABC transporter proteins. However, since they were not part of the used reference set, they were not identified with this approach.

Lastly, besides contextualizing the homologs for each species separately, we compared the identified homologs to each other via orthologous groups computed **Orthofinder** (Emms and Kelly [2019](#)) to analyze the consistency of the gene expression in a cross-species manner. Based on the orthogroups, we could identify that the majority of homologs within the same orthogroup not only shared the up-regulation, but showed similar absolute expression values. This indicates that, independently of each other, the species show a similar expression profile of the homologs.

The identified orthogroups also enabled a comparison between gene sequence similarity and the phenotypic outcomes of the co-culture assays with *S. aureus*, which tested the ability of different species to utilize staphyloferrin A and/or B. However, none of the orthogroups mirrored the phenotypic profiles observed experimentally. As previously mentioned, Zhao *et al.* ([2024](#)) identified the molecular mechanisms of siderophore uptake in *C. hesseae*, reporting two candidate genes: one mediating the uptake of both siderophores and another

5. Results

showing no functional involvement in this process. In our dataset, both genes were assigned to the same orthogroup, despite showing different phenotypes in the experiments. This discrepancy shows a limitation of our orthogroup-based comparison, which may not account for the fine-grained genetic aspects needed for siderophore specificity.

The final objective of this project was to investigate the regulatory mechanisms of the identified homologs by analyzing their putative promoter regions (Fig. 5.15, bottom part). Motif enrichment analysis can reveal overrepresented sequences that likely correspond to transcription factor binding sites and thereby provide insight into the regulation of gene expression (Parise *et al.* 2020). Since siderophore uptake genes are organized in operons (Cotton *et al.* 2009; Dale, Sebalsky, *et al.* 2004; Ghssein, Brutesco, *et al.* 2016), they are expected to be co-regulated by the same transcription factor (Osbourn and Field 2009). To account for this, we predicted operons across all corynebacterial genomes using a conservative threshold of 50bp intergenic distance, and included the *S. aureus* strains in the analysis for validation purposes. For each operon, the upstream region of the first gene was extracted and analyzed using XSTREME to identify and annotate enriched motifs (Grant and Bailey 2021). Filtering the resulting motifs to only those known to be associated with iron homeostasis from the literature revealed binding sites corresponding to two transcription factors: the Ferric uptake regulator (*Fur*) (Sheldon and Heinrichs 2012) and the Iron-dependent regulator (*IdeR*) (Lee and Holmes 2000; Marcos-Torres *et al.* 2021).

To compare the occurrence of binding sites across species, we counted the number of binding sites upstream of the identified genes per strain and summarized the results at the species level (Fig. 5.17). In *S. aureus*, *Fur* binding sites were predominant, consistent with their well-established role as the primary regulator of iron-uptake mechanisms in this species (Sheldon and Heinrichs 2012). In contrast, corynebacterial species showed a higher frequency of *IdeR* binding sites compared to *Fur*. *IdeR*, a member of the *DtxR* regulation family, has been described in the literature as the main regulator for iron uptake in *Corynebacterium diphtheria* (Marcos-Torres *et al.* 2021). Our analysis therefore enforces the idea that, in *Corynebacteria*, *IdeR* rather than *Fur* serves as the primary transcriptional regulator of iron uptake.

In this study, the main goal was to characterize the uptake mechanisms of siderophores and their regulation in a set of isolates of corynebacterial species. To achieve this, data from multiple layers across the central dogma was analyzed using computational approaches and the results were integrated for a detailed characterization (Fig. 5.18). Starting with genomic sequences and annotations, the focus was to identify putative lipoproteins involved in this process. Further processing through sequence-based clustering and prediction of protein structure led to the identification of structural homologs to the provided references, thereby bridging the layers of genomics and proteomics. To contextualize the function of these findings, the transcriptomic data was analyzed both generally for each strain and specifically for the identified homologs. In addition, cross-species comparisons between corynebacterial species allowed the assessment of expression consistency through orthogroups. Lastly, promoter analyzes revealed candidate regulatory binding sites and thus provided insight into species-specific mechanisms of iron homeostasis. Together, these results demonstrate that characterizing complex uptake systems requires a multi-layered approach, where insights emerge only from the integration of genomic, transcriptomic, and proteomic perspectives.

In summary, integration of structural homology detection, orthogroup analysis, transcriptomic profiling, and motif enrichment allowed the characterization of putative siderophore uptake mechanisms in nasal *Corynebacteria*, using information across all layers of the central dogma of molecular biology. The results showed a large repertoire of structurally conserved lipoproteins that are strongly up-regulated under iron limitation.

5.5 Characterization of Siderophore Uptake Mechanisms and their Regulation in *Corynebacteria*

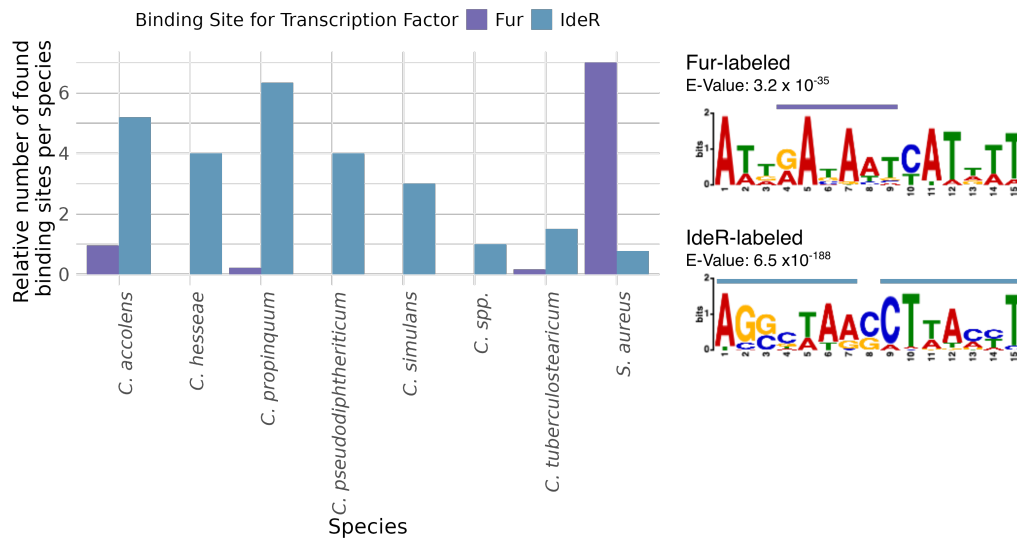


Figure 5.17: Overview of the results for motif enrichment across the analyzed species. (left) Relative abundance of transcription factor binding sites associated with iron homeostasis, identified per species. The number of identified motifs was normalized by the number of samples per species, providing an estimate of each transcription factor's overall relevance within the respective species. (right) Motifs identified as binding sites for *Fur* or *IdeR*, known iron-regulators. The corresponding consensus binding sequences have been highlighted for each motif (bar on top of the subsequence). The original figure can be found in Fig. [V.10](#).

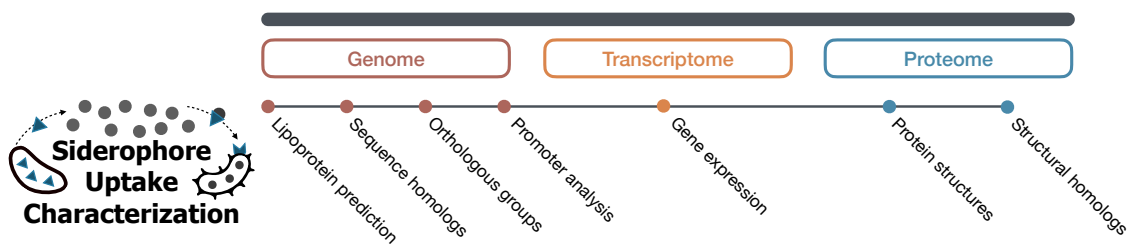


Figure 5.18: Framing of the data integrated for the analysis of siderophore uptake in corynebacterial species. This study analyzed and computed data across all three layers of the central dogma of molecular biology to provide insight into the investigated mechanisms.

However, their precise functional functions could not be resolved, as illustrated by the discrepancies between orthogroup assignments and phenotypic uptake profiles. Moreover, the identification of *IdeR* rather than *Fur* as the main regulator across corynebacterial species highlights that even conserved functional traits can be embedded in divergent regulatory frameworks. Together, these findings provide a high-resolution but still incomplete picture of how corynebacteria exploit staphylococcal siderophores, highlighting candidate genes for future experimental validation and the methodological challenges of linking sequence and structural homology to substrate specificity.

Chapter 6

Discussion and Conclusion

The best means of controlling microbes is to understand their capabilities and, more importantly, the mechanisms behind them. In order to address this, the optimization of omic data generation has led to a rapid increase in available data that provides insights into multiple types of microbes. However, the generation of insights from such data is highly dependent on its processing. In this thesis, new applications have been developed to advance the characterization of prokaryotes. The underlying theme that connects all projects is the idea that, instead of analyzing single data sources independently, methods that integrate multiple sources into a workflow should be developed to deepen insights within prokaryotic characterization across the central dogma of molecular biology. Furthermore, these projects demonstrate how methodological advances, such as advances in accessibility, data exploration, and reproducibility, directly enhance the quality of insights in microbial research. Here, we will discuss how the projects complement each other, and uniquely address certain aspects to fulfill these mentioned goals, as well as the main goal of the thesis: to enhance the characterization of prokaryotes for microbe control. The chapter concludes with a critical assessment of the current challenges for the developed tools as well as with an overview of emerging trends in the field.

Each of the projects presented integrates more than one type of input data for the characterization of prokaryotes. However, for some of the projects, their input data focus on one layer of the central dogma of molecular biology, while others have a cross-layer scope.

The project presenting the characterization of siderophores uptake mechanisms in *Corynebacteria* species (Project [V](#)) has the largest scope across layers of the central dogma, as it required the processing and analysis of genomic and transcriptomic data. Following common approaches to integrate multiple omics layers, each data input was analyzed independently to be then integrated at a later point for insight generation. First, the Nextflow-based pipeline **PRESERVE** provided insight into the protein structurome and allowed the homolog identification of multiple lipoproteins involved in iron-uptake based on genomic data, while late integration of transcriptomic data provided a further important perspective on the identified structural homologs. The transcriptome data revealed that most of the identified homologs are highly expressed in iron-deprived environments and also helped to discard homologs that might not be involved in these iron-uptake mechanisms.

The other projects presented in this thesis employ a complementary integration strategy, leveraging multiple data sources within a single layer to gain insights into that same layer or characterize downstream layers of the central dogma. For the characterization of metallophore synthesis and uptake across staphylococcal species (Project [IV](#)), genomic data characterizing biosynthetic gene clusters and homologous lipoprotein receptors were analyzed together to elucidate the diversity of metabolite production and utilization across the genus. This demonstrated how the combination of multiple perspectives within the same layer can reveal complex phenotypical potentials. **Evidente** (Project [II](#)) follows a similar approach by unifying genomic features, in this case phylogenetic and SNP data, with functional annotations into a single, accessible framework for prokaryotic characterization. By transforming complex genomic and phylogenetic data into an accessible and interpretable format, **Evidente** lays the groundwork for a deeper investigation across other layers of the central dogma, such as the effect of mutations on specific biological processes. In addition,

6. Discussion and Conclusion

users can include metadata, such as transcriptomic or proteomic information, to connect genomic findings with the corresponding layers.

For the remaining two projects presented in this thesis, the integration explicitly links multiple layers of the central dogma, with information from one layer directly translating in insights of another layer. In Project [II](#), based on transcriptomic data, **TSSpredator-Web** allows the user to generate insights into genomic positions linked to the start of gene expression, such as the exploration of surrounding genes or upstream regulatory regions. With such possibilities, researchers can generate insights into the genomic aspects of regulatory mechanisms behind specific genes. Building on this concept of cross-layer insight, **TSS-Captur** (Project [III](#)) also connects genomic positions that have been identified via transcriptomic data with the genomic sequence information to reconstruct putative transcripts, including previously possibly overlooked non-coding RNAs and small protein-coding genes. Given such a detailed characterization, **TSS-Captur** provides a foundation for further functional analysis at the transcriptomic and proteomic levels. Taken together, while the developed applications differ in their outreach on input data across the central dogma, they consistently act as bridges between the molecular layers. Collectively, they allow researchers to gain a broader and more interconnected view of bacterial function.

The majority of the projected presented here also highlight a common methodological strategy followed in this thesis that enhances the insightful generation via enabling users to access and interact with the tools without technical barriers. In particular, in bioinformatics, it is a common practice that new methods are made publicly available through command-line tools that require users to install several dependencies before analyzing their data with novel approaches (Liaw and Gehlenborg [2022](#)). This is also the case for the tools presented in this thesis, as they all depend on other programs or modules to work properly. For example, **Evidente** relies on **CLASSICO** to analyze the SNP distribution along the phylogenetic tree, and **TSS-Captur** bases its predictions of possible genes on the results of **CNIT** and **BLAST**. Although solving such dependencies has become a routine task for computational experts, this is a barrier for many biologists, who are the domain experts. To address these accessibility-related issues, the Tübingen Visualization (TueVis) server was developed^{[1](#)}. The idea behind this platform is to provide an overview of accessible tools for a broad spectrum of researchers, including computational researchers and domain experts, similar to the platform **Galaxy** (Jalili *et al.* [2020](#)). The focus of the TueVis is on the visual analysis of omics data. It categorizes the tools into six groups depending on the main type of input data: Genomics, transcriptomics and proteomics, similarly as defined by the central dogma of molecular biology, with the additional categories of multi-omics, phylogenomics and systems biology. Three of the software tools presented in this thesis (**Evidente**, **TSSpredator-Web** and **TSS-Captur**) are part of the nineteen tools currently hosted on TueVis, as well as three other web-applications to which I contributed. These are: **GO-Compass**, which allows the visual exploration of GO enrichment results (Harbig *et al.* [2023](#)), **PTM-Vision**, an application for the analysis of post-translational modifications in bacteria (Hackl *et al.* [2025](#)), and **ProtEgoNist**, an ego-network centered approach for exploring large networks (Brich *et al.* [2024](#)). Such a user-friendly platform allows experts of different fields to take profit of the tools, as it is also the case in **Galaxy**.

In contrast to **Galaxy**, the focus of the tools provided via TueVis, including those presented in this thesis, is to provide an explorative approach to the data, a further methodological feature for enhanced insight generation explored within this thesis. The main approaches followed by the tools presented in this thesis are visual exploration, in the cases of **Evidente** and **TSSpredator-Web**, and accessibility of the rationale behind

¹<https://tuevis.cs.uni-tuebingen.de/> accessed on February 17, 2026

specific predictions, as in the case of **TSS-Captur**. These applications differ from the traditional hypothesis-driven approaches, where computational methods should directly answer a specific research question, as they allow for the generation of new hypothesis via data exploration. **Evidente** enables visual exploration of the SNP distribution throughout the phylogenetic tree so that users can identify correlations between them and further metadata. Similarly, **TSSpredator-Web** enables the researcher to create a summary of several TSS positions in order to subsequently formulate hypotheses, such as whether a certain class of TSS accumulates in specific genomic regions or if there are differences within the promoter regions that can then be further investigated with other visualizations provided. Lastly, **TSS-Captur** provides the possibility to improve genome annotation starting from identified TSS positions and thus allows for an improved understanding of prokaryotic genomes.

The focus of these three projects was to provide accessible and user-friendly tools for data exploration. As the **PRESERVE** pipeline and the computational analyses applied for the exploration of the metallophore landscape were developed to answer a domain-specific research question rather than a general web-based deployment, they are currently optimized for computational experts and not domain experts.

The pipeline **PRESERVE** was optimized for the characterization of *Corynebacteria* species, prioritizing the depth of knowledge on the specific aspects of siderophore characterization over generating a broad and generalizable approach. During this process, **PRESERVE** has been developed as a ready-to-use Nextflow pipeline², which can be used by any experienced user, with sufficient computational resources (e.g. graphic cards for faster computation). While the tools found on the TueVis prioritize immediate accessibility for domain experts, **PRESERVE** addresses a different relevant methodological aspect mentioned throughout this thesis: reproducibility. As a Nextflow pipeline, it easily connects various analysis steps to get consistent results over multiple runs. In addition to that, it enables the installation of all dependencies using **conda** environments. Similarly, **TSS-Captur** allows for high reproducibility since the analyses are also computed through a Nextflow pipeline and depend on a Docker image for the containerization of all dependencies.

To increase the reproducibility and user-friendly exploration of metallophore synthesis and uptake analyzes, the prototype pipeline **SID-Chart** (SIDerophore CHARacterization Tool) was developed in the context of a supervised bachelor thesis³. Similar to **PRESERVE** and **TSS-Captur**, **SID-Chart** encapsulates all steps for the characterization within a Nextflow pipeline to enhance reproducibility. In addition, an interactive interface allows users to get an overview of the results, as well as identify specific strains with the use of xenosiderophore. This prototype provides a basis for future development of a web-application that facilitates the overall exploration of metallophore synthesis and uptake in other bacteria.

Currently, **Evidente** and **TSSpredator-Web** are designed to prioritize interactive real-time exploration of complex biological data. Both tools focus on enabling a visual analysis rather than providing a full reproducible environment of the exploratory sessions. While **TSSpredator-Web** supports re-analysis through configuration files and allows temporary sharing of results via URL links, downstream interactive exploration steps are not yet stored for later reuse. A similar design applies to **Evidente**. The reproducibility of interactive analysis workflows is commonly addressed through visualization provenance

²https://github.com/Integrative-Transcriptomics/preserve-nextflow/tree/Raffael_BA - the latest version by Raffael Rack, Bachelor thesis, supervised by Mathias Witte Paz and Simon Hackl.

³<https://github.com/Integrative-Transcriptomics/SID-Chart>, developed by Lukas Lorentz, Bachelor thesis, supervised by Mathias Witte Paz.

6. Discussion and Conclusion

approaches (Gratzl *et al.* 2016). Integrating such features would further enhance the reproducibility of exploratory analyses in future tool iterations. At the same time, both tools support reproducible deployment by providing fully containerized versions of web applications through their respective repositories. Beyond making local execution possible, containerization preserves a stable software environment and facilitates long-term reuse, which is particularly important in web-based applications that depend on fast evolving libraries and frameworks (Cito *et al.* 2016).

In the discussion so far, three key aspects for methodological advances were pointed out: web-based deployment for better accessibility, enabling interactive and integrative exploration of data, and reproducibility by standardized workflows and containerization of tools. These advances demonstrate how computational design decisions impact not only speed, but also depth of insight in biological inquiry. Each of the strategies developed has its unique strengths, but together they highlight a shift from static, hypothesis-driven analyses to dynamic, user-focused, and high-throughput exploration of omics data. This methodological shift is necessary because new biological hypotheses are emerging from the intersection between interactive data exploration and robust reproducible software. Having established how these approaches help researchers overcome barriers in data analysis, the next section will focus on the biological insights that have been or could be gained through these methods for characterizing and controlling microbes.

In alignment with the goals of the cluster of excellence *Controlling Microbes to Fight Infection* (CMFI), which funded the research work presented in this thesis, the aim of characterizing microbes should provide a means of generating new methods for their control. The use cases presented for each project establish a proof-of-concept for the characterization of prokaryotes. However, the tools presented can be applied to various control scenarios, offering the generation of broader insights for this goal.

A use-case with **Evidente** showed how easily the tool can be used to replicate the identification of SNP positions that correlate with antibiotic resistance phenotypes. Interestingly, other studies have indicated that an integrative analysis of genome-wide SNP mutations and metadata is useful for developing control strategies against different microbes. For example, SNP data and phylogenetic analyses have been integrated with patient metadata to trace the transmission of epidemic strains and find highly contagious individuals (Kumar *et al.* 2016). Researchers using **Evidente** could investigate the SNP data deeper and determine which SNP mutations may be associated with these individuals and the molecular functions affected by them. In addition, SNP data have helped identify resistance microbes (Alam *et al.* 2014; Stubberfield *et al.* 2021), which in turn helps to advance patient treatment, such as alteration of antibiotic usage or dose due to variations in the minimum inhibitory concentration. Integration and visualization of data within **Evidente** may help to develop methods to find the best patient treatment based on the given bacterial isolate.

Regarding the transcriptomic layer, increased availability and data processing can provide ways to regulate pathogenicity of microbes. For example, we have shown how genome-wide TSS maps generated with **TSSpredator-Web** can be used to identify the effects of multiple antibiotics on the transcriptome and its architecture. In particular, such TSS maps enable the identification of the exact position of promoter regions, which harbor regulatory mechanisms and provide insight into the corresponding transcription factors. Repression of such transcription factors has been shown to help regulate the pathogenic characteristics of specific organisms, such as *Clostridioides difficile* (Soutourina *et al.* 2020) or *Listeria monocytogenes* (Al-Tohamy and Grove 2025). In such studies, the interface of **TSSpredator-Web** can provide a more accessible analysis as well as an intuitive manner

to explore the generated genome-wide TSS set and hence speed-up the insight generation process. In addition to insight generation, TSS maps also allow the discovery of new small non-coding RNA, some of which have been related to several pathogenicity factors (Khaledi *et al.* [2025]; Zhukova *et al.* [2017]) or resistance mechanisms, such as *ompX* in *Salmonella Typhimurium* (Briones *et al.* [2022]). Thus, a genome-wide follow-up characterization of TSS with **TSS-Captur** has the potential to identify other overlooked key players within the genome of pathogens.

Of the projects described in this thesis, both projects focusing on metallophore characterization clearly illustrate how new computational methods can open completely new directions for microbial control. For example, first approaches have already exploited siderophores to increase antibiotic efficiency (Górska *et al.* [2014]). Moreover, both projects help to understand how ecological niches are formed and whether commensals can be used as probiotics for the inhibition of pathogens (Kramer *et al.* [2020]). Also, the **PRESERVE** pipeline can be applied to identify structural homologs of different proteins needed for interactions with the environment, such as quorum sensing proteins (Bassler [1999]; Mukherjee and Bassler [2019]). As this pipeline also offers information on regulatory mechanisms, it can lead to the identification of compounds that might enhance or repress specific transcription factors involved in the iron-uptake mechanism, as has been identified for other prokaryotic pathways (Binsfeld *et al.* [2025]).

Finally, the applications developed in this thesis might, in principle, be combined to characterize multiple molecular layers of a single organism, thus complementing standard annotation pipelines and allowing the construction of more accurate computational models, such as genome-scale metabolic models (Bäuerle *et al.* [2023]). By integrating genomic variation, transcriptomic regulation, and functional or structural protein information, such models may better reflect fitness, pathogenicity, and interaction potential under different environmental conditions. In that respect, the tools introduced here do not intend to replace existing approaches but to complement them by enabling exploratory, hypothesis-generating analyses that are directly linked to possible microbial control strategies. Together, these projects demonstrate how integrative, accessible bioinformatics methods can translate complex multi-omics data into mechanistic insights. Hence, they align with the objectives of the cluster of excellence *CMFI* and point to the potential of accessible and exploratory analyses to drive a new era of discovery in antimicrobial strategies. However, the route from computational insight to clinical or environmental application involves limitations that extend beyond the current implementation of these tools. These limitations do not reduce the methodological contributions of this work, but instead identify clear boundaries for their interpretation and provide an avenue for further development in the characterization of bacteria.

First, all tools were designed with integration in mind, while interoperability would be a focus for future developments. While currently the tools work in a standalone manner, providing a closer technical coupling between these tools and the existing TueVis applications would enhance the potential of integrative microbial analysis. For example, a direct transfer of data from **TSSpredator-Web** to **TSS-Captur** will provide the user with a direct analysis of all uncharacterized positions. Further analysis of predicted loci encoding uncharacterized proteins could be extended by a structural search with **PRESERVE** against protein databases, such as the PDB (Burley *et al.* [2017]) or the **AlphaFoldDB** (Varadi *et al.* [2022]). This extension would overcome the limitations of standard sequence homology pipelines, as it enforces the idea that structural homology reflects a higher functional similarity. Another extension would be to connect **Evidente** with the tool **MUSIAL**⁴ (Multi

⁴<https://musial-tuevis.cs.uni-tuebingen.de/>, developed by Simon Hackl

Sample variant AnaLysis), as **MUSIAL** can produce the input SNP table based on variant calling files (VCF), providing an even faster analysis of the data. Moreover, the GO-enrichment analyses computed within **Evidente** could be enhanced by connecting them to **GO-Compass**, as it would deliver improved results by taking into account the hierarchy of GO terms and facilitate a more explorative interaction of the results. An example of such a GO-enrichment analysis was performed as a use-case within the **GO-Compass** (Harbig *et al.* 2023). These technical extensions would not only improve the analysis of the individual tools, but also establish the TueVis platform as a robust, multi-layered environment for the integrative characterization of prokaryotic organism.

A central contribution of this work is the enhancement of tool accessibility and usability for the microbial research community. While these tools establish a robust basis for interactive data analysis, future developments will provide the opportunity to include formal large-scale evaluations, such as user studies with a diverse cohort of researchers. As the recruitment of domain experts for such studies remains a proven challenge in bioinformatics software development (Davis *et al.* 2023), this thesis has addressed usability based on exploratory approaches of published articles. By identifying the exploratory requirements of the literature as described in the projects I and II, we ensured that the tools are able to provide an easy exploration of the complex steps needed for prokaryotic characterization.

Although the applications in this thesis have been proven with known biological data, their true value lies in their ability to help in the generation of new hypotheses. By providing a way to explore complex biological data, the novel findings identified by these tools serve as a guide for future hypothesis-driven experimental validation. For example, the use case presented in the context of **TSSpredator-Web** identified a transcript originating from an orphan TSS. This signal represents a valuable candidate for further characterization with established experimental approaches such as RT-qPCR (Freeman *et al.* 1999), Western blots (Towbin *et al.* 1979), or Northern blots (Alwine *et al.* 1977). Such experiments would allow the confirmation of the expression of the transcripts and the evaluation of their coding or non-coding potential, similar to the analysis presented by Vockenhuber *et al.* (2011).

In the same way, the insights described in the context of **TSS-Captur** and by the metallophore characterization analyses provide a road-map for targeted experimental validation. In the study of siderophore uptake in *Corynebacterium* species (Project V), the computational analyses successfully provided a list of high-confidence candidates from a diverse genomic landscape, significantly narrowing the focus of future studies. Extending this analysis to include proteomic quantification would offer a more complete view of how these transcriptional signals translate into functional mechanisms and complete the characterization across the central dogma. Such a complete characterization of the siderophore mechanisms would allow highly precise knock-out experiments and co-culture assays with *S. aureus* to quantify the essentiality of the genes, as presented by Zhao *et al.* (2024). Moreover, the predicted role of DtxR as a regulator of siderophore uptake genes in *Corynebacteria* provides a clear next step for experimental validation, with corresponding studies being currently conducted in collaboration with Alina Bitzer. A targeted deletion of this transcription factor should confirm the role within regulatory mechanisms, as identified through our integrative computational analysis.

Lastly, while the projects presented in this thesis are all based on the central dogma of molecular biology, they prioritize the genomic and transcriptomic layers as the primary drivers of integrative analysis. The scope of this thesis covers the proteomic layer via **PRESERVE**, which provides structural insight based on the genomic layer, and via

TSS-Captur, which is able to identify previously overlooked protein-coding regions, based on genomic positions supported by experimental data. While the high-throughput quantification of protein expression levels, derived from mass-spectrometry data, remained outside the primary scope of this thesis, I contributed to the conceptualization of **PTM-Vision** (Hackl *et al.* [2025]), a platform for the exploration of post-translational modifications. Integrating more proteomic data analysis into the developed workflows represents a next step, offering the potential to further refine the functional characterization of prokaryotic organisms.

Moreover, instead of focusing on the characterization of single organisms, future focus could take advantage of current methods that allow simultaneous analysis of multiple species. For example, integrative analysis of the composition of whole bacterial communities can be explored using whole-genome metagenomics (Tyson *et al.* [2004]) and could also profit from the analysis of metatranscriptomics approaches (Frias-Lopez *et al.* [2008]). Moreover, even simpler compositions could benefit from exploratory analyses of multiple organisms, such as the analysis of the bacterium's transcriptome when interacting with another bacterium (Tognon *et al.* [2019]), or during bacteriophage infection (Westermann *et al.* [2012]), which are currently on the rise as resources for new antimicrobial strategies (Skurnik [2022]). By providing explorative and integrative analyses of such datasets, researchers would be able to reflect a more realistic environment to which organisms are exposed.

As the amount of data to be analyzed increases, the complexity of integration also increases. For these reasons, newer applications based on machine learning methods facilitate the analysis and integration of biological data. Recent advances have enabled highly accurate predictions in many biological domains, such as genomic language models that predict gene expression only based on sequence (Barbadilla-Martínez *et al.* [2025]) or machine learning methods that identify mechanisms for antimicrobial resistance based on transcriptomic data (Shahreen *et al.* [2025]). Integrating machine learning models would, therefore, allow future extensions of the presented tools to benefit from these recent computational developments. For example, further developments of **TSSpredator-Web** could integrate predictive models for base-specific TSS identification using only standard RNA-seq libraries. This would significantly lower experimental and financial barriers, as enriched libraries, while ideal, are not always accessible due to technical constraints or limited funding. Moreover, protein language models (Liu *et al.* [2024]) could be used to identify potential homologs and replace the highly-demanding computational step of prediction of protein structures. However, their applicability in microbiology remains limited by the lack of interpretability. To generate insights, researchers should be able to trace the results back to underlying biological features, such as genomic loci, regulatory elements, or structural motifs. Hence, further research should be directed towards the exploration and visualization of these methods to better interpret their results.

A final next step beyond the development of individual analysis tools is the aggregation of their results into explorable interfaces, such as specialized atlases of microbial knowledge. This differs from raw data repositories, as these interfaces allow the exploration of results, without the need of a reanalysis, as seen in the BGC atlas (Bagci *et al.* [2025]) or the **PhageExpressionAtlas**⁵. Applying this atlas-based approach to the outputs of **Evidente**, **TSSpredator-Web**, **TSS-Captur**, and **SID-Chart** would increase access to these complex datasets, allowing a wider research community to uncover novel biological insights that might otherwise remain hidden in raw data.

In conclusion, this thesis demonstrates that user-centered and integrative computational tools are needed to navigate the complexity of omics datasets and support hypothesis

⁵<https://phageexpressionatlas.cs.uni-tuebingen.de/>, developed mainly by Maik Wolfram-Schauerte

6. Discussion and Conclusion

generation in microbiology. By bridging the individual layers of the central dogma through a focus on accessibility, reproducibility, and exploration, the work presented here provided a methodological framework for a more complete understanding of prokaryotic systems. Future efforts focusing on tool interoperability, large-scale experimental validation, and community-level analyses will be crucial to translate these exploratory insights into actionable strategies to control microbial behavior. As omics technologies continue to grow in volume and complexity, the integration of all molecular layers through transparent and interactive computational frameworks will remain a key aspect to characterizing microbes and ultimately controlling their infections.

Bibliography

- Aebersold, R. and Mann, M. (2016). “Mass-Spectrometric Exploration of Proteome Structure and Function”. In: *Nature* vol. 537, no. 7620, pp. 347–355. ISSN: 1476-4687. DOI: [10.1038/nature19949](https://doi.org/10.1038/nature19949).
- Alam, M. T., Petit, R. A., Crispell, E. K., Thornton, T. A., Conneely, K. N., Jiang, Y., Satola, S. W., and Read, T. D. (2014). “Dissecting Vancomycin-Intermediate Resistance in Staphylococcus Aureus Using Genome-Wide Association”. In: *Genome Biology and Evolution* vol. 6, no. 5, pp. 1174–1185. ISSN: 1759-6653. DOI: [10.1093/gbe/evu092](https://doi.org/10.1093/gbe/evu092).
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). “Basic local alignment search tool”. In: *Journal of Molecular Biology* vol. 215, no. 3, pp. 403–410. ISSN: 0022-2836. DOI: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2). URL: <https://www.sciencedirect.com/science/article/pii/S0022283605803602>.
- Alwine, J. C., Kemp, D. J., and Stark, G. R. (1977). “Method for Detection of Specific RNAs in Agarose Gels by Transfer to Diazobenzyloxymethyl-Paper and Hybridization with DNA Probes.” In: *Proceedings of the National Academy of Sciences* vol. 74, no. 12, pp. 5350–5354. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.74.12.5350](https://doi.org/10.1073/pnas.74.12.5350).
- Amman, F., Wolfinger, M. T., Lorenz, R., Hofacker, I. L., Stadler, P. F., and Findeiß, S. (2014). “TSSAR: TSS Annotation Regime for dRNA-seq Data”. In: *BMC Bioinformatics* vol. 15, no. 1, p. 89. ISSN: 1471-2105. DOI: [10.1186/1471-2105-15-89](https://doi.org/10.1186/1471-2105-15-89).
- Anscombe, F. J. (1973). “Graphs in Statistical Analysis”. In: *The American Statistician* vol. 27, no. 1, pp. 17–21. ISSN: 0003-1305. DOI: [10.1080/00031305.1973.10478966](https://doi.org/10.1080/00031305.1973.10478966).
- Arber, W. (2000). “Genetic Variation: Molecular Mechanisms and Impact on Microbial Evolution”. In: *FEMS Microbiology Reviews* vol. 24, no. 1, pp. 1–7. ISSN: 0168-6445. DOI: [10.1111/j.1574-6976.2000.tb00529.x](https://doi.org/10.1111/j.1574-6976.2000.tb00529.x).
- Arora, N., Schuenemann, V. J., Jäger, G., Peltzer, A., Seitz, A., Herbig, A., Strouhal, M., Grillová, L., Sánchez-Busó, L., Kühnert, D., *et al.* (2016). “Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster”. In: *Nature Microbiology* vol. 2, no. 1, pp. 1–6.
- Backofen, R., Amman, F., Costa, F., Findeiß, S., Richter, A. S., and Stadler, P. F. (2014). “Bioinformatics of Prokaryotic RNAs”. In: *RNA Biology* vol. 11, no. 5, pp. 470–483. DOI: [10.4161/rna.28647](https://doi.org/10.4161/rna.28647).
- Bagci, C., Nuhamunada, M., Goyat, H., Ladanyi, C., Sehnal, L., Blin, K., Kautsar, S. A., Tagirdzhanov, A., Gurevich, A., Mantri, S., von Mering, C., Udworthy, D., Medema, M. H., Weber, T., and Ziemert, N. (2025). “BGC Atlas: A Web Resource for Exploring the Global Chemical Diversity Encoded in Bacterial Genomes”. In: *Nucleic Acids Research* vol. 53, no. D1, pp. D618–D624. ISSN: 1362-4962. DOI: [10.1093/nar/gkae953](https://doi.org/10.1093/nar/gkae953).
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). “MEME SUITE: Tools for Motif Discovery and Searching”. In: *Nucleic Acids Research* vol. 37, no. suppl_2, W202–W208. DOI: [10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335).
- Balkin, A., Plotnikov, A., Konnova, T., Shagimardanova, E., Hamo, H., Gogolev, Y., and Gogoleva, N. (2025). “Cappable-Seq RNA-sequencing Data Sets of *Escherichia coli* K-12 MG1655 Treated with Novobiocin, Tetracycline, and Rifampicin”. In: *Microbiology Resource Announcements* vol. 14, no. 2, e01194–24. DOI: [10.1128/mra.01194-24](https://doi.org/10.1128/mra.01194-24).

- Barbadilla-Martínez, L., Klaassen, N., van Steensel, B., and de Ridder, J. (2025). “Predicting Gene Expression from DNA Sequence Using Deep Learning Models”. In: *Nature Reviews Genetics* vol. 26, no. 10, pp. 666–680. ISSN: 1471-0064. DOI: [10.1038/s41576-025-00841-2](https://doi.org/10.1038/s41576-025-00841-2).
- Barquist, L., Burge, S. W., and Gardner, P. P. (2016). “Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families”. In: *Current Protocols in Bioinformatics* vol. 54, no. 1. ISSN: 1934-340X. DOI: [10.1002/cpbi.4](https://doi.org/10.1002/cpbi.4).
- Bassler, B. L. (1999). “How Bacteria Talk to Each Other: Regulation of Gene Expression by Quorum Sensing”. In: *Current Opinion in Microbiology* vol. 2, no. 6, pp. 582–587. ISSN: 1369-5274. DOI: [10.1016/S1369-5274\(99\)00025-9](https://doi.org/10.1016/S1369-5274(99)00025-9).
- Bäuerle, F., Döbel, G. O., Camus, L., Heilbronner, S., and Dräger, A. (2023). “Genome-Scale Metabolic Models Consistently Predict in Vitro Characteristics of *Corynebacterium Striatum*”. In: *Frontiers in Bioinformatics* vol. 3, p. 1214074. ISSN: 2673-7647. DOI: [10.3389/fbinf.2023.1214074](https://doi.org/10.3389/fbinf.2023.1214074).
- Beasley, F. C., Marolda, C. L., Cheung, J., Buac, S., and Heinrichs, D. E. (2011). “Staphylococcus aureus Transporters Hts, Sir, and Sst Capture Iron Liberated from Human Transferrin by staphyloferrin A, staphyloferrin B, and Catecholamine Stress Hormones, Respectively, and Contribute to Virulence”. In: *Infection and Immunity* vol. 79, no. 6, pp. 2345–2355. ISSN: 1098-5522. DOI: [10.1128/IAI.00117-11](https://doi.org/10.1128/IAI.00117-11).
- Beasley, F. C., Vinés, E. D., Grigg, J. C., Zheng, Q., Liu, S., Lajoie, G. A., Murphy, M. E. P., and Heinrichs, D. E. (2009). “Characterization of Staphyloferrin A Biosynthetic and Transport Mutants in *Staphylococcus aureus*”. In: *Molecular Microbiology* vol. 72, no. 4, pp. 947–963. ISSN: 1365-2958. DOI: [10.1111/j.1365-2958.2009.06698.x](https://doi.org/10.1111/j.1365-2958.2009.06698.x).
- Becker, K., Heilmann, C., and Peters, G. (2014). “Coagulase-Negative Staphylococci”. In: *Clinical Microbiology Reviews* vol. 27, no. 4, pp. 870–926. DOI: [10.1128/cmr.00109-13](https://doi.org/10.1128/cmr.00109-13).
- Bentley, S. D. and Parkhill, J. (2004). “Comparative Genomic Structure of Prokaryotes”. In: *Annual Review of Genetics* vol. 38, no. Volume 38, 2004, pp. 771–791. ISSN: 0066-4197, 1545-2948. DOI: [10.1146/annurev.genet.38.072902.094318](https://doi.org/10.1146/annurev.genet.38.072902.094318).
- Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., and Trees, E. (2018). “Next-Generation Sequencing Technologies and Their Application to the Study and Control of Bacterial Infections”. In: *Clinical Microbiology and Infection* vol. 24, no. 4, pp. 335–341. ISSN: 1198-743X. DOI: [10.1016/j.cmi.2017.10.013](https://doi.org/10.1016/j.cmi.2017.10.013).
- Binsfeld, C., Olayo-Alarcon, R., Jiménez, L. P., Wartel, M., Stadler, M., Mateus, A., Müller, C., and Brochado, A. R. (2025). “Systematic Screen Uncovers Regulator Contributions to Chemical Cues in *Escherichia Coli*”. In: *PLOS Biology* vol. 23, no. 7, e3003260. ISSN: 1545-7885. DOI: [10.1371/journal.pbio.3003260](https://doi.org/10.1371/journal.pbio.3003260).
- Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J. N., van Wezel, G. P., Medema, M. H., and Weber, T. (2023). “antiSMASH 7.0: New and Improved Predictions for Detection, Regulation, Chemical Structures and Visualisation”. In: *Nucleic Acids Research* vol. 51, no. W1, W46–W50. ISSN: 0305-1048. DOI: [10.1093/nar/gkad344](https://doi.org/10.1093/nar/gkad344).
- Bognár, B., Spohn, R., and Lázár, V. (2024). “Drug Combinations Targeting Antibiotic Resistance”. In: *npj Antimicrobials and Resistance* vol. 2, no. 1, p. 29. ISSN: 2731-8745. DOI: [10.1038/s44259-024-00047-2](https://doi.org/10.1038/s44259-024-00047-2).
- Bomar, L., Brugger, S. D., and Lemon, K. P. (2018). “Bacterial Microbiota of the Nasal Passages across the Span of Human Life”. In: *Current Opinion in Microbiology* vol. 41, pp. 8–14. ISSN: 1879-0364. DOI: [10.1016/j.mib.2017.10.023](https://doi.org/10.1016/j.mib.2017.10.023).
- Brich, N., Harbig, T. A., Witte Paz, M., Nieselt, K., and Krone, M. (2024). “ProtEGOnist: Visual Analysis of Interactions in Small World Networks Using Ego-graphs”. In: *Computer Graphics Forum* vol. 43, no. 3, e15078. DOI: [10.1111/cgf.15078](https://doi.org/10.1111/cgf.15078).

- Briones, A. C., Lorca, D., Cofre, A., Cabezas, C. E., Krüger, G. I., Pardo-Esté, C., Baquedano, M. S., Salinas, C. R., Espinoza, M., Castro-Severyn, J., Remonsellez, F., Hidalgo, A. A., Morales, E. H., and Saavedra, C. P. (2022). “Genetic Regulation of the ompX Porin of *Salmonella Typhimurium* in Response to Hydrogen Peroxide Stress”. In: *Biological Research* vol. 55, no. 1, p. 8. ISSN: 0717-6287. DOI: [10.1186/s40659-022-00377-3](https://doi.org/10.1186/s40659-022-00377-3).
- Brown, E. D. and Wright, G. D. (2016). “Antibacterial Drug Discovery in the Resistance Era”. In: *Nature* vol. 529, no. 7586, pp. 336–343. ISSN: 1476-4687. DOI: [10.1038/nature17042](https://doi.org/10.1038/nature17042).
- Brozyna, J. R., Sheldon, J. R., and Heinrichs, D. E. (2014). “Growth Promotion of the Opportunistic Human Pathogen, *Staphylococcus lugdunensis*, by Heme, Hemoglobin, and Coculture with *Staphylococcus aureus*”. In: *MicrobiologyOpen* vol. 3, no. 2, pp. 182–195. ISSN: 2045-8827. DOI: [10.1002/mbo3.162](https://doi.org/10.1002/mbo3.162).
- Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). “Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive”. In: *Protein Crystallography*, pp. 627–641.
- Chauvier, A. and Walter, N. G. (2024). “Regulation of Bacterial Gene Expression by Non-Coding RNA: It Is All about Time!” In: *Cell Chemical Biology* vol. 31, no. 1, pp. 71–85. ISSN: 2451-9456, 2451-9448. DOI: [10.1016/j.chembiol.2023.12.011](https://doi.org/10.1016/j.chembiol.2023.12.011). PMID: [38211587](https://pubmed.ncbi.nlm.nih.gov/38211587/).
- Cheah, H.-L., Citartan, M., Lee, L.-P., Ahmed, S. A., Salleh, M. Z., Teh, L. K., and Tang, T.-H. (2024). “Exploring the Transcription Start Sites and Other Genomic Features Facilitates the Accurate Identification and Annotation of Small RNAs across Multiple Stress Conditions in *Mycobacterium Tuberculosis*”. In: *Functional & Integrative Genomics* vol. 24, no. 5, p. 160. ISSN: 1438-7948. DOI: [10.1007/s10142-024-01437-5](https://doi.org/10.1007/s10142-024-01437-5).
- Cito, J., Ferme, V., and Gall, H. C. (2016). “Using Docker Containers to Improve Reproducibility in Software and Web Engineering Research”. In: *Web Engineering*. Ed. by Bozzon, A., Cudre-Maroux, P., and Pautasso, C. Vol. 9671. Cham: Springer International Publishing, pp. 609–612. ISBN: 978-3-319-38791-8. DOI: [10.1007/978-3-319-38791-8_58](https://doi.org/10.1007/978-3-319-38791-8_58).
- Cordero, O. X., Ventouras, L.-A., DeLong, E. F., and Polz, M. F. (2012). “Public Good Dynamics Drive Evolution of Iron Acquisition Strategies in Natural Bacterioplankton Populations”. In: *Proceedings of the National Academy of Sciences* vol. 109, no. 49, pp. 20059–20064. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1213344109](https://doi.org/10.1073/pnas.1213344109).
- Cotton, J. L., Tao, J., and Balibar, C. J. (2009). “Identification and Characterization of the *Staphylococcus Aureus* Gene Cluster Coding for Staphyloferrin A”. In: *Biochemistry* vol. 48, no. 5, pp. 1025–1035. ISSN: 1520-4995. DOI: [10.1021/bi801844c](https://doi.org/10.1021/bi801844c).
- Crick, F. (1958). “On Protein Synthesis”. In: *Symposia of the Society for Experimental Biology* vol. 12, pp. 138–163. ISSN: 0081-1386. PMID: [13580867](https://pubmed.ncbi.nlm.nih.gov/13580867/).
- (1970). “Central Dogma of Molecular Biology”. In: *Nature* vol. 227, no. 5258, pp. 561–563. ISSN: 1476-4687. DOI: [10.1038/227561a0](https://doi.org/10.1038/227561a0). URL: <https://doi.org/10.1038/227561a0>.
- Croucher, N. J. and Thomson, N. R. (2010). “Studying Bacterial Transcriptomes Using RNA-seq”. In: *Current Opinion in Microbiology*. Antimicrobials/Genomics vol. 13, no. 5, pp. 619–624. ISSN: 1369-5274. DOI: [10.1016/j.mib.2010.09.009](https://doi.org/10.1016/j.mib.2010.09.009).
- Cui, W. (2019). “Visual Analytics: A Comprehensive Overview”. In: *IEEE Access* vol. 7, pp. 81555–81573. ISSN: 2169-3536. DOI: [10.1109/ACCESS.2019.2923736](https://doi.org/10.1109/ACCESS.2019.2923736).
- Čuklina, J., Hahn, J., Imakaev, M., Omasits, U., Förstner, K. U., Ljubimov, N., Goebel, M., Pessi, G., Fischer, H.-M., Ahrens, C. H., Gelfand, M. S., and Evgueniev-Hackenberg, E. (2016). “Genome-Wide Transcription Start Site Mapping of *Bradyrhizobium japonicum*

- Grown Free-Living or in Symbiosis – a Rich Resource to Identify New Transcripts, Proteins and to Study Gene Regulation”. In: *BMC Genomics* vol. 17, no. 1, p. 302. ISSN: 1471-2164. DOI: [10.1186/s12864-016-2602-9](https://doi.org/10.1186/s12864-016-2602-9).
- Dale, S. E., Doherty-Kirby, A., Lajoie, G., and Heinrichs, D. E. (2004). “Role of Siderophore Biosynthesis in Virulence of *Staphylococcus aureus*: Identification and Characterization of Genes Involved in Production of a Siderophore”. In: *Infection and Immunity* vol. 72, no. 1, pp. 29–37. ISSN: 0019-9567. DOI: [10.1128/IAI.72.1.29-37.2004](https://doi.org/10.1128/IAI.72.1.29-37.2004).
- Dale, S. E., Sebulsky, M. T., and Heinrichs, D. E. (2004). “Involvement of SirABC in Iron-Siderophore Import in *Staphylococcus Aureus*”. In: *Journal of Bacteriology* vol. 186, no. 24, pp. 8356–8362. ISSN: 0021-9193. DOI: [10.1128/JB.186.24.8356-8362.2004](https://doi.org/10.1128/JB.186.24.8356-8362.2004).
- Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., and Sorek, R. (2016). “Term-Seq Reveals Abundant Ribo-Regulation of Antibiotics Resistance in Bacteria”. In: *Science* vol. 352, no. 6282, aad9822. DOI: [10.1126/science.aad9822](https://doi.org/10.1126/science.aad9822).
- Davis, M. C., Aghayi, E., Latoza, T. D., Wang, X., Myers, B. A., and Sunshine, J. (2023). “What’s (Not) Working in Programmer User Studies?” In: *ACM Trans. Softw. Eng. Methodol.* vol. 32, no. 5, 120:1–120:32. ISSN: 1049-331X. DOI: [10.1145/3587157](https://doi.org/10.1145/3587157).
- DeLano, W. L. *et al.* (2002). “Pymol: An open-source molecular graphics tool”. In: *CCP4 Newsl. protein crystallogr* vol. 40, no. 1, pp. 82–92.
- Di Salvo, M., Puccio, S., Peano, C., Lacour, S., and Alifano, P. (2019). “RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on *Escherichia coli*, *Bacillus subtilis* and *Salmonella enterica* databases”. In: *BMC Bioinformatics* vol. 20, no. 1, p. 117. ISSN: 1471-2105. DOI: [10.1186/s12859-019-2704-x](https://doi.org/10.1186/s12859-019-2704-x). URL: <https://doi.org/10.1186/s12859-019-2704-x>.
- Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). “Nextflow enables reproducible computational workflows”. In: *Nature Biotechnology* vol. 35, no. 4, pp. 316–319. ISSN: 1546-1696. DOI: [10.1038/nbt.3820](https://doi.org/10.1038/nbt.3820). URL: <https://doi.org/10.1038/nbt.3820>.
- Drazic, A., Myklebust, L. M., Ree, R., and Arnesen, T. (2016). “The World of Protein Acetylation”. In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* vol. 1864, no. 10, pp. 1372–1401. ISSN: 1570-9639. DOI: [10.1016/j.bbapap.2016.06.007](https://doi.org/10.1016/j.bbapap.2016.06.007).
- Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., and Sharma, C. M. (2013). “High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple *Campylobacter jejuni* Isolates”. In: *PLOS Genetics* vol. 9, no. 5, e1003495. DOI: [10.1371/journal.pgen.1003495](https://doi.org/10.1371/journal.pgen.1003495). URL: <https://doi.org/10.1371/journal.pgen.1003495>.
- Emms, D. M. and Kelly, S. (2019). “OrthoFinder: Phylogenetic Orthology Inference for Comparative Genomics”. In: *Genome Biology* vol. 20, no. 1, p. 238. ISSN: 1474-760X. DOI: [10.1186/s13059-019-1832-y](https://doi.org/10.1186/s13059-019-1832-y).
- Ettwiller, L., Buswell, J., Yigit, E., and Schildkraut, I. (2016). “A Novel Enrichment Strategy Reveals Unprecedented Number of Novel Transcription Start Sites at Single Base Resolution in a Model Prokaryote and the Gut Microbiome”. In: *BMC Genomics* vol. 17, no. 1, pp. 1–14. DOI: [10.1186/s12864-016-2539-z](https://doi.org/10.1186/s12864-016-2539-z).
- Faison, W. J., Rostovtsev, A., Castro-Nallar, E., Crandall, K. A., Chumakov, K., Simonyan, V., and Mazumder, R. (2014). “Whole Genome Single-Nucleotide Variation Profile-Based Phylogenetic Tree Building Methods for Analysis of Viral, Bacterial and Human Genomes”. In: *Genomics* vol. 104, no. 1, pp. 1–7. ISSN: 0888-7543. DOI: [10.1016/j.ygeno.2014.06.001](https://doi.org/10.1016/j.ygeno.2014.06.001).

- Flicek, P. and Birney, E. (2009). "Sense from Sequence Reads: Methods for Alignment and Assembly". In: *Nature Methods* vol. 6, no. 11, S6–S12. ISSN: 1548-7105. DOI: [10.1038/nmeth.1376](https://doi.org/10.1038/nmeth.1376).
- Ford, C. B., Lin, P. L., Chase, M. R., Shah, R. R., Iartchouk, O., Galagan, J., Mohaideen, N., Ioerger, T. R., Sacchettini, J. C., Lipsitch, M., Flynn, J. L., and Fortune, S. M. (2011). "Use of Whole Genome Sequencing to Estimate the Mutation Rate of Mycobacterium Tuberculosis during Latent Infection". In: *Nature Genetics* vol. 43, no. 5, pp. 482–486. ISSN: 1546-1718. DOI: [10.1038/ng.811](https://doi.org/10.1038/ng.811).
- Forde, B. M. and O'Toole, P. W. (2013). "Next-Generation Sequencing Technologies and Their Impact on Microbial Genomics". In: *Briefings in Functional Genomics* vol. 12, no. 5, pp. 440–453. ISSN: 2041-2649. DOI: [10.1093/bfgp/els062](https://doi.org/10.1093/bfgp/els062).
- Freeman, W. M., Walker, S. J., and Vrana, K. E. (1999). "Quantitative RT-PCR: Pitfalls and Potential". In: *BioTechniques* vol. 26, no. 1, pp. 112–125. ISSN: 0736-6205, 1940-9818. DOI: [10.2144/99261rv01](https://doi.org/10.2144/99261rv01).
- Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and DeLong, E. F. (2008). "Microbial Community Gene Expression in Ocean Surface Waters". In: *Proceedings of the National Academy of Sciences* vol. 105, no. 10, pp. 3805–3810. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.0708897105](https://doi.org/10.1073/pnas.0708897105).
- Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). "A Brief History of Bioinformatics". In: *Briefings in Bioinformatics* vol. 20, no. 6, pp. 1981–1996. ISSN: 1477-4054. DOI: [10.1093/bib/bby063](https://doi.org/10.1093/bib/bby063).
- Ghssein, G., Brutesco, C., Ouerdane, L., Fojcik, C., Izaute, A., Wang, S., Hajjar, C., Lobinski, R., Lemaire, D., Richaud, P., Voulhoux, R., Espaillet, A., Cava, F., Pignol, D., Borezée-Durant, E., and Arnoux, P. (2016). "Biosynthesis of a Broad-Spectrum Nicotianamine-like Metallophore in *Staphylococcus aureus*". In: *Science* vol. 352, no. 6289, pp. 1105–1109. ISSN: 1095-9203. DOI: [10.1126/science.aaf1018](https://doi.org/10.1126/science.aaf1018).
- Ghssein, G. and Ezzeddine, Z. (2022). "The Key Element Role of Metallophores in the Pathogenicity and Virulence of *Staphylococcus aureus*: A Review". In: *Biology* vol. 11, no. 10, p. 1525. ISSN: 2079-7737. DOI: [10.3390/biology11101525](https://doi.org/10.3390/biology11101525).
- Gilchrist, C. L. M. and Chooi, Y.-H. (2021). "Clinker & Clustermat: Automatic Generation of Gene Cluster Comparison Figures". In: *Bioinformatics* vol. 37, no. 16. Ed. by Robinson, P., pp. 2473–2475. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btab007](https://doi.org/10.1093/bioinformatics/btab007).
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). "Coming of Age: Ten Years of next-Generation Sequencing Technologies". In: *Nature Reviews Genetics* vol. 17, no. 6, pp. 333–351. ISSN: 1471-0064. DOI: [10.1038/nrg.2016.49](https://doi.org/10.1038/nrg.2016.49).
- Gordillo Altamirano, F. L. and Barr, J. J. (2019). "Phage Therapy in the Postantibiotic Era". In: *Clinical Microbiology Reviews* vol. 32, no. 2, 10.1128/cmr.00066–18. DOI: [10.1128/cmr.00066-18](https://doi.org/10.1128/cmr.00066-18).
- Górska, A., Sloderbach, A., and Marszał, M. P. (2014). "Siderophore–Drug Complexes: Potential Medicinal Applications of the 'Trojan Horse' Strategy". In: *Trends in Pharmacological Sciences* vol. 35, no. 9, pp. 442–449. ISSN: 0165-6147. DOI: [10.1016/j.tips.2014.06.007](https://doi.org/10.1016/j.tips.2014.06.007).
- Gottesman, S. (2005). "Micros for microbes: non-coding regulatory RNAs in bacteria". In: *Trends in Genetics* vol. 21, no. 7, pp. 399–404. ISSN: 0168-9525. DOI: [10.1016/j.tig.2005.05.008](https://doi.org/10.1016/j.tig.2005.05.008). URL: <https://www.sciencedirect.com/science/article/pii/S0168952505001381>.
- Grant, C. E. and Bailey, T. L. (2021). "XSTREME: Comprehensive Motif Analysis of Biological Sequence Datasets". In: *bioRxiv*, p. 2021.09.02.458722. DOI: [10.1101/2021.09.02.458722](https://doi.org/10.1101/2021.09.02.458722).

- Gratzl, S., Lex, A., Gehlenborg, N., Cosgrove, N., and Streit, M. (2016). "From Visual Exploration to Storytelling and Back Again". In: *Computer Graphics Forum* vol. 35, no. 3, pp. 491–500. ISSN: 0167-7055, 1467-8659. DOI: [10.1111/cgf.12925](https://doi.org/10.1111/cgf.12925).
- Gu, S., Shao, Z., Qu, Z., Zhu, S., Shao, Y., Zhang, D., Allen, R., He, R., Shao, J., Xiong, G., Jousset, A., Friman, V.-P., Wei, Z., Kümmerli, R., and Li, Z. (2025). "Siderophore Synthetase-Receptor Gene Coevolution Reveals Habitat- and Pathogen-Specific Bacterial Iron Interaction Networks". In: *Science Advances* vol. 11, no. 3. ISSN: 2375-2548. DOI: [10.1126/sciadv.adq5038](https://doi.org/10.1126/sciadv.adq5038).
- Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., and Serrano, L. (2009). "Transcriptome Complexity in a Genome-Reduced Bacterium". In: *Science* vol. 326, no. 5957, pp. 1268–1271. DOI: [10.1126/science.1176951](https://doi.org/10.1126/science.1176951).
- Guo, J.-C., Fang, S.-S., Wu, Y., Zhang, J.-H., Chen, Y., Liu, J., Wu, B., Wu, J.-R., Li, E.-M., Xu, L.-Y., Sun, L., and Zhao, Y. (2019). "CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition". In: *Nucleic Acids Research* vol. 47, no. W1, W516–W522. ISSN: 0305-1048. DOI: [10.1093/nar/gkz400](https://doi.org/10.1093/nar/gkz400).
- Hackl, S., Jachmann, C., Witte Paz, M., Harbig, T. A., Martens, L., and Nieselt, K. (2025). "PTMVision: An Interactive Visualization Webserver for Post-translational Modifications of Proteins". In: *Journal of Proteome Research* vol. 24, no. 2, pp. 919–928. DOI: [10.1021/acs.jproteome.4c00679](https://doi.org/10.1021/acs.jproteome.4c00679).
- Harbig, T. A. (2023). *Integrative-Transcriptomics/rnaSeq-pipeline: V1.0.0*. Zenodo. DOI: [10.5281/ZENODO.7500898](https://doi.org/10.5281/ZENODO.7500898).
- Harbig, T. A., Witte Paz, M., and Nieselt, K. (2023). "GO-Compass: Visual Navigation of Multiple Lists of GO Terms". In: *Computer Graphics Forum* vol. 42, no. 3, pp. 271–281. DOI: [10.1111/cgf.14829](https://doi.org/10.1111/cgf.14829).
- Hasin, Y., Seldin, M., and Lusis, A. (2017). "Multi-Omics Approaches to Disease". In: *Genome Biology* vol. 18, no. 1, p. 83. ISSN: 1474-760X. DOI: [10.1186/s13059-017-1215-1](https://doi.org/10.1186/s13059-017-1215-1).
- Hauser, M., Steinegger, M., and Söding, J. (2016). "MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets". In: *Bioinformatics* vol. 32, no. 9, pp. 1323–1330. ISSN: 1367-4811, 1367-4803. DOI: [10.1093/bioinformatics/btw006](https://doi.org/10.1093/bioinformatics/btw006).
- Hegemann, J. D., Birkelbach, J., Walesch, S., and Müller, R. (2022). "Current Developments in Antibiotic Discovery". In: *EMBO Reports* vol. 24, no. 1, e56184. ISSN: 1469-221X. DOI: [10.15252/embr.202256184](https://doi.org/10.15252/embr.202256184).
- Heilbronner, S. (2021). "Staphylococcus Lugdunensis". In: *Trends in Microbiology* vol. 29, no. 12, pp. 1143–1145. ISSN: 0966-842X, 1878-4380. DOI: [10.1016/j.tim.2021.07.008](https://doi.org/10.1016/j.tim.2021.07.008). PMID: [34462188](https://pubmed.ncbi.nlm.nih.gov/34462188/).
- Heilbronner, S., Holden, M. T. G., van Tonder, A., Geoghegan, J. A., Foster, T. J., Parkhill, J., and Bentley, S. D. (2011). "Genome Sequence of *Staphylococcus lugdunensis* N920143 Allows Identification of Putative Colonization and Virulence Factors". In: *FEMS microbiology letters* vol. 322, no. 1, pp. 60–67. ISSN: 1574-6968. DOI: [10.1111/j.1574-6968.2011.02339.x](https://doi.org/10.1111/j.1574-6968.2011.02339.x).
- Hocq, R., Jagtap, S., Boutard, M., Tolonen, A. C., Duval, L., Pirayre, A., Lopes Ferreira, N., and Wasels, F. (2022). "Genome-Wide TSS Distribution in Three Related Clostridia with Normalized Capp-Switch Sequencing". In: *Microbiology Spectrum* vol. 10, no. 2, e02288–21. DOI: [10.1128/spectrum.02288-21](https://doi.org/10.1128/spectrum.02288-21).

- Howden, B. P., Giulieri, S. G., Wong Fok Lung, T., Baines, S. L., Sharkey, L. K., Lee, J. Y. H., Hachani, A., Monk, I. R., and Stinear, T. P. (2023). “Staphylococcus Aureus Host Interactions and Adaptation”. In: *Nature Reviews Microbiology* vol. 21, no. 6, pp. 380–395. ISSN: 1740-1534. DOI: [10.1038/s41579-023-00852-y](https://doi.org/10.1038/s41579-023-00852-y).
- Innocenti, N., Golumbeanu, M., d’Hérouël, A. F., Lacoux, C., Bonnin, R. A., Kennedy, S. P., Wessner, F., Serror, P., Bouloc, P., Repoila, F., *et al.* (2015). “Whole-Genome Mapping of 5’ RNA Ends in Bacteria by Tagged Sequencing: A Comprehensive View in *Enterococcus faecalis*”. In: *RNA* vol. 21, no. 5, pp. 1018–1030. DOI: [10.1261/rna.048470.114](https://doi.org/10.1261/rna.048470.114).
- Jalili, V., Afgan, E., Gu, Q., Clements, D., Blankenberg, D., Goecks, J., Taylor, J., and Nekrutenko, A. (2020). “The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2020 Update”. In: *Nucleic Acids Research* vol. 48, no. W1, W395–W402. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkaa434](https://doi.org/10.1093/nar/gkaa434).
- Jeong, Y., Kim, J.-N., Kim, M. W., Bucca, G., Cho, S., Yoon, Y. J., Kim, B.-G., Roe, J.-H., Kim, S. C., Smith, C. P., and Cho, B.-K. (2016). “The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2)”. In: *Nature Communications* vol. 7, no. 1, p. 11605. ISSN: 2041-1723. DOI: [10.1038/ncomms11605](https://doi.org/10.1038/ncomms11605).
- Joppich, M. and Zimmer, R. (2019). “From Command-Line Bioinformatics to bioGUI”. In: *PeerJ* vol. 7, e8111. ISSN: 2167-8359. DOI: [10.7717/peerj.8111](https://doi.org/10.7717/peerj.8111).
- Jorjani, H. and Zavolan, M. (2013). “TSSer: An Automated Method to Identify Transcription Start Sites in Prokaryotic Genomes from Differential RNA Sequencing Data”. In: *Bioinformatics* vol. 30, no. 7, pp. 971–974. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btt752](https://doi.org/10.1093/bioinformatics/btt752).
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). “Highly Accurate Protein Structure Prediction with AlphaFold”. In: *Nature* vol. 596, no. 7873, pp. 583–589. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2).
- Khaledi, M., Khatami, M., Hemmati, J., Bakhti, S., Hoseini, S. A., and Ghahramanpour, H. (2025). “Role of Small Non-Coding RNA in Gram-Negative Bacteria: New Insights and Comprehensive Review of Mechanisms, Functions, and Potential Applications”. In: *Molecular Biotechnology* vol. 67, no. 8, pp. 2978–2996. ISSN: 1073-6085, 1559-0305. DOI: [10.1007/s12033-024-01248-w](https://doi.org/10.1007/s12033-024-01248-w).
- Kingsford, C. L., Ayanbule, K., and Salzberg, S. L. (2007). “Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake”. In: *Genome Biology* vol. 8, no. 2, R22. ISSN: 1474-760X. DOI: [10.1186/gb-2007-8-2-r22](https://doi.org/10.1186/gb-2007-8-2-r22). URL: <https://doi.org/10.1186/gb-2007-8-2-r22>.
- Ko, D. and Choi, S. H. (2021). “Comparative Genomics Reveals an SNP Potentially Leading to Phenotypic Diversity of Salmonella Enterica Serovar Enteritidis”. In: *Microbial Genomics* vol. 7, no. 5, p. 000572. ISSN: 2057-5858. DOI: [10.1099/mgen.0.000572](https://doi.org/10.1099/mgen.0.000572).
- Koonin, E. V. (2001). “Computational Genomics”. In: *Current Biology* vol. 11, no. 5, R155–R158. ISSN: 0960-9822. DOI: [10.1016/S0960-9822\(01\)00081-1](https://doi.org/10.1016/S0960-9822(01)00081-1).
- Kraemer, S. M., Duckworth, O. W., Harrington, J. M., and Schenkeveld, W. D. C. (2015). “Metallophores and Trace Metal Biogeochemistry”. In: *Aquatic Geochemistry* vol. 21, no. 2, pp. 159–195. ISSN: 1573-1421. DOI: [10.1007/s10498-014-9246-7](https://doi.org/10.1007/s10498-014-9246-7).

- Kramer, J., Özkaya, Ö., and Kümmerli, R. (2020). “Bacterial Siderophores in Community and Host Interactions”. In: *Nature Reviews Microbiology* vol. 18, no. 3, pp. 152–163. ISSN: 1740-1526. DOI: [10.1038/s41579-019-0284-4](https://doi.org/10.1038/s41579-019-0284-4).
- Kumar, N., Miyajima, F., He, M., Roberts, P., Swale, A., Ellison, L., Pickard, D., Smith, G., Molyneux, R., Dougan, G., Parkhill, J., Wren, B. W., Parry, C. M., Pirmohamed, M., and Lawley, T. D. (2016). “Genome-Based Infection Tracking Reveals Dynamics of *Clostridium Difficile* Transmission and Disease Recurrence”. In: *Clinical Infectious Diseases* vol. 62, no. 6, pp. 746–752. ISSN: 1058-4838, 1537-6591. DOI: [10.1093/cid/civ1031](https://doi.org/10.1093/cid/civ1031).
- Lee, J. H. and Holmes, R. K. (2000). “Characterization of Specific Nucleotide Substitutions in DtxR-specific Operators of *Corynebacterium diphtheriae* That Dramatically Affect DtxR Binding, Operator Function, and Promoter Strength”. In: *Journal of Bacteriology* vol. 182, no. 2, pp. 432–438. ISSN: 0021-9193. DOI: [10.1128/jb.182.2.432-438.2000](https://doi.org/10.1128/jb.182.2.432-438.2000).
- Letunic, I. and Bork, P. (2019). “Interactive Tree Of Life (iTOL) v4: recent updates and new developments”. In: *Nucleic Acids Research* vol. 47, no. W1, W256–W259.
- Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. (2014). “UpSet: Visualization of Intersecting Sets”. In: *IEEE Transactions on Visualization and Computer Graphics* vol. 20, no. 12, pp. 1983–1992. ISSN: 1077-2626. DOI: [10.1109/TVCG.2014.2346248](https://doi.org/10.1109/TVCG.2014.2346248).
- Liaw, T. S. and Gehlenborg, N. (2022). “User-Centric Process of Designing a Molecular & Cellular Query Interface for Biomedical Research”. In: *Proceedings of the Design Society* vol. 2, pp. 221–230. ISSN: 2732-527X. DOI: [10.1017/pds.2022.24](https://doi.org/10.1017/pds.2022.24).
- Liu, W., Wang, Z., You, R., Xie, C., Wei, H., Xiong, Y., Yang, J., and Zhu, S. (2024). “PLMSearch: Protein Language Model Powers Accurate and Fast Sequence Search for Remote Homology”. In: *Nature Communications* vol. 15, no. 1, p. 2775. ISSN: 2041-1723. DOI: [10.1038/s41467-024-46808-5](https://doi.org/10.1038/s41467-024-46808-5).
- Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). “ViennaRNA Package 2.0”. In: *Algorithms for Molecular Biology* vol. 6, no. 1, p. 26. ISSN: 1748-7188. DOI: [10.1186/1748-7188-6-26](https://doi.org/10.1186/1748-7188-6-26).
- Love, M. I., Huber, W., and Anders, S. (2014). “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2”. In: *Genome Biology* vol. 15, no. 12, p. 550. ISSN: 1474-760X. DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). “Transcriptomics Technologies”. In: *PLOS Computational Biology* vol. 13, no. 5, e1005457. ISSN: 1553-7358. DOI: [10.1371/journal.pcbi.1005457](https://doi.org/10.1371/journal.pcbi.1005457).
- LYi, S., Wang, Q., Lekschas, F., and Gehlenborg, N. (2022). “Gosling: A Grammar-Based Toolkit for Scalable and Interactive Genomics Data Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* vol. 28, no. 1, pp. 140–150. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2021.3114876](https://doi.org/10.1109/TVCG.2021.3114876).
- Maier, L., Stein-Thoeringer, C., Ley, R. E., Brötz-Oesterhelt, H., Link, H., Ziemert, N., Wagner, S., and Peschel, A. (2024). “Integrating Research on Bacterial Pathogens and Commensals to Fight Infections—an Ecological Perspective”. In: *The Lancet Microbe* vol. 5, no. 8. ISSN: 2666-5247. DOI: [10.1016/S2666-5247\(24\)00049-1](https://doi.org/10.1016/S2666-5247(24)00049-1).
- Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., Hill, B., Brito, J., Littman, R. J., Statz, B., Lam, A. K.-M., Dayama, G., Grieneisen, L., Martin, L. S., Flint, J., Eskin, E., and Blekhman, R. (2019). “Challenges and Recommendations to Improve the Installability and Archival Stability of Omics Computational Tools”. In: *PLOS Biology* vol. 17, no. 6, e3000333. ISSN: 1545-7885. DOI: [10.1371/journal.pbio.3000333](https://doi.org/10.1371/journal.pbio.3000333).

- Marcos-Torres, F. J., Maurer, D., Juniar, L., and Griesse, J. J. (2021). “The Bacterial Iron Sensor IdeR Recognizes Its DNA Targets by Indirect Readout”. In: *Nucleic Acids Research* vol. 49, no. 17, pp. 10120–10135. ISSN: 0305-1048. DOI: [10.1093/nar/gkab711](https://doi.org/10.1093/nar/gkab711).
- Medema, M. H., Cimermancic, P., Sali, A., Takano, E., and Fischbach, M. A. (2014). “A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis”. In: *PLOS Computational Biology* vol. 10, no. 12, e1004016. ISSN: 1553-7358. DOI: [10.1371/journal.pcbi.1004016](https://doi.org/10.1371/journal.pcbi.1004016).
- Messner, C. B., Demichev, V., Wang, Z., Hartl, J., Kustatscher, G., Mülleder, M., and Ralser, M. (2023). “Mass Spectrometry-Based High-Throughput Proteomics and Its Role in Biomedical Studies and Systems Biology”. In: *PROTEOMICS* vol. 23, no. 7-8, p. 2200013. ISSN: 1615-9861. DOI: [10.1002/pm.202200013](https://doi.org/10.1002/pm.202200013).
- Miller, W. B., Baluška, F., and Reber, A. S. (2023). “A Revised Central Dogma for the 21st Century: All Biology Is Cognitive Information Processing”. In: *Progress in Biophysics and Molecular Biology* vol. 182, pp. 34–48. ISSN: 0079-6107. DOI: [10.1016/j.pbiomolbio.2023.05.005](https://doi.org/10.1016/j.pbiomolbio.2023.05.005).
- Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). “ColabFold: Making Protein Folding Accessible to All”. In: *Nature Methods* vol. 19, no. 6, pp. 679–682. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/s41592-022-01488-1](https://doi.org/10.1038/s41592-022-01488-1).
- Mukherjee, S. and Bassler, B. L. (2019). “Bacterial Quorum Sensing in Complex and Dynamically Changing Environments”. In: *Nature Reviews Microbiology* vol. 17, no. 6, pp. 371–382. ISSN: 1740-1534. DOI: [10.1038/s41579-019-0186-5](https://doi.org/10.1038/s41579-019-0186-5).
- Munzner, T. (2014). *Visualization analysis and design: Abstraction, Principles and Methods*. CRC press.
- Nachtergaele, S. and He, C. (2017). “The Emerging Biology of RNA Post-Transcriptional Modifications”. In: *RNA Biology* vol. 14, no. 2, pp. 156–163. ISSN: 1547-6286. DOI: [10.1080/15476286.2016.1267096](https://doi.org/10.1080/15476286.2016.1267096).
- Navarro-Muñoz, J. C., Selem-Mojica, N., Mallowney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., and Medema, M. H. (2020). “A Computational Framework to Explore Large-Scale Biosynthetic Diversity”. In: *Nature chemical biology* vol. 16, no. 1, pp. 60–68. ISSN: 1552-4450. DOI: [10.1038/s41589-019-0400-9](https://doi.org/10.1038/s41589-019-0400-9).
- Neu, H. C. (1992). “The Crisis in Antibiotic Resistance”. In: *Science* vol. 257, no. 5073, pp. 1064–1073. DOI: [10.1126/science.257.5073.1064](https://doi.org/10.1126/science.257.5073.1064).
- Neumann, W., Gulati, A., and Nolan, E. M. (2017). “Metal Homeostasis in Infectious Disease: Recent Advances in Bacterial Metallophores and the Human Metal-Withholding Response”. In: *Current Opinion in Chemical Biology* vol. 37, pp. 10–18. ISSN: 1879-0402. DOI: [10.1016/j.cbpa.2016.09.012](https://doi.org/10.1016/j.cbpa.2016.09.012).
- Okeke, I. N., Kraker, M. E. A. de, Boeckel, T. P. V., Kumar, C. K., Schmitt, H., Gales, A. C., Bertagnolio, S., Sharland, M., and Laxminarayan, R. (2024). “The Scope of the Antimicrobial Resistance Challenge”. In: *The Lancet* vol. 403, no. 10442, pp. 2426–2438. ISSN: 0140-6736, 1474-547X. DOI: [10.1016/S0140-6736\(24\)00876-6](https://doi.org/10.1016/S0140-6736(24)00876-6). PMID: [38797176](https://pubmed.ncbi.nlm.nih.gov/38797176/).
- Osborn, A. E. and Field, B. (2009). “Operons”. In: *Cellular and Molecular Life Sciences* vol. 66, no. 23, pp. 3755–3775. ISSN: 1420-9071. DOI: [10.1007/s00018-009-0114-3](https://doi.org/10.1007/s00018-009-0114-3).
- Parise, M. T. D., Parise, D., Kato, R. B., Pauling, J. K., Tauch, A., Azevedo, V. A. d. C., and Baumbach, J. (2020). “CoryneRegNet 7, the Reference Database and Analysis Platform for Corynebacterial Gene Regulatory Networks”. In: *Scientific Data* vol. 7, no. 1, p. 142. ISSN: 2052-4463. DOI: [10.1038/s41597-020-0484-9](https://doi.org/10.1038/s41597-020-0484-9).

- Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2022). “GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy”. In: *Nucleic Acids Research* vol. 50, no. D1, pp. D785–D794. ISSN: 0305-1048. DOI: [10.1093/nar/gkab776](https://doi.org/10.1093/nar/gkab776).
- Reitz, Z. L. and Medema, M. H. (2022). “Genome Mining Strategies for Metallophore Discovery”. In: *Current Opinion in Biotechnology* vol. 77, p. 102757. ISSN: 0958-1669. DOI: [10.1016/j.copbio.2022.102757](https://doi.org/10.1016/j.copbio.2022.102757).
- Revell, L. J., Schliep, K., Valderrama, E., and Richardson, J. E. (2018). “Graphs in Phylogenetic Comparative Analysis: Anscombe’s Quartet Revisited”. In: *Methods in Ecology and Evolution* vol. 9, no. 10, pp. 2145–2154. ISSN: 2041-210X. DOI: [10.1111/2041-210X.13067](https://doi.org/10.1111/2041-210X.13067).
- Rivas, E. and Eddy, S. R. (2001). “Noncoding RNA gene detection using comparative sequence analysis”. In: *BMC Bioinformatics* vol. 2, no. 1, p. 8. ISSN: 1471-2105. DOI: [10.1186/1471-2105-2-8](https://doi.org/10.1186/1471-2105-2-8). URL: <https://doi.org/10.1186/1471-2105-2-8>.
- Rogozin, I. B. (2004). “Computational Approaches for the Analysis of Gene Neighbourhoods in Prokaryotic Genomes”. In: *Briefings in Bioinformatics* vol. 5, no. 2, pp. 131–149. ISSN: 1467-5463, 1477-4054. DOI: [10.1093/bib/5.2.131](https://doi.org/10.1093/bib/5.2.131).
- Ryan, D., Bornet, E., Prezsa, G., Alampalli, S. V., Franco de Carvalho, T., Felchle, H., Ebbecke, T., Hayward, R. J., Deutschbauer, A. M., Barquist, L., and Westermann, A. J. (2024). “An expanded transcriptome atlas for *Bacteroides thetaiotaomicron* reveals a small RNA that modulates tetracycline sensitivity”. In: *Nature Microbiology* vol. 9, no. 4, pp. 1130–1144. ISSN: 2058-5276. DOI: [10.1038/s41564-024-01642-9](https://doi.org/10.1038/s41564-024-01642-9).
- Ryan, D., Jenniches, L., Reichardt, S., Barquist, L., and Westermann, A. J. (2020). “A High-Resolution Transcriptome Map Identifies Small RNA Regulation of Metabolism in the Gut Microbe *Bacteroides thetaiotaomicron*”. In: *Nature communications* vol. 11, no. 1, p. 3557. DOI: [10.1038/s41467-020-17348-5](https://doi.org/10.1038/s41467-020-17348-5).
- Schuenemann, V. J., Avanzi, C., Krause-Kyora, B., Seitz, A., Herbig, A., Inskip, S., Bonazzi, M., Reiter, E., Urban, C., Dangvard Pedersen, D., *et al.* (2018). “Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe”. In: *PLOS Pathogens* vol. 14, no. 5, e1006997.
- Sebulsky, M. T., Hohnstein, D., Hunter, M. D., and Heinrichs, D. E. (2000). “Identification and Characterization of a Membrane Permease Involved in Iron-Hydroxamate Transport in *Staphylococcus aureus*”. In: *Journal of Bacteriology* vol. 182, no. 16, pp. 4394–4400. ISSN: 0021-9193. DOI: [10.1128/JB.182.16.4394-4400.2000](https://doi.org/10.1128/JB.182.16.4394-4400.2000).
- Sebulsky, M. T., Speziali, C. D., Shilton, B. H., Edgell, D. R., and Heinrichs, D. E. (2004). “FhuD1, a Ferric Hydroxamate-Binding Lipoprotein in *Staphylococcus aureus*: A Case of Gene Duplication and Lateral Transfer”. In: *The Journal of Biological Chemistry* vol. 279, no. 51, pp. 53152–53159. ISSN: 0021-9258. DOI: [10.1074/jbc.M409793200](https://doi.org/10.1074/jbc.M409793200).
- Shahreen, N., Shahid, S. A., Subhani, M., Al-Siyabi, A., and Saha, R. (2025). “Minimal Gene Signatures Enable High-Accuracy Prediction of Antibiotic Resistance in *Pseudomonas Aeruginosa*”. In: *npj Systems Biology and Applications* vol. 11, no. 1, p. 108. ISSN: 2056-7189. DOI: [10.1038/s41540-025-00584-0](https://doi.org/10.1038/s41540-025-00584-0).
- Shapiro, J. A. (2009). “Revisiting the Central Dogma in the 21st Century”. In: *Annals of the New York Academy of Sciences* vol. 1178, no. 1, pp. 6–28. ISSN: 1749-6632. DOI: [10.1111/j.1749-6632.2009.04990.x](https://doi.org/10.1111/j.1749-6632.2009.04990.x).
- Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010).

- “The Primary Transcriptome of the Major Human Pathogen *Helicobacter pylori*”. In: *Nature* vol. 464, no. 7286, pp. 250–255. ISSN: 1476-4687. DOI: [10.1038/nature08756](https://doi.org/10.1038/nature08756).
- Sheldon, J. and Heinrichs, D. (2012). “The Iron-Regulated Staphylococcal Lipoproteins”. In: *Frontiers in Cellular and Infection Microbiology* vol. 2. ISSN: 2235-2988. DOI: [10.3389/fcimb.2012.00041](https://doi.org/10.3389/fcimb.2012.00041).
- Shneiderman, B. (1996). “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”. In: *Proceedings of IEEE Symposium on Visual Languages*. Vol. 96.
- Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M., and Carriço, J. A. (2018). “chewBBACA: A Complete Suite for Gene-by-Gene Schema Creation and Strain Identification”. In: *Microbial Genomics* vol. 4, no. 3, e000166. ISSN: 2057-5858. DOI: [10.1099/mgen.0.000166](https://doi.org/10.1099/mgen.0.000166).
- Singh, I., Contreras, A., Cordero, J., Rubio, K., Dobersch, S., Günther, S., Jeratsch, S., Mehta, A., Krüger, M., Graumann, J., Seeger, W., Dobrev, G., Braun, T., and Barreto, G. (2018). “MiCEE Is a ncRNA-protein Complex That Mediates Epigenetic Silencing and Nucleolar Organization”. In: *Nature Genetics* vol. 50, no. 7, pp. 990–1001. ISSN: 1546-1718. DOI: [10.1038/s41588-018-0139-3](https://doi.org/10.1038/s41588-018-0139-3).
- Skurnik, M. (2022). “Can Bacteriophages Replace Antibiotics?” In: *Antibiotics* vol. 11, no. 5, p. 575. ISSN: 2079-6382. DOI: [10.3390/antibiotics11050575](https://doi.org/10.3390/antibiotics11050575).
- Sorek, R. and Cossart, P. (2010). “Prokaryotic Transcriptomics: A New View on Regulation, Physiology and Pathogenicity”. In: *Nature Reviews Genetics* vol. 11, no. 1, pp. 9–16. DOI: [10.1038/nrg2695](https://doi.org/10.1038/nrg2695).
- Soutourina, O., Dubois, T., Monot, M., Shelyakin, P. V., Saujet, L., Boudry, P., Gelfand, M. S., Dupuy, B., and Martin-Verstraete, I. (2020). “Genome-Wide Transcription Start Site Mapping and Promoter Assignments to a Sigma Factor in the Human Enteropathogen *Clostridioides Difficile*”. In: *Frontiers in Microbiology* vol. 11. ISSN: 1664-302X. DOI: [10.3389/fmicb.2020.01939](https://doi.org/10.3389/fmicb.2020.01939).
- Spratt, B. G. (1994). “Resistance to Antibiotics Mediated by Target Alterations”. In: *Science* vol. 264, no. 5157, pp. 388–393. DOI: [10.1126/science.8153626](https://doi.org/10.1126/science.8153626).
- Stazic, D. and Voß, B. (2016). “The Complexity of Bacterial Transcriptomes”. In: *Journal of Biotechnology* vol. 232, pp. 69–78. ISSN: 0168-1656. DOI: [10.1016/j.jbiotec.2015.09.041](https://doi.org/10.1016/j.jbiotec.2015.09.041).
- Stein, L. (2001). “Genome Annotation: From Sequence to Biology”. In: *Nature Reviews Genetics* vol. 2, no. 7, pp. 493–503. ISSN: 1471-0064. DOI: [10.1038/35080529](https://doi.org/10.1038/35080529).
- Stubberfield, E., Sheldon, J., Card, R. M., AbuOun, M., Rogers, J., Williamson, S., Kay, G. L., Pallen, M. J., and Anjum, M. F. (2021). “Whole-Genome Sequencing of *Brachyspira Hyodysenteriae* Isolates From England and Wales Reveals Similarities to European Isolates and Mutations Associated With Reduced Sensitivity to Antimicrobials”. In: *Frontiers in Microbiology* vol. 12, p. 713233. ISSN: 1664-302X. DOI: [10.3389/fmicb.2021.713233](https://doi.org/10.3389/fmicb.2021.713233).
- Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W.-H. (2019). “Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees”. In: *Nucleic Acids Research* vol. 47, no. W1, W270–W275.
- Tam, C. and Iwasaki, W. (2023). “AlphaCutter: Efficient Removal of Non-Globular Regions from Predicted Protein Structures”. In: *Proteomics* vol. 23, no. 16, p. 2300176. ISSN: 1615-9861. DOI: [10.1002/pmic.202300176](https://doi.org/10.1002/pmic.202300176).
- Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2022). “SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models”. In: *Nature Biotechnology* vol. 40, no. 7, pp. 1023–1025. ISSN: 1087-0156. DOI: [10.1038/s41587-021-01156-3](https://doi.org/10.1038/s41587-021-01156-3).

- Thieffry, D. and Sarkar, S. (1998). “Forty Years under the Central Dogma”. In: *Trends in Biochemical Sciences* vol. 23, no. 8, pp. 312–316. ISSN: 0968-0004. DOI: [10.1016/S0968-0004\(98\)01244-4](https://doi.org/10.1016/S0968-0004(98)01244-4). PMID: [9757833](https://pubmed.ncbi.nlm.nih.gov/9757833/).
- Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). “Integrative Genomics Viewer (IGV): High-performance Genomics Data Visualization and Exploration”. In: *Briefings in Bioinformatics* vol. 14, no. 2, pp. 178–192. ISSN: 1467-5463. DOI: [10.1093/bib/bbs017](https://doi.org/10.1093/bib/bbs017).
- Tognon, M., Köhler, T., Luscher, A., and Van Delden, C. (2019). “Transcriptional Profiling of *Pseudomonas Aeruginosa* and *Staphylococcus Aureus* during in Vitro Co-Culture”. In: *BMC Genomics* vol. 20, no. 1, p. 30. ISSN: 1471-2164. DOI: [10.1186/s12864-018-5398-y](https://doi.org/10.1186/s12864-018-5398-y).
- Al-Tohamy, A. and Grove, A. (2025). “Targeting Bacterial Transcription Factors for Infection Control: Opportunities and Challenges”. In: *Transcription* vol. 16, no. 1, pp. 141–168. ISSN: 2154-1264, 2154-1272. DOI: [10.1080/21541264.2023.2293523](https://doi.org/10.1080/21541264.2023.2293523).
- Towbin, H., Staehelin, T., and Gordon, J. (1979). “Electrophoretic Transfer of Proteins from Polyacrylamide Gels to Nitrocellulose Sheets: Procedure and Some Applications.” In: *Proceedings of the National Academy of Sciences* vol. 76, no. 9, pp. 4350–4354. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.76.9.4350](https://doi.org/10.1073/pnas.76.9.4350).
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). “Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment”. In: *Nature* vol. 428, no. 6978, pp. 37–43. ISSN: 0028-0836, 1476-4687. DOI: [10.1038/nature02340](https://doi.org/10.1038/nature02340).
- van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2024). “Fast and Accurate Protein Structure Search with Foldseek”. In: *Nature Biotechnology* vol. 42, no. 2, pp. 243–246. ISSN: 1546-1696. DOI: [10.1038/s41587-023-01773-0](https://doi.org/10.1038/s41587-023-01773-0).
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., *et al.* (2022). “AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models”. In: *Nucleic Acids Research* vol. 50, no. D1, pp. D439–D444. DOI: [10.1093/nar/gkab1061](https://doi.org/10.1093/nar/gkab1061).
- Vignal, A., Milan, D., SanCristobal, M., and Eggen, A. (2002). “A Review on SNP and Other Types of Molecular Markers and Their Use in Animal Genetics”. In: *Genetics Selection Evolution* vol. 34, no. 3, pp. 275–305. ISSN: 0999-193X, 1297-9686. DOI: [10.1051/gse:2002009](https://doi.org/10.1051/gse:2002009).
- Vockenhuber, M.-P., Sharma, C. M., Statt, M. G., Schmidt, D., Xu, Z., Dietrich, S., Liesegang, H., Mathews, D. H., and Suess, B. (2011). “Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*”. In: *RNA biology* vol. 8, no. 3, pp. 468–477. ISSN: 1547-6286. DOI: [10.4161/rna.8.3.14421](https://doi.org/10.4161/rna.8.3.14421).
- Weinberg, E. D. (2000). “Microbial Pathogens with Impaired Ability to Acquire Host Iron”. In: *Biomaterials: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine* vol. 13, no. 1, pp. 85–89. ISSN: 0966-0844. DOI: [10.1023/a:1009293500209](https://doi.org/10.1023/a:1009293500209).
- Wells, J. N. and Marsh, J. A. (2018). “Experimental Characterization of Protein Complex Structure, Dynamics, and Assembly”. In: *Protein Complex Assembly: Methods and Protocols*. Ed. by Marsh, J. A. New York, NY: Springer, pp. 3–27. ISBN: 978-1-4939-7759-8. DOI: [10.1007/978-1-4939-7759-8_1](https://doi.org/10.1007/978-1-4939-7759-8_1).

- Westermann, A. J., Gorski, S. A., and Vogel, J. (2012). “Dual RNA-seq of Pathogen and Host”. In: *Nature Reviews Microbiology* vol. 10, no. 9, pp. 618–630. ISSN: 1740-1526, 1740-1534. DOI: [10.1038/nrmicro2852](https://doi.org/10.1038/nrmicro2852).
- Witte Paz, M.**, Bitzer, A., Nieselt, K., and Heilbronner, S. (2025). “A Landscape of Metallophore Synthesis and Uptake Potential of the Genus *Staphylococcus*”. In: *NAR Genomics and Bioinformatics* vol. 7, no. 4, lqaf183. DOI: [10.1093/nargab/lqaf183](https://doi.org/10.1093/nargab/lqaf183).
- Witte Paz, M.**, Harbig, T. A., and Nieselt, K. (2022). “Evidente—a Visual Analytics Tool for Data Enrichment in SNP-based Phylogenetic Trees”. In: *Bioinformatics Advances* vol. 2, no. 1, vbac075. DOI: [10.1093/bioadv/vbac075](https://doi.org/10.1093/bioadv/vbac075).
- Witte Paz, M.**, Herbig, A., and Nieselt, K. (2026). “TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration”. In: *PLOS ONE* vol. 21, no. 3, e0326299. DOI: [10.1371/journal.pone.0326299](https://doi.org/10.1371/journal.pone.0326299).
- Witte Paz, M.**, Vogel, T., and Nieselt, K. (2024). “TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA Transcripts”. In: *NAR Genomics and Bioinformatics* vol. 6, no. 4, lqae168. DOI: [10.1093/nargab/lqae168](https://doi.org/10.1093/nargab/lqae168).
- Wolański, M., Krawiec, M., Nieselt, K., Schwarz, T., Dere, D., Krismer, B., Cano-Prieto, C., Gross, H., and Zakrzewska-Czerwińska, J. (2025). “Multiple Regulators Control the Biosynthesis of Brasilicardin in *Nocardia Terpenica*”. In: *Applied Microbiology and Biotechnology* vol. 109, no. 1, p. 150. ISSN: 1432-0614. DOI: [10.1007/s00253-025-13485-3](https://doi.org/10.1007/s00253-025-13485-3).
- Wolfram-Schauerte, M., Moskalchuk, A., Pozhydaieva, N., Rojas, A. A. R., Schindler, D., Kaiser, S., Pazcia, N., and Höfer, K. (2024). “T4 Phage RNA Is NAD-capped and Alters the NAD-cap Epitranscriptome of *Escherichia coli* during Infection through a Phage-Encoded Decapping Enzyme”. In: *bioRxiv*, pp. 2024–04. DOI: [10.1101/2024.04.04.588121](https://doi.org/10.1101/2024.04.04.588121).
- Wright, G. D. (2005). “Bacterial Resistance to Antibiotics: Enzymatic Degradation and Modification”. In: *Advanced Drug Delivery Reviews*. Mechanisms of Antimicrobial Resistance: Opportunities for New Targeted Therapies vol. 57, no. 10, pp. 1451–1470. ISSN: 0169-409X. DOI: [10.1016/j.addr.2005.04.002](https://doi.org/10.1016/j.addr.2005.04.002).
- Yamano, Y. (2019). “In Vitro Activity of Cefiderocol Against a Broad Range of Clinically Important Gram-negative Bacteria”. In: *Clinical Infectious Diseases* vol. 69, no. Supplement_7, S544–S551. ISSN: 1058-4838. DOI: [10.1093/cid/ciz827](https://doi.org/10.1093/cid/ciz827).
- Zackular, J. P., Chazin, W. J., and Skaar, E. P. (2015). “Nutritional Immunity: S100 Proteins at the Host-Pathogen Interface”. In: *The Journal of Biological Chemistry* vol. 290, no. 31, pp. 18991–18998. ISSN: 1083-351X. DOI: [10.1074/jbc.R115.645085](https://doi.org/10.1074/jbc.R115.645085).
- Zhang, W., Li, F., and Nie, L. (2010). “Integrating Multiple ‘Omics’ Analysis for Microbial Biology: Application and Methodologies”. In: *Microbiology* vol. 156, no. 2, pp. 287–301. ISSN: 1465-2080. DOI: [10.1099/mic.0.034793-0](https://doi.org/10.1099/mic.0.034793-0).
- Zhang, Z., Wu, S., Stenoien, D. L., and Paša-Tolić, L. (2014). “High-Throughput Proteomics”. In: *Annual Review of Analytical Chemistry* vol. 7, no. Volume 7, 2014, pp. 427–454. ISSN: 1936-1327, 1936-1335. DOI: [10.1146/annurev-anchem-071213-020216](https://doi.org/10.1146/annurev-anchem-071213-020216).
- Zhao, Y., Bitzer, A., Power, J. J., Belikova, D., Torres Salazar, B. O., Adolf, L. A., Gerlach, D., Krismer, B., and Heilbronner, S. (2024). “Nasal Commensals Reduce *Staphylococcus aureus* Proliferation by Restricting Siderophore Availability”. In: *The ISME Journal* vol. 18, no. 1, wrae123. DOI: [10.1093/ismejo/wrae123](https://doi.org/10.1093/ismejo/wrae123).
- Zhukova, A., Fernandes, L. G., Hugon, P., Pappas, C. J., Sismeiro, O., Coppée, J.-Y., Becavin, C., Malabat, C., Eshghi, A., Zhang, J.-J., Yang, F. X., and Picardeau, M. (2017). “Genome-Wide Transcriptional Start Site Mapping and sRNA Identification in the Pathogen *Leptospira interrogans*”. In: *Frontiers in Cellular and Infection Microbiology* vol. 7. ISSN: 2235-2988. DOI: [10.3389/fcimb.2017.00010](https://doi.org/10.3389/fcimb.2017.00010).

Appendices

Appendix I

Evidente – A visual analytics tool for data enrichment in SNP-based phylogenetic trees

Mathias Witte Paz, Theresa Anisja Harbig, Kay Nieselt

Published in *Bioinformatic Advances*, Volume 2, Issue 1, October 2022, vbac075, DOI: [10.1093/bioadv/vbac075](https://doi.org/10.1093/bioadv/vbac075).

Abstract

Motivation: A common practice in the analysis of pathogens and their strains is using single nucleotide polymorphisms (SNPs) to reconstruct their evolutionary history. However, genome-wide SNP-based phylogenetic trees are rarely analyzed without any further information. Including the underlying SNP data together with further metadata on the respective samples in the exploration process can facilitate linking the genomic and phenotypic properties of the samples.

Results: We introduce **Evidente**, a web-application that provides an interactive visual analysis interface for the simultaneous interrogation of phylogenetic relationships, genome-wide SNP data and metadata for samples of an organism. Besides visualizing the phylogenetic tree, **Evidente** classifies SNPs as supporting or non-supporting of the tree structures and shows the distribution of both types of SNPs among samples and clades of interest. Furthermore, additional metadata can be included in the visualization. Lastly, **Evidente** includes an enrichment analysis to identify over-represented genomic features encoded by GO-terms within the clades of the tree. We demonstrate the usability of **Evidente** with the data of the pathogens *Treponema pallidum* and *Mycobacterium leprae*.

Availability: **Evidente** is available at the TueVis visualization web-server at <https://evidente-tuevis.cs.uni-tuebingen.de/>, it can also be run locally.

Keywords: Visualization, Phylogenetics, SNPs, Web services, Comparative genomics

Introduction

The rise of next-generation sequencing (NGS) techniques allows faster and more efficient procedures for the sequencing of genomes (Goodwin *et al.* 2016). This has not only lead to the sequencing of thousands of different species across all kingdoms, but also to the deciphering of the variability of genomes within a species. In particular the massive sequencing of microbial and viral genomes now offer exciting abilities to query the variability of different strains within a bacterial or viral species. Many fields, in particular that of molecular epidemiology but also the field of ancient pathogenomics (see Spyrou *et al.* 2019 for a review), have made enormous progress and created amazing insight in the evolutionary history of human pathogens. NGS techniques in particular have enabled the study of point mutations called *single-nucleotide polymorphisms* or *variants* depending on the frequencies in the populations (SNPs/SNVs, from this point on SNPs) in strains within a species. SNPs are then often used for example to reconstruct the evolutionary relationship of the individuals. For this, a common procedure is to first map the NGS data to a common reference genome, followed by the SNP-calling step using a genotyping method. An alternative approach is to align sample genomes with a reference genome and to identify all positions where a sample differs by a mutation from the reference with respect to the coordinate system of the reference genome. The detected SNPs of all samples can be summarized into one *SNP-table* (see Figure I.1a for an example). A SNP-table, and hence the underlying identified SNPs, can then be used as a basis for the evolutionary reconstruction to compute a SNP-based phylogenetic tree using, for example, a maximum parsimony approach.

However, the evolutionary reconstruction does not provide a direct relationship between the reconstructed phylogenetic tree and the underlying SNPs. One way of linking these two data sets and also providing a measurement of accuracy of the tree is to compute the clade-specificity of the SNPs. A clade-specific or clade-unique SNP appears only and in all samples of the clade's leaf-nodes. Hence, clade-specific SNPs support the choice for the reconstructed phylogeny and therefore can be called supporting SNPs. An example for a clade-specific, i.e., supporting SNP, is the SNP found at position 1 in Figure I.1b. If a SNP appears in many clades, we refer to it as a non-supporting SNP. An example for a non-supporting SNP is the SNP found at position 2 in Figure I.1b.

SNP-based phylogenetic trees are often used for the analysis of pathogens and their strains (Filliol *et al.* 2006; Foster *et al.* 2009; Pandya *et al.* 2009). For example, SNP-based phylogenies were used in the analysis of the spread of the pathogen *Mycobacterium leprae* (Schuenemann *et al.* 2018) as well as in the analysis of the rise of antibiotic resistance of *Treponema pallidum* (Arora *et al.* 2016). In these studies, the analysis required the inclusion of the identified SNP-data as well as further metadata (e.g. the origin place of the sample or its antibiotic resistance) to identify significant correlations among the data sets. To simultaneously interact with the phylogenetic data, the identified genome-wide SNPs and further metadata within one visualisation interface can help in identifying possible correlations. Furthermore, the underlying genomic data is often analyzed from the functional perspective or to classify, for example, microbial samples after disease outbreaks. A prominent example for this is Nextstrain (Hadfield *et al.* 2018), a visual analytics platform to track the evolution of viruses. It integrates genome sequence data with phylogenetic reconstruction of the strains and metadata such as geographic origin or host species. Nextstrain has been instrumental in understanding the evolution and spread of SARS-CoV-2, the virus that has caused the COVID-19 pandemic.

There are many pipelines for the phylogenetic reconstruction from SNP data, such as

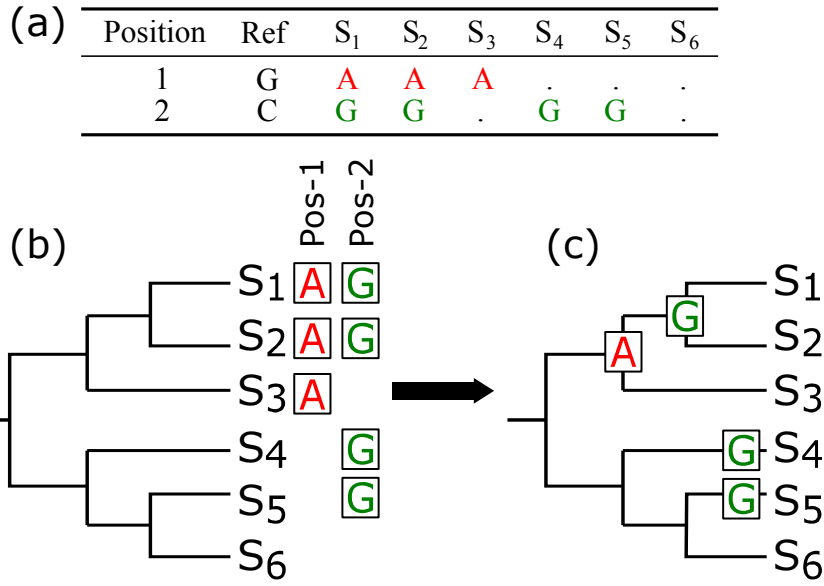


Figure I.1: (a) Example of a SNP-table with toy data on genomes of 6 strains. Each row describes a position with SNPs at the strains when compared to the reference allele. A dot character indicates that no mutation appears within the given strain, while any character indicates the presence of a SNP. (b) The SNPs at positions 1 and 2 are aligned to the leaves of the phylogenetic tree. (c) When using **CLASSICO**, the SNP at position 1 propagates up to the lowest common ancestor (LCA) of S_1 , S_2 and S_3 . Since the SNP is found only in one clade, it is clade-specific, thus supporting. The SNP found at position 2 is found after propagation at the LCA of S_1 and S_2 , as well as in the clades formed only by S_4 and S_5 , respectively. Hence, this SNP is not clade-specific, i.e. non-supporting.

snpTree (Leekitcharoenphon *et al.* [2012]), **PhyloSNP** (Faison *et al.* [2014]), **SNPhylo** (Lee *et al.* [2014; Yang *et al.* [2018]) or **SNVPhyl** (Petkau *et al.* [2017]). However, they either do not offer any kind of visualization or they lack essential interactivity for a subsequent exploration process. On the other hand, tools such as **Evolview** (Subramanian *et al.* [2019]), **iTOL** (Letunic and Bork [2019]) or **PhyD3** (Kreft *et al.* [2017]) provide the visualization of phylogenetic data with additional metadata for the analyzed samples. Still, these tools do not focus on SNP-based trees and hence they neither provide an intuitive interaction with available SNP data nor allow the identification of enriched molecular features, such as Gene Ontology terms, in genes with either clade-specific SNPs or non-clade-specific SNPs.

To cover this gap of a visual analytics tool, this paper presents **Evidente** (Efficient Visual analytics tool for Data ENrichment in phylogenetic TreEs). **Evidente** enables a simultaneous visualization of genome-wide SNP data and phylogenetic data. Furthermore, **Evidente** allows to integrate additional metadata on the analyzed samples into the visualization. By providing an interactive visual interface, **Evidente** facilitates the identification of correlations between the phylogenetic clades, their respective SNPs and metadata. Moreover, **Evidente** allows the user to run an enrichment analysis on the Gene Ontology (GO, Ashburner *et al.* [2000]) terms associated with the genes in which SNPs occur.

We demonstrate the effectiveness of our approach through an analysis of two case studies using the pathogens *Treponema pallidum* and *Mycobacterium leprae*, and thereby show how **Evidente** can be used to identify clade-specific SNPs as well as visually identify possible correlations between the genotype and metadata.

Approach

Before we developed the visual interface of **Evidente**, we first identified four main tasks for the analysis of SNP data together with metadata in a phylogenetic context in collaboration with two domain experts.

Task T_1 **Find clade-specific SNPs**: Analyze how well the SNPs support the tree structure since the tree structure itself does not give insights into the underlying SNP distribution. By viewing SNPs specific for a clade users can tell how well it is supported by SNPs.

Task T_2 **Analyze the underlying SNP/Metadata distribution of a clade**: It is not always of interest to visualize the SNPs and metadata for all individual strains separately, but summarize the data for a clade of interest to see how often the different alleles appear at a position or how the metadata is distributed.

Task T_3 **Link Metadata and non-supporting SNPs**: Visually identify correlations that are not supporting the tree. Non-supporting SNPs by definition do not correlate perfectly with the tree structure. However, they might correlate with metadata. For instance, a correlation could hint at a connection between genotype and environmental factors.

Task T_4 **Analyze a clade functionally**: Find out if any gene functions are over-represented in a clade compared to other clades. For example, one can conduct a GO enrichment analysis to find out if any GO terms linked to genes with identified SNPs are significantly over-represented in a clade compared to the rest of the tree.

With the design of **Evidente** we aim to enable the tasks defined above and allow simultaneous interrogation of phylogenetic and genotype relationships, with additional metadata expressed through colorings. To achieve this, **Evidente** visualizes a phylogenetic tree that has been computed from a set of samples from a single species together with the underlying SNPs and additional metadata aligned to the leaves of the tree. In the following the design and functionalities offered in **Evidente** are described in detail.

Data Preprocessing and Input Data

The minimum requirement to run **Evidente** are two input files which need to be provided by the user: a SNP-table in tsv-format and a Newick file representing the phylogenetic tree of the same samples as represented in the SNP-table. An example pipeline that computes a SNP-table from NGS data is EAGER (Peltzer *et al.* 2016) in conjunction with the multi-sample analyzer MUSIAL (<https://github.com/Integrative-Transcriptomics/MUSIAL>). The columns of the SNP-table (see Figure 1.1a) represent each strain or analyzed sample. The rows represent the positions in the genome where a SNP has been found in at least one sample. Each entry containing a nucleotide in IUPAC-code indicates the presence of a SNP in the corresponding position. An N character represents an unresolved base. A dot symbol symbolizes that the corresponding sample shares the same base with the reference at the given position.

The phylogenetic tree on the same samples as used for the SNP-table can be computed by any method of choice. For example, we use the SNP data and a maximum parsimony approach for this step. In the two use cases presented below, this has been done using MEGAX (Kumar *et al.* 2018).

In addition to the phylogenetic tree and the SNP data, **Evidente** is able to visualize metadata provided by the user. This information should be encoded in a CSV or a TSV-file, with the columns representing the metadata and the rows containing the name of the analyzed samples. For a correct interpretation and visualization of the metadata features,

the second row of the metadata file needs to assign each column to a data type category. For this classification, the data types presented by Munzner (2014) were chosen, and should therefore be labelled as *Categorical*, *Ordinal* or *Numerical*. **Evidente** orders categorical values by their frequency. Ordinal values are sorted alphabetically by default, though the user can re-sort them in **Evidente**.

Further optional input files are an annotation file of the reference organism in GFF format and a list of the genes with their corresponding GO-terms. The usage of these files will be explained below in the Subsection *GO Enrichment Analysis*.

Computation of the SNP clade-specificity

The goal of **Evidente** is to link the phylogenetic data and the genome-wide SNP data by an interactive visualization. Therefore, for each SNP **Evidente** assesses if it is clade-specific, i.e., the SNP appears only in the respective samples of the clade's descendant leaf-nodes. For the calculation of the clade-specificity **Evidente** includes a module called **CLASSICO** (CLAdE Specific Snp IdentifiCatOr), which computes the distribution of the SNPs from the input files with respect to the least common ancestors (LCAs) that define a clade in the phylogenetic tree (Figure I.1c). The algorithm first distributes the SNPs within the reconstructed clades by the following method: The identified SNPs are propagated from the leaves towards the root. An internal node receives a SNP from its children, as long as the SNP is present among all descendants with the same allele. This process is repeated through post-ordering of the nodes up to the root of the tree. It is important to note that some SNPs might not propagate at all and will be allocated only to one of the leaves of the phylogenetic tree. If a SNP is allocated to only one node, meaning it is a clade-specific SNP, then it is labelled as supporting of the tree structure, otherwise as non-supporting. Unresolved bases **N** do not undergo this process but are directly labelled as non-supporting SNPs, since they could be labelled as any base which prevents a clear classification. **CLASSICO** produces two lists of SNPs, one for the supporting and one for the non-supporting SNPs which are used for the visualizations in **Evidente**.

Design and Features

In this section the main visual interface as well as design choices of **Evidente** are described. To illustrate this, we have chosen an artificial phylogeny and exemplary toy SNP and metadata. The corresponding input data can be found in the supplementary material (Suppl. Tables I.S1 and I.S2).

Evidente's Graphical User Interface

The key feature of **Evidente** is to interactively and simultaneously view and explore a phylogenetic tree and whole-genome SNP data as well as metadata of the samples of one organism. The central part of the graphical user interface (GUI) of **Evidente** consists of three visualisation components that are horizontally aligned: a horizontal dendrogram, and two heat map-like visualisations of SNP and metadata (see Figure I.2I-III). In addition, the user interface includes a sidebar with graphical control elements for access to various functions as well as a settings panel for the colour encoding of the SNP data and metadata (see Figure I.2IV).

For the visualisation of the phylogenetic tree, we chose a horizontal linear dendrogram. It is created using a modified version of the D3-based library **PhyloTree** (Shank *et al.* 2018). The resulting interactive dendrogram provides typical interactions for phylogenetic

trees, such as selection of branches or nodes, collapsing subtrees on clade-nodes and the hiding of subtrees or leaves. This is described in more detail in subsection *Functionalities*.

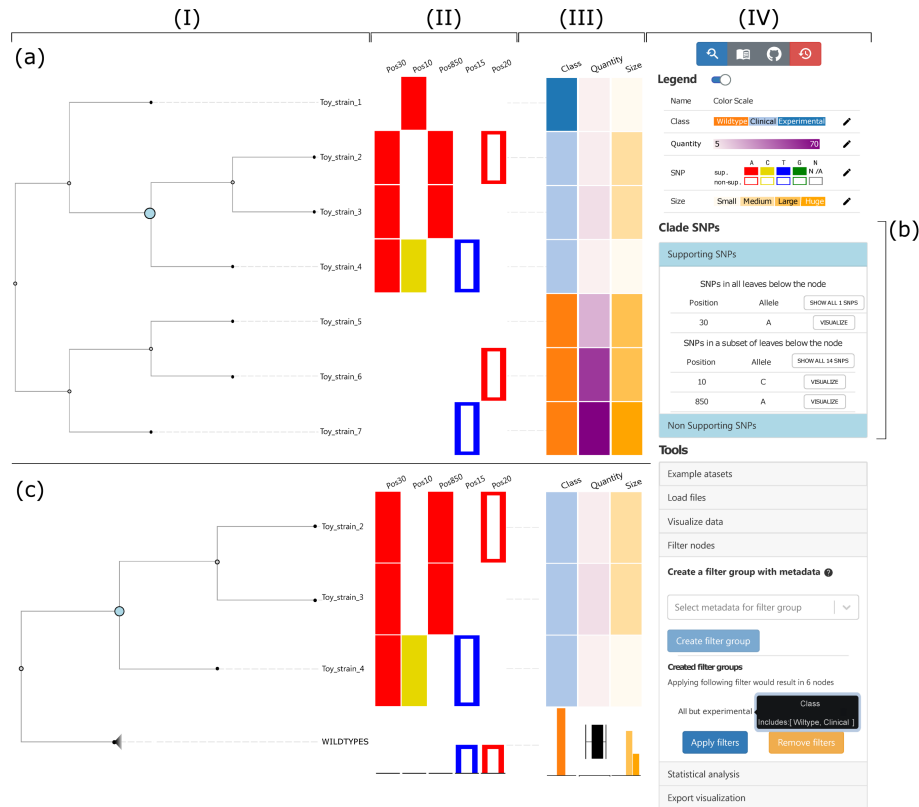
As a horizontal dendrogram was chosen, screen space is used efficiently since heat map-like visualizations of sample-specific data can be aligned horizontally to the leaves (see Figure [I.2I](#)). Two separate heat map-like visualisations are aligned to the tree encoding SNPs and metadata using color (see Figure [I.2II-III](#)). Separating SNPs and metadata facilitates differentiating the used color scales within both components more easily and helps finding patterns in the SNP data. The heat map visualizing the SNPs shows one position of the genome per column and a rectangle encodes the presence of a SNP in the corresponding sample (see Figure [I.2II](#)). The nucleotide of the SNP is visualized using a different color. Cells encoding supporting SNPs receive a full color fill, while non-supporting SNPs are encoded by cells only with a colored frame instead of a fill. Blank spaces encode the absence of a SNP at the given position, i.e. the sample in question showing the reference allele. In the metadata visualization each column of the heat map shows different sample-specific information that can be either of the numerical, ordinal, or categorical data type (see Figure [I.2III](#)). For numerical data, a continuous sequential color scale was chosen. Similarly, the ordinal data has a sequential color scale as default. Lastly, the categorical values get a default color scale of twenty colors defined by **D3** (Bostock *et al.* [2011](#)). An overview of the color scales of currently displayed data can be found in the legend component (see Figure [I.2IV](#)). The color scales can be changed interactively.

Since the distances between leaves, their labels, and the heat maps can be too large to easily identify the rows corresponding to the leaves, guiding-lines connect the different visual components. Via hovering interactions of any element, the guiding-lines are highlighted to underline the connection between associated elements. Similarly, vertical guiding-lines are displayed when a user hovers over a cell in one of the heat maps. Furthermore, users can highlight branches in the phylogenetic tree to emphasize their corresponding elements in the heat maps.

Functionalities

Interactive functionalities of **Evidente** are offered both in the dendrogram itself as well as in the sidebar (see Figure [I.2IV](#)). Initially no SNPs or metadata is selected and only the dendrogram is displayed after loading the files. Though the default view uses a dendrogram (i.e., a tree visualization that includes the branch lengths in the visualization), the user can interactively visualize the data as a cladogram, a visualization of the tree that focuses only in the descendent-ancestor relationship without including distances. SNPs belonging to a clade of interest can be explored and added to the view by selecting a node. This interaction is the key for the simultaneous exploration of the SNP and phylogenetic data (T_1). Figure [I.2a](#) shows such an interaction. After selecting a node, the SNPs corresponding to the clade are listed in a table in the sidebar (Figure [I.2b](#)) categorized by their clade-specificity and the presence of the SNP among the leaves of the clade (either present in all leaves or only in a subset of them). The selected node and the respective side bar elements for the clade SNPs are visually linked. Specific SNPs or all SNPs can be selected to be visualized from the clade of interest. Alternatively, using the sidebar, SNPs can be selected by their position and also metadata can be added to the visualization. With the simultaneous visualization of both data sources, it is possible to visually link SNP and metadata patterns (T_3).

Besides the extraction of SNPs for specific clades, further interactions such as the aggregation and filtering of the data are possible. Any internal node can be collapsed and afterwards labelled as the user desires. This functionality is helpful to get an overview



of the characteristics within a summarized clade (T_2) or to simplify the visualization by summarizing clades with similar characteristics into one element depending on their data type. Numerical values are summarized into horizontal box plots, while SNP, categorical and ordinal data are transformed into bar plots. All aggregated plots are scaled to the maximal value within each column to facilitate the comparison of aggregated data. An example for aggregation can be seen in Figure [I.2c](#).

On the other hand, the filtering function allows the user to focus the exploration to only a subset of data. For example when a research question focuses only on a specific strain of the species or that not all available samples have metadata at one's disposal. There are two alternative filtering strategies for this: either directly from the phylogenetic tree or by using available metadata. For the first option, any node but the root can be selected to be hidden together with all its descendants. The latter option is used from the corresponding section of the side bar by selecting the metadata categories that should be considered for the filtering. For each of the selected categories, the user specifies the desired characteristics of the nodes, i.e. the required range for values within the selected numerical data, or the categories from ordinal or categorical data. With this information the user defines a filter group that can be annotated at will. Every condition within a filter group needs to be fulfilled by a node in order for it to be shown. If many filter groups are created, a node needs to fulfill only one of them in order to be shown. Figure [I.2c](#) shows an example of the results of a filtering process.

The phylogenetic tree as well as the heat maps allow dragging and zooming both vertically as well as horizontally, which is especially useful for viewing cells of collapsed clades in more detail when a large tree is displayed. The visualizations and legends produced with **Evidente** can be exported in PNG or PDF format.

GO Enrichment Analysis

Besides the interactive features mentioned above, **Evidente** offers an analysis of GO-terms associated with the genes with detected SNPs to identify significantly enriched terms within a clade of the phylogenetic tree (T_4). Before running an enrichment analysis, each SNP that is found within a gene needs to be linked with the GO terms corresponding to the gene. This is achieved by using the annotation GFF file together with the file containing the GO-terms relationships. Since GO-terms are modelled in a directed acyclic graph (Ashburner *et al.* [2000](#)), each gene with one or more detected SNPs is not only annotated with the corresponding term, but also with all ancestor terms. This propagation is achieved by using the library **GOATOOLS** (Klopfenstein *et al.* [2018](#)) and an up-to-date GO hierarchy. The enrichment analysis can be run for a selected clade or for all clades that contain at least one supporting SNP. The enrichment analysis runs independently for each clade as follows: **Evidente** first extracts all GO-terms that are associated with genes affected by any kind of SNP within the selected clade. For every GO-term a right-sided Fisher's Exact Test (Fisher [1922](#)) is conducted. A scheme of the contingency table can be found in Suppl. Table I.S3. The sum of all cells of the contingency table is equal to the total number of SNPs found within annotated genes. Bonferroni's method (Holm [1979](#)) is applied to correct the resulting p -values from the statistical tests, using the number of GO-terms identified within the selected clade as the parameter for the correction.

All clades that contain enriched GO-terms after the correction are reported within a *pop-up* modal dialogue with the enriched terms, their description and their respective p -value. Using this modal the user is able to highlight the analyzed clade and add SNPs related with each over-represented GO-term to the visualization. Furthermore, the significantly enriched GO-terms can be exported as a CSV file.

Implementation and Availability

Evidente is a client-server application that uses established libraries in the backend and the frontend. The frontend of **Evidente** is written in JavaScript using React (<https://reactjs.org/>) as a structural framework and the library **D3** (Bostock *et al.* [2011]) for the rendering of the visualizations. **Evidente** uses a python server based on **Flask** (Grinberg [2018]) for data-parsing and the computation of the enrichment statistics. In order to create the background used for the GO enrichment **Evidente** makes use of **GOATOOLS** (Klopfenstein *et al.* [2018]). The module **CLASSICO** for the calculation of SNP clade-specificity is written in Java and is run as a python subprocess.

Evidente is available on the TueVis server (<https://evidente-tuevis.cs.uni-tuebingen.de/>), but can also be run locally. The stand-alone version requires **NodeJS** (Tilkov and Vinoski [2010]) and **npm** (<https://www.npmjs.com/>) to be installed.

Results

The goal of the case studies is to demonstrate how the proposed visual analysis approach enables users to execute the four tasks described in Section *Approach*. For this, **Evidente** is used to visualize the data of two studies analyzing pathogens: *Mycobacterium leprae*, which causes Hansen's disease also known as leprosy (Monot *et al.* [2009]), and *Treponema pallidum*, which causes the syphilis (subspecies TPA), yaws (subspecies TPE) or bejel (subspecies TEN) disease (Fraser *et al.* [1998]).

The first use case was performed on the data of the study by Schuenemann *et al.* ([2018]), which analyzed modern and ancient samples of *Mycobacterium leprae* extracted from different species around the world. The data consists of a total of 159 samples with 3,124 identified SNP positions. The metadata file was manually created with the information provided by the authors in the original publication. To explore the tree with the SNP data and metadata, we used **Evidente** to first collapse all clades except one clade (labeled *2F*, see Figure [L.3a]) according to the branch structures presented in the original paper of Schuenemann *et al.*. These branches represent clades of *M. leprae* strains based on different SNP (sub)types. The collapsing allows a detailed inspection of the most ancient clade of *M. leprae*. One of the findings of Schuenemann *et al.* was that the ancient *M. leprae* strains are not present in one common clade, but that they can be found in several branches. Figure [L.3a] shows this distribution by aggregating the metadata. The box plots representing the origin date of the samples indicate that ancient samples also appear in Branch 4, Branch 0 and Branch 3. The bar plots reveal that no clade contains samples from one specific continent of origin. This analysis of the metadata distribution is a typical task in the field, as described in Task T_2 .

Next, a specific branch (labelled B2) has been analyzed in detail (see Figure [L.3b]). The results of the study of Schuenemann *et al.* ([2018]) suggest that B2 should be split into three different branches (branches 2E/F/H). For a detailed analysis of the branches below B2, all other subtrees were hidden. In order to further analyze the novel branch splits discussed in the paper, the root of clade B2 is selected to identify any SNPs supporting the original split, which is an example for Task T_1 (identifying clade specific SNPs). For the root of the clade B2, **Evidente** returns that no SNP is supporting this split. This speaks in favour of the decision of splitting this branch into three different subbranches.

Furthermore, Figure [L.3b] shows that within this clade, Ethiopian samples are not found within a common clade, but are distributed across the subtree. To understand which SNPs support the splits of these clades, we used **Evidente** to select the root nodes of each clade containing only Ethiopian samples to identify and visualize these SNPs. Figure [L.3b] shows

that the first upper two Ethiopian clades have more SNPs supporting this arrangement when compared to the rest of the clades. Finally, we selected the root of the branch $2F$ to identify SNPs supporting this branch that surprisingly consists of two modern Ethiopian samples and many ancient samples. **Evidente** lists only two supporting SNPs, which underlines the heterogeneity of the strains. These two SNPs could be analyzed in detail to identify their location in the respective genomes and analyze their possible effects.

The data of the second use case consists of 75 *Treponema pallidum* samples with a total of 2,550 different SNP positions (Pla-Díaz *et al.* 2022). In addition, we used metadata regarding the antibiotic resistance of some samples against the macrolide azithromycin (Arora *et al.* 2016) as well as geographical origin, year of isolation and subspecies classification (TPA, TPE or TEN). The original phylogenetic tree contains a long branch that makes the visualization difficult to visualize as a dendrogram. Hence, the data is visualized instead as a cladogram (see Supplementary Figure I.S1).

When visualizing the *macrolide resistant* data, Figure S1a shows how resistant samples are distributed among the phylogenetic tree. In order to find a possible correlation of this metadata with SNPs, the non-supporting SNPs can be visualized, as stated in Task T_3 . By selecting, for example, the leaf labelled *NE20*, **Evidente** returns a list of 38 non-supporting SNPs. All SNPs that visually correlate only with the samples with macrolide resistance were kept in the visualisation. We tolerated the correlation with unresolved bases (N), since they could refer to any allele at the position. Figure S1a shows two SNPs having a clear correlation with resistant samples, since all sensitive samples show the reference allele at these positions. These two SNPs are the only ones of the 38 identified SNPs that show a correlation with the resistant samples. These positions (genomic positions 235,204 and 283,649) are found within the 23S rRNA operon, whose modification is associated with the resistance against azithromycin (Stamm 2010). With this finding, researchers could now test the samples for which the macrolide resistance data is not available but show these SNPs for resistance.

Finally, to demonstrate how **Evidente** can be used to fulfill task T_4 , the clade SS14 of *T. pallidum* was analyzed to identify enriched GO-terms within the clade. The GFF annotation file of the Nichols strain (RefSeq NC_021490) was used, since this reference genome was used to compute the SNP-table (Pla-Díaz *et al.* 2022). The GO-term IDs related to the coding genes were predicted using **FACoP** (<http://facop.molgenrug.nl/>). This analysis returned 61 significantly enriched GO-terms within the SS14 clade. The four most significant results were related to the cell's membrane (GO:0061024, GO:0019867, GO:0071709, GO:0009279, $p\text{-value} = 6.29 \times 10^{-31}$). **Evidente** also allows the user to visualize the distribution of (supporting or non-supporting) SNPs found within genes labelled with the GO-term of interest. Figure S1b shows the distribution of all SNPs in those genes that are associated with the GO-term *membrane organization* (GO:0061024). However, the SNP distribution hints that the over-representation is not unique for clade SS14, since a second clade (with the TPE/TEN samples) also shows SNPs in genes associated with this GO-term.

Discussion

Evidente is designed to simultaneously visualize the phylogenetic information, genome-wide SNP data and further aspects of an evolutionary reconstruction of different samples or strains from a single species. As a true visual analytics tool, it does not only provide a visual link by aligning the SNPs to the dendrogram, but it also links them computationally with its module **CLASSICO**, which adds the structural information “supporting of the tree

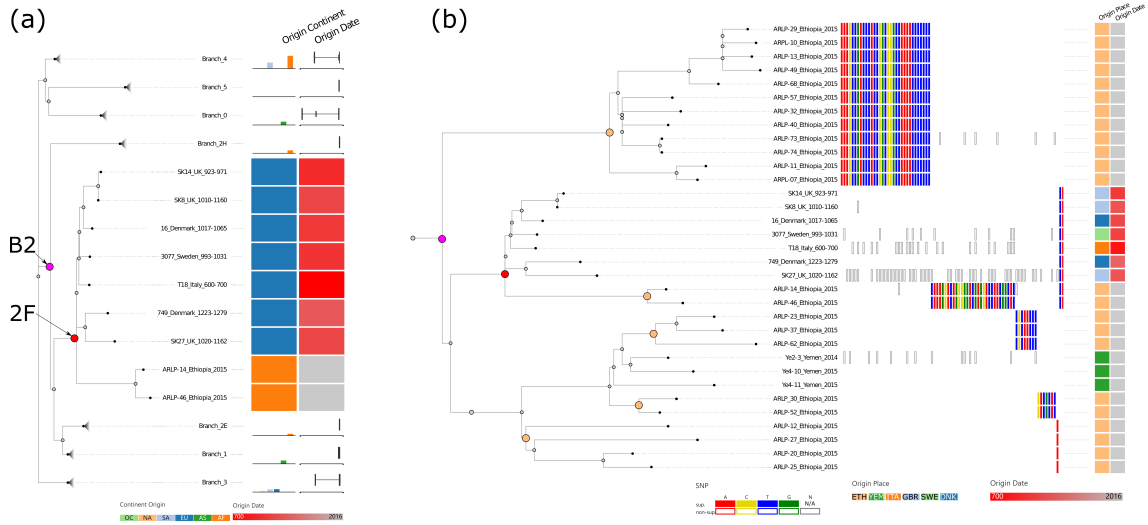


Figure I.3: Use case for **Evidente** with *Mycobacterium leprae*'s data presented by Schuenemann *et al.* (2018). (a) Visualization of the phylogenetic tree with collapsed branches as presented in the original paper. Only the clade 2F (red node) has been left uncollapsed for detailed visualization. (b) Detailed analysis of clades found within branch 2 (B2 - purple node). The rest of the clades have been hidden. Using the root of each clade containing only Ethiopian samples (orange nodes), the supporting SNPs allocated to them were selected and visualized. The last 2 SNPs represent those SNPs supporting the clade 2F (red node).

structure” and “non-supporting of the structure” to the SNPs. Moreover, this type of tree visualization is one of the most commonly used in related work and therefore is easy to interpret for users. Another layer of information is added with the GO-term enrichment analysis, which aims to find significantly enriched functions for genes affected by SNPs in a clade of interest. Further features such as filtering, aggregation and selection facilitate the interaction with the tool.

With the two case-studies we showed that the simultaneous visualization of SNP-data and metadata, and the integration of an enrichment analysis provided in **Evidente** supports the identification of interesting patterns, within analyses focused on the spread of a pathogen or in antibiotic resistance development. In the first case study, where modern and ancient samples of the *M. leprae* were analyzed, we showed that **Evidente** can effectively be used to summarize metadata by collapsing clades. Moreover, we analyzed how strongly a certain clade is supported by the SNPs. This distinction helps finding SNPs that are unique to a clade, which might be candidates for a closer inspection by a domain expert.

In the second case study an exploration process was shown for finding SNPs that are non-supporting for the tree structure but correlate with metadata for samples of *T. pallidum*. Such correlations can, for example, indicate convergent evolution, where strains in different clades in the tree develop similar mutation due to a shared environmental factor. Furthermore, we were able to show how SNPs in genes associated with a specific GO-term are distributed in the tree to find GO-terms specific for the genes affected by SNPs in two of the three main clades in the phylogenetic tree of *T. pallidum*.

In both case studies, phylogenies of bacterial strains were used. While this data source is suitable since SNP-based phylogenetic trees are often used for the analysis of the evolution of (bacterial but also viral) pathogens and their strains, the interest on eukaryotic pangenomes has started to increase in the last years (Richard 2020). Due to the simple data formats, Evidente can be applied to any properly formatted phylogenetic tree and SNP data table and thus also deal with eukaryotic pangenomes.

I. Evidente – A visual analytics tool for data enrichment in SNP-based phylogenetic trees

In our use-cases the phylogenetic tree was based on the SNPs themselves, but it might also be interesting to use the tool for analyzing a tree that is not SNP-based to investigate how well, for instance, a tree reconstructed using the whole genome is supported by SNPs. Currently, the usage of **Evidente** is restricted to the analysis of samples with a common genomic coordinate system. Hence, insertions and deletions are not considered.

In the current implementation only SNPs that are found uniquely in clades, i.e. monophyletic groups are considered supporting in order to ensure clade-conformation. However, a further category, which could be considered supporting are SNPs found within a paraphyletic group. Such a third category of SNPs will be a future work for **Evidente**'s pre-processing module **CLASSICO**.

Presently **Evidente** provides a user-friendly interaction with phylogenetic trees up to a couple of hundred genomes, like the examples shown here. on a laptop screen. The largest phylogenetic tree visualized so far in **Evidente** is the one of *M. leprae*. An overview of the data that consists of 159 samples is provided due to the possibility of collapsing clades. To facilitate the visualization of even larger trees, we plan to work on enhancing the tree scalability. For such trees, one future implementation could provide automatic collapsing of clades to reduce the number of nodes shown. With this, the user is first shown an overview of the data, before having the possibility of decollapsing clades to zoom in and interact with more detailed data, thus implementing the visual information seeking mantra by textcitesneiderman96eyes. With respect to the scalability of the SNPs, a them, the number of SNPs that can be displayed at the same time depends on the screen width. For the *Mycobacterium leprae*'s use-case, around 100 SNPs were studied simultaneously. Further SNPs can be explored by horizontally dragging the SNP heat map.

With the inclusion of GO-terms in the analysis, **Evidente** offers already the exploration of a wide range of functions. Further future work will expand the statistical computations of **Evidente**, such as for example by integrating KEGG pathways as another resource for an enrichment analysis. Moreover, we will provide further methods for statistical correction. In addition, we plan to include the metadata into the computational analysis, by either conducting an enrichment based on the metadata, or by automatically identifying SNP patterns that correlate with metadata.

Conclusion

We have presented **Evidente**, a web-application that enables an interactive and simultaneous visualization of phylogenetic, genome-wide SNP data and further metadata. The linked dendrogram of the phylogenetic tree with the two heat map-like table visualisations make the visual assessment of clade-specific or non-clade-specific SNPs and the connection of SNPs with specific metadata as well as enriched GO terms straightforward. With these visualization approaches and the functionalities implemented in **Evidente**, the identified tasks required for the interaction with the underlying data could be addressed successfully as demonstrated in the case studies. Hence, **Evidente** allows the exploration and identification of possible associations that may not be immediately evident in the data.

Acknowledgements

We would like to thank Sophie Pesch for the implementation of the statistical methods for **Evidente**. Furthermore, thanks to Katrin Fisher for the development of **CLASSICO** and Alexander Seitz for the joint work in **Evidente**'s first prototype. We thank our two

domain experts Natasha Arora and Marta Pla-Díaz. Funding for TH was provided by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — TRR261 [project ID 398967434]. MW and KN are supported by infrastructural funding from the Cluster of Excellence EXC 2124 ‘Controlling Microbes to Fight Infections’ [project ID 390838134] from the DFG. We acknowledge support by the Open Access Publishing Fund of the University of Tübingen.

Competing interests

No competing interest is declared.

Author contributions statement

MW and KN conceived and presented the idea. MW and TH worked out the technical details and carried out the implementation. MW created the figures. All authors wrote and reviewed the manuscript.

References

- Arora, N., Schuenemann, V. J., Jäger, G., Peltzer, A., Seitz, A., Herbig, A., Strouhal, M., Grillová, L., Sánchez-Busó, L., Kühnert, D., *et al.* (2016). “Origin of modern syphilis and emergence of a pandemic *Treponema pallidum* cluster”. In: *Nature Microbiology* vol. 2, no. 1, pp. 1–6.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium”. In: *Nature genetics* vol. 25, no. 1, pp. 25–29. ISSN: 1061-4036. DOI: [10.1038/75556](https://doi.org/10.1038/75556).
- Bostock, M., Ogievetsky, V., and Heer, J. (2011). “D³ data-driven documents”. In: *IEEE Transactions on Visualization and Computer Graphics* vol. 17, no. 12, pp. 2301–2309.
- Faison, W. J., Rostovtsev, A., Castro-Nallar, E., Crandall, K. A., Chumakov, K., Simonyan, V., and Mazumder, R. (2014). “Whole genome single-nucleotide variation profile-based phylogenetic tree building methods for analysis of viral, bacterial and human genomes”. In: *Genomics* vol. 104, no. 1, pp. 1–7.
- Filliol, I., Motiwala, A. S., Cavatore, M., Qi, W., Hazbón, M. H., Del Valle, M. B., Fyfe, J., García-García, L., Rastogi, N., Sola, C., *et al.* (2006). “Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set”. In: *Journal of Bacteriology* vol. 188, no. 2, pp. 759–772.
- Fisher, R. A. (1922). “On the interpretation of χ^2 from contingency tables, and the calculation of P”. In: *Journal of the Royal Statistical Society* vol. 85, no. 1, pp. 87–94.
- Foster, J. T., Beckstrom-Sternberg, S. M., Pearson, T., Beckstrom-Sternberg, J. S., Chain, P. S., Roberto, F. F., Hnath, J., Brettin, T., and Keim, P. (2009). “Whole-genome-based phylogeny and divergence of the genus *Brucella*”. In: *Journal of Bacteriology* vol. 191, no. 8, pp. 2864–2870.
- Fraser, C. M., Norris, S. J., Weinstock, G. M., White, O., Sutton, G. G., Dodson, R., Gwinn, M., Hickey, E. K., Clayton, R., Ketchum, K. A., *et al.* (1998). “Complete genome sequence of *Treponema pallidum*, the syphilis spirochete”. In: *Science* vol. 281, no. 5375, pp. 375–388.

- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). “Coming of age: ten years of next-generation sequencing technologies”. In: *Nature Reviews Genetics* vol. 17, no. 6, p. 333.
- Grinberg, M. (2018). *Flask web development: developing web applications with python*. O'Reilly Media, Inc.
- Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., and Neher, R. A. (2018). “Nextstrain: real-time tracking of pathogen evolution”. In: *Bioinformatics* vol. 34, no. 23, pp. 4121–4123.
- Holm, S. (1979). “A Simple Sequentially Rejective Multiple Test Procedure”. In: *Scandinavian Journal of Statistics* vol. 6, no. 2, pp. 65–70. ISSN: 03036898. URL: <http://www.jstor.org/stable/4615733>.
- Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., and Tang, H. (2018). “GOATOOLS: A Python library for Gene Ontology analyses”. In: *Scientific Reports* vol. 8, no. 1, p. 10872. ISSN: 2045-2322. DOI: [10.1038/s41598-018-28948-z](https://doi.org/10.1038/s41598-018-28948-z).
- Kreft, L., Botzki, A., Coppens, F., Vandepoele, K., and Van Bel, M. (2017). “PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization”. In: *Bioinformatics* vol. 33, no. 18, pp. 2946–2947.
- Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). “MEGA X: molecular evolutionary genetics analysis across computing platforms”. In: *Molecular biology and evolution* vol. 35, no. 6, p. 1547.
- Lee, T.-H., Guo, H., Wang, X., Kim, C., and Paterson, A. H. (2014). “SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data”. In: *BMC Genomics* vol. 15, no. 1, p. 162.
- Leekitcharoenphon, P., Kaas, R. S., Thomsen, M. C. F., Friis, C., Rasmussen, S., and Aarestrup, F. M. (2012). “snpTree-a web-server to identify and construct SNP trees from whole genome sequence data”. In: *BMC Genomics* vol. 13, no. 7, pp. 1–8.
- Letunic, I. and Bork, P. (2019). “Interactive Tree Of Life (iTOL) v4: recent updates and new developments”. In: *Nucleic Acids Research* vol. 47, no. W1, W256–W259.
- Monot, M., Honoré, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G. M., Donoghue, H. D., Bouwman, A., *et al.* (2009). “Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*”. In: *Nature Genetics* vol. 41, no. 12, p. 1282.
- Munzner, T. (2014). *Visualization analysis and design: Abstraction, Principles and Methods*. CRC press.
- Pandya, G. A., Holmes, M. H., Petersen, J. M., Pradhan, S., Karamycheva, S. A., Wolcott, M. J., Molins, C., Jones, M., Schriefer, M. E., Fleischmann, R. D., *et al.* (2009). “Whole genome single nucleotide polymorphism based phylogeny of *Francisella tularensis* and its application to the development of a strain typing assay”. In: *BMC microbiology* vol. 9, no. 1, p. 213.
- Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). “EAGER: efficient ancient genome reconstruction”. In: *Genome biology* vol. 17, no. 1, pp. 1–14.
- Petkau, A., Mabon, P., Sieffert, C., Knox, N. C., Cabral, J., Iskander, M., Iskander, M., Weedmark, K., Zaheer, R., Katz, L. S., *et al.* (2017). “SNVPhyl: a single nucleotide variant phylogenomics pipeline for microbial genomic epidemiology”. In: *Microbial genomics* vol. 3, no. 6.

- Pla-Díaz, M., Sánchez-Busó, L., Giacani, L., Šmajš, D., Bosshard, P. P., Bagheri, H. C., Schuenemann, V. J., Nieselt, K., Arora, N., and González-Candelas, F. (2022). “Evolutionary processes in the emergence and recent spread of the syphilis agent, *Treponema pallidum*”. In: *Molecular biology and evolution* vol. 39, no. 1, msab318.
- Richard, G.-F. (2020). “Eukaryotic Pangenomes”. In: *The pangenome: Diversity, dynamics and evolution of genomes*. Ed. by Tettelin, H. and Medini, D. Springer Nature.
- Schuenemann, V. J., Avanzi, C., Krause-Kyora, B., Seitz, A., Herbig, A., Inskip, S., Bonazzi, M., Reiter, E., Urban, C., Dangvard Pedersen, D., *et al.* (2018). “Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe”. In: *PLOS Pathogens* vol. 14, no. 5, e1006997.
- Shank, S. D., Weaver, S., and Pond, S. L. K. (2018). “phylotree.js—a JavaScript library for application development and interactive data visualization in phylogenetics”. In: *BMC Bioinformatics* vol. 19, no. 1, p. 276.
- Spyrou, M. A., Bos, K. I., Herbig, A., and Krause, J. (2019). “Ancient pathogen genomics as an emerging tool for infectious disease research”. In: *Nature Reviews Genetics* vol. 20, no. 6, pp. 323–340.
- Stamm, L. V. (2010). “Global challenge of antibiotic-resistant *Treponema pallidum*”. In: *Antimicrobial Agents and Chemotherapy* vol. 54, no. 2, pp. 583–589.
- Subramanian, B., Gao, S., Lercher, M. J., Hu, S., and Chen, W.-H. (2019). “Evolview v3: a webserver for visualization, annotation, and management of phylogenetic trees”. In: *Nucleic Acids Research* vol. 47, no. W1, W270–W275.
- Tilkov, S. and Vinoski, S. (2010). “Node.js: Using JavaScript to build high-performance network programs”. In: *IEEE Internet Computing* vol. 14, no. 6, pp. 80–83.
- Yang, H.-l., Le, D., Hui, W., LIU, C.-l., Fang, L., and XIE, C.-x. (2018). “A simple way to visualize detailed phylogenetic tree of huge genome-wide SNP data constructed by SNPhylo”. In: *Journal of Integrative Agriculture* vol. 17, no. 9, pp. 1972–1978.

Supplementary Material

Table I.S1: Excerpt of the SNP-table with toy data used for the visualizations of **Evidente**.

Position	Ref	Toy_strain_1	Toy_strain_2	Toy_strain_3	Toy_strain_4	Toy_strain_5	Toy_strain_6	Toy_strain_7
5	A	T	.	.	.	.	T	.
10	G	A	.	.	C	.	.	.
15	A	.	.	.	.	T	.	T
20	T	.	A	.	.	.	A	.
25	C	G	G	G	G	T	T	T
30	C	.	A	A	A	.	.	.
...	...	...	...	...	...	...	...	...

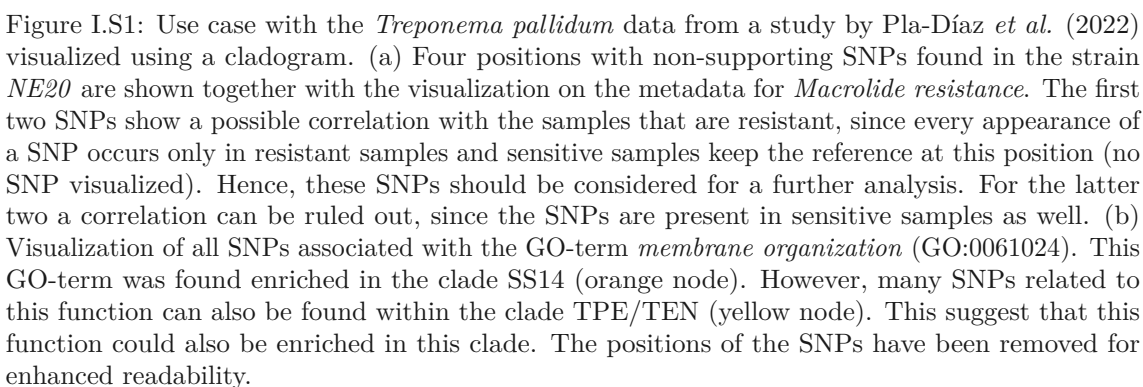
I. Evidente – A visual analytics tool for data enrichment in SNP-based phylogenetic trees

Table I.S2: Example of the input table for **Evidente** that contains the metadata of the analyzed samples. Each column indicates a metadata feature, while the rows show the strains analyzed. The type of each metadata is specified in the second row named “Type”.

Information	Weight	Quantity	Class	Size
Type	Numerical	Numerical	Categorical	Ordinal
Toy_strain_1	0.17	5	Experimental	Small
Toy_strain_2	0.9	6	Clinical	Medium
Toy_strain_3	0.8	10	Clinical	Medium
Toy_strain_4	0.012	5	Clinical	Small
Toy_strain_5	1100	22	Wild Type	Large
Toy_strain_6	200	55	Wild Type	Large
Toy_strain_7	6000	70	Wild Type	Huge

Table I.S3: Distributions of all SNPs regardless of their classification that are found within a gene within the contingency table. Each SNP is categorized depending if they are related to the GO-term G and a clade C in question for an over-representation analysis using a right-sided Fisher’s Exact Test.

	# SNPs found in C	# SNPs not found in C	
# SNPs related to G	a	b	$a + b$
# SNPs not rel. to G	c	d	$c + d$
	$a + c$	$b + d$	$N = a + b + c + d$





Appendix II

TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

Mathias Witte Paz, Alexander Herbig, Kay Nieselt

Published in *PLOS ONE*, Volume 21, Issue 3, March 2026, e0326299, DOI: [10.1371/journal.pone.0326299](https://doi.org/10.1371/journal.pone.0326299)

Abstract

Background: With the rapid development of high-throughput RNA-seq technologies, the transcriptome of prokaryotes can now be studied in unprecedented detail. Transcription start site (TSS) identification provides critical insights into transcriptional regulation. Still, current command-line tools for the prediction of TSS remain challenging with respect to their usability and lack of integrated exploration features.

Results: We introduce **TSSpredator-Web**, an interactive web application that enhances the usability of the established yet unpublished tool, **TSSpredator**. **TSSpredator-Web** facilitates TSS prediction from non-enriched and enriched RNA-seq data, classifies TSS relative to annotated genes, and allows users to explore results through dynamic visualizations and interactive tables. For the visualizations, we provide an UpSet plot summarizing the TSS distribution across experiments or classes, and a genome viewer that integrates transcriptomic and genomic data, which contextualizes the insights of the TSS predictions. To illustrate the usage of **TSSpredator-Web**, we provide a use case with Cappable-seq data from *Escherichia coli*. **TSSpredator-Web** is available on the TueVis visualization web-server at <https://tsspredator-tuevis.cs.uni-tuebingen.de/>

Conclusions: By combining user-friendly accessibility with interactive data exploration, **TSSpredator-Web** significantly facilitates genome-wide TSS analysis and interpretation in prokaryotes, empowering a broader range of researchers to generate biological insights from transcriptomic data.

Keywords: Transcription Start Sites, Web Application, Transcriptomics, Enriched RNA-seq

Introduction

Due to the development of high-throughput RNA-seq technologies, the transcriptome of a prokaryotic organism can now be studied in unprecedented detail. Far beyond the quantification of known transcripts, the precision of the data allows for the detection of novel transcripts or the definition of transcript starting boundaries (Ettwiller *et al.* 2016; Sharma *et al.* 2010).

To clearly define the base-specific starting boundary of a transcript, researchers require the identification of transcription start sites (TSS). Their identification plays a crucial role in understanding prokaryotic transcriptional regulation, as they define the exact positions where the RNA polymerase initiates transcription as well as the location of the promoter region. Promoter regions contain essential regulatory elements such as binding sites for various sigma and transcription factors (Backofen *et al.* 2014; Kazmierczak *et al.* 2005). This facilitates the generation of insights on how gene expression is regulated under varying environmental conditions (Paget 2015; Ryan *et al.* 2020). Beyond this, the base-specific TSS identification also defines untranslated regions (UTRs), which contain riboswitches and RNA thermometers (Sorek and Cossart 2010) and helps to identify other regulatory instances in bacteria, such as non-coding RNAs (ncRNAs) or antisense RNAs (Storz *et al.* 2011). Despite the critical role of TSS identification, accurately predicting these sites remains challenging. Most genome annotations only provide predicted translation start sites and coding regions. Moreover, relying solely on standard RNA-seq data for a base-exact genome-wide TSS prediction does not produce comprehensive results, as this protocol does not distinguish unprocessed, so-called *primary* transcripts from processed transcripts (i.e., transcripts with a degraded 5'-end). Reads originating from processed RNA transcripts will shift the coverage profile towards the 3'-direction and thus bias the exact TSS prediction (Sharma *et al.* 2010; Westermann and Vogel 2021). To overcome these challenges, special library preparation protocols have been developed for the determination of genome-wide TSS maps.

In 2010, Sharma *et al.* presented *differential RNA-seq* (dRNA-seq), an experimental approach to enrich for reads originating from the 5' ends of primary transcripts (i.e., transcripts containing a 5'-triphosphate instead of a monophosphate) in prokaryotes (Sharma *et al.* 2010). This is achieved by treating the cDNA with a terminator exonuclease (TEX), which specifically degrades processed RNAs with a 5'-monophosphate. A more recent method for enrichment is Cappable-seq (Ettwiller *et al.* 2016). Instead of degrading processed RNAs, a vaccinia capping enzyme (VCE) is used to positively enrich primary RNAs with a 5'-triphosphate end. A more general approach is tagRNA-seq (Innocenti *et al.* 2015), which attaches specific *tags* to primary and processed RNA molecules. Besides the difference in methodology, all enriched libraries can be sequenced to produce enriched expression profiles, thereby increasing the sensitivity and specificity of TSS annotation strategies, making genome-wide TSS identification possible. Such a comprehensive view of transcriptional activity enables the identification of not only local regulatory elements but also global regulatory patterns throughout the organism, such as unannotated ncRNAs throughout a genome (Witte Paz *et al.* 2024). With the genome-wide TSS mapping, a high-resolution view of gene regulation is provided by identifying differences in promoter usage, identification of the length of 5' untranslated regions (5'-UTRs), and the presence of novel transcripts across species, going beyond what the typical RNA-seq experiments offer, in terms of comparative transcriptomics (Dugar *et al.* 2013). However, as the scale of the data generated by these approaches grows, manual curation becomes impractical, necessitating the development of fully automated computational methods to process and

analyze TSS data efficiently.

Various tools have been developed to automate the prediction of TSS from enriched RNA sequencing libraries. Here, we focus on one of the first methods developed for TSS prediction, **TSSpredator** (Dugar *et al.* 2013). Other tools, such as TSSAR (Amman *et al.* 2014) and TSSer (Jorjani and Zavolan 2013), have been developed to identify TSS from standard and dRNA-seq data using statistical models based on the Poisson and Skellam distributions or Bayesian approximation based on the binomial distribution, respectively, to identify significantly enriched primary transcripts. A newer approach, ToNER (Promworn *et al.* 2017), has been developed for the analysis of Cappable-seq data. Similarly to the previous methods, it relies on the computation of an enrichment score for each position and the identification of those that are statistically significant. Lastly, a support vector machine learning-based approach has also been explored to improve predictive accuracy (Čuklina *et al.* 2016). Differently from all previous methods, **TSSpredator** follows a heuristic approach by mimicking the manual TSS annotation process originally described by Sharma *et al.* (2010). **TSSpredator** identifies peaks in expression across the genome by comparing the values to thresholds set by the user. Afterwards, enrichment scores are computed by comparing both RNA-seq libraries (enriched and standard) to identify transcription start sites, without relying on a statistical approach. The latest version of **TSSpredator** now includes support for three prominent experimental protocols (dRNA-seq (Sharma *et al.* 2010), Cappable-seq (Ettwiller *et al.* 2016), and tagRNA-seq (Innocenti *et al.* 2015)), alongside new features, such as the analysis of multi-contig assemblies. While neither **TSSpredator** nor its further developments have been released as a stand-alone publication, the tool has proven effective in various studies. For instance, it has been used to compare the transcriptome of one organism under different conditions (e.g., Bischler *et al.* (2015) and Ryan *et al.* (2020)), as well as cross-strain analyses to identify conserved and divergent transcriptional characteristics (e.g., Dugar *et al.* (2013) and Rohmer *et al.* (2022)).

Still, given the complexity and volume of genome-wide TSS data, **TSSpredator** as well as the other command-line tools mentioned above, often fail to provide the efficient data exploration needed for meaningful interpretation and insight generation. Interactive web applications address this challenge by simplifying both the use of the tool and the exploration of results through user-friendly interfaces. By including dynamic and interactive representations of TSS distributions, promoter sequences, and expression patterns, such platforms enable users to quickly inspect genomic regions, compare transcriptional landscapes, and identify patterns that might otherwise remain hidden. Features such as summary visualizations, interactive genome browsers, expression data overlays, and customizable filtering options enhance accessibility and facilitate hypothesis generation, making genome-wide TSS analysis more accessible, efficient, and informative.

To achieve this, we introduce **TSSpredator-Web**, a web application designed to predict and explore TSS identified by **TSSpredator**. The web interface facilitates the interaction with **TSSpredator**, for example, by reducing the hurdles of dependencies installation and by providing enhanced interactions for data upload and data sharing. Moreover, **TSSpredator-Web** enhances the exploration of results by offering interactive visualizations, providing an overview of genome-wide TSS maps, as well as detailed views on the TSS predictions, thus facilitating deeper insights into transcriptional regulation.

Methods

Comparative detection of TSS

Before introducing **TSSpredator-Web**, we introduce the underlying algorithm of **TSSpredator** for TSS detection. **TSSpredator** is able to analyze different sequencing protocols for TSS detection in prokaryotes, as long as they produce a control and an enriched library. So far, it has been tested with data provided by all established protocols mentioned above: dRNAseq (Sharma *et al.* [2010]), Cappable-seq (Ettwiller *et al.* [2016]) and tagRNA-seq (Innocenti *et al.* [2015]). The sequencing results need to be processed to produce coverage profiles of the complete reads for both libraries, either in *wiggle* or in *bedGraph* format, using mapping workflows such as **READemption** (Förstner *et al.* [2014]). **TSSpredator** then identifies the TSS by comparing both profiles and expecting an increase in expression for the TSS in the enriched library versus its non-enriched counterpart, independently for each genomic strand (forward and reverse). For this, **TSSpredator** expects the expression data pairs (enriched and non-enriched) to be normalized beforehand. For example, one possibility provided by **READemption** is to normalize the coverage by the total number of aligned reads and then multiply each position by the lowest number of aligned reads of all considered libraries (*coverage-tnoar_min_normalized*) or by a million (*coverage-tnoar_mil_normalized*). **TSSpredator** then conducts a further normalization between libraries (see Supplementary Algorithm II.S1). For each enriched library, by default, the 90th percentile (Q_{PL}) is computed and used as a factor for percentile normalization of both libraries—enriched and its corresponding non-enriched counterpart. To recover the original data range, the minimal percentile value across samples (Q_{min}) is multiplied by the normalized values. Moreover, to normalize for different enrichment factors across replicates and sets, the median enrichment factor (i.e., enriched value divided by non-enriched library) is computed by default for all library pairs (enriched and non-enriched). The largest enrichment factor EF_{max} is then used to normalize all non-enriched libraries. While not activated by default, **TSSpredator** can export the normalized libraries for further downstream analyses after TSS prediction.

After normalization, for each position i in each enriched expression graph, **TSSpredator** extracts the expression height $e(i)$, and computes, with respect to the previous position $e(i-1)$, the height change $e(i) - e(i-1)$ and the factor of height change $\frac{e(i)}{e(i-1)}$ (Fig. II.1, Suppl. Algorithms II.S2 and II.S3). These values are compared to predefined thresholds. If a position i exceeds the thresholds for step height T_h and step factor T_f , it is considered a TSS candidate and classified as *detected*. Here, it is important to note that the step height threshold is relative to the 90th percentile value used for the *inter-library* normalization. If too many of the detected TSS are found close to each other within a window size W , they are reduced by selecting either only the first TSS or the one with the highest expression. This produces a set of putative TSS per strand and per replicate.

In case of more than one replicate per experiment (that is, an organism's genome or a tested condition), predicted TSS that are within a distance of 1bp when comparing replicates are considered equal to allow a cross-replicate shift (Δ_r). This default value can be changed by the user. If a TSS is labeled as *detected* in one replicate, the corresponding positions in the other replicates are re-evaluated by reducing the thresholds by predefined *reduction values* for the step height and step factor (ρ_h and ρ_f , respectively). This increases the number of detected TSS across replicates. A TSS needs to be detected in a minimal number of replicates (R_{min} , default: 1) to be included in the next steps. However, this lower threshold can be increased for higher specificity. For all detected TSS at any position j , the enrichment factor $EF = \frac{e_{enr}(j)}{e_{non}(j)}$ (Fig. II.1, Supplementary Algorithm II.S2) is computed for

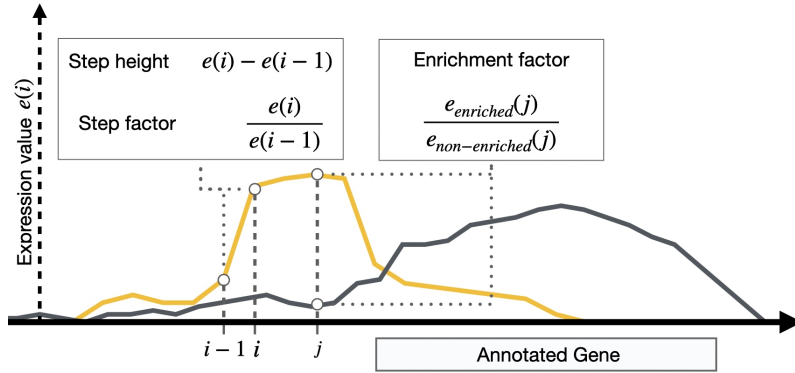


Figure II.1: Sketch of the computation of the three parameters *step height*, *step factor* and *enrichment factor* required for TSS prediction. The black line represents the non-enriched RNA-seq profile (e_{non}), covering the whole length of an annotated gene. The yellow line shows the enriched profile (e_{enr}), with an increased expression value upstream of the annotated gene.

all replicates with respect to the same position in the non-enriched library, and the maximal value across replicates is compared to the enrichment fact threshold T_{EF} . For example, if the selected threshold is 2, it means that the double expression value is expected in the enriched library compared to the control library for the TSS to be labeled as enriched. Analogously to the cross-replicate shift, all TSS found within the cross-condition shift value (default: 1bp), are considered equal. This returns one set of detected and enriched TSS per strand and experiment.

Lastly, the sets of enriched TSS per experiment are compared to each other. The comparison of TSS in a *multiple-condition* experiment is straightforward since the TSS are already in the same coordinate system. For a *multiple-strain* experiment, a common coordinate system is computed using the concept of the *SuperGenome* based on a whole-genome alignment (Dugar *et al.* [2013]), such as the XMFA file provided by Mauve (Darling *et al.* [2010]). With the alignment, TSS are compared to each other and clustered together if they are found in close chromosomal distance. This returns, per experiment, a set of detected and enriched TSS that can be further analyzed. All mentioned thresholds can be modified by the user to lower or increase the specificity and/or sensitivity of the prediction. As the prediction step is influenced by the chosen thresholds and parameters, five predefined parameter sets have been introduced to span a range of sensitivities and specificities, from *very sensitive* to *very specific* (Suppl. Fig. II.S1). An overview of how the parameter sets influence the TSS-calling process for an experiment with one condition and one replicate is provided in Suppl. Figure II.S2.

Finally, all detected TSS are classified according to their locations relative to annotated genes. While other classification methods exist (Boutard *et al.* [2016]), here we classify TSS as defined in Sharma *et al.* [2010] (Fig. II.2). 5 different classes are defined: Primary and secondary TSS are found up to 300 nt upstream of a gene's annotated translation start, where the strongest or the first signal are classified as primaries, and the rest of signals as secondary TSS. Internal TSS are found within an annotated genes, and antisense TSS are found on the antisense strand within a distance of less than 150 nt to an annotated gene. Lastly, TSS that are not associated with either of the other four classes are called orphan TSS. Note that a TSS can also get more than one class assignment (e.g., it can be a primary TSS of a gene and at the same time an antisense TSS for a gene on the opposite strand). The resulting predictions are stored in a TSV-file called *MasterTable*, which summarizes all relevant information per TSS. This table describes in detail all predicted TSS positions, such as their enrichment factor, their class, the gene to which they might be

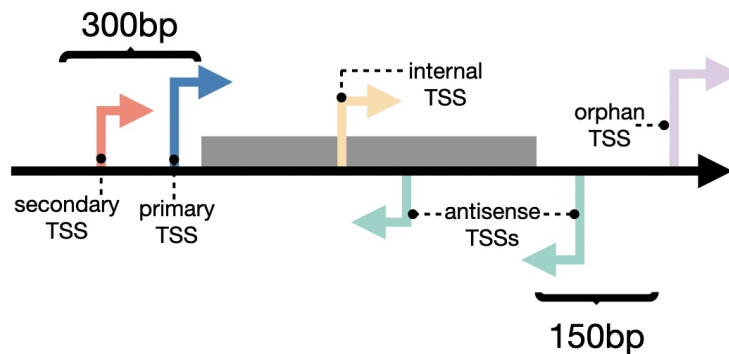


Figure II.2: Classification of TSS based on the distance to annotated genes as defined by Sharma *et al.* (2010). Primary and secondary TSS are located upstream of annotated genes, where secondary TSS show a lower enriched expression signal compared to the respective primary TSS. Internal TSS are located within the genes themselves, while antisense TSS are located on the antisense strand close to a gene (within 150 bp) or within the gene itself. Lastly, all other identified TSS are called orphan TSS.

associated, among others. This is reported for every experiment in the analysis, showing multiple lines per TSS and classification. Moreover, **TSSpredator** provides GFF files for each experiment aligned into one coordinate system. This is especially useful when analyzing multiple strains of a bacterium.

Design process of TSSpredator-Web

The workflow for TSS prediction and their association with a gene is run independently for each replicate, condition, and strand, as described above. For each replicate, four input files are required—one for each strand (forward and reverse) and experimental type (non-enriched control and enriched). Since **TSSpredator** expects each file to be correctly categorized, every file must be uploaded separately.

To facilitate this interaction, the latest version of **TSSpredator** offers a **JAVA**-based GUI to assist users in allocating files and setting the required parameters for the TSS prediction (<https://it.inf.uni-tuebingen.de/tsspredicateor>, accessed on May 2025). From this GUI, the data processing can be initiated, and the predicted results are generated and saved in external files for downstream exploration.

While effective for TSS prediction, this process presents several usability challenges. First, since **TSSpredator** is implemented in **JAVA**, it requires users to install additional dependencies, limiting platform independence and posing difficulties for users with limited technical expertise. Second, the manual file allocation process becomes increasingly cumbersome, and possible error-prone, as the number of replicates and conditions grows. Finally, downstream analysis is not integrated into the tool, so that users must export the results to other platforms to gain a comprehensive overview or biological insight.

By reviewing several genome-wide TSS studies, we identified common strategies used to explore such data sets. For example, many studies provide overview visualizations with respect to the TSS distribution across their classes or analyzed conditions (Čuklina *et al.* 2016; Hocq *et al.* 2022; Ryan *et al.* 2020). Additionally, these studies often integrate transcriptomic and genomic data, especially gene annotations and upstream regulatory regions, within one view for enhanced data exploration (Sharma *et al.* 2010).

Based on the limitations of **TSSpredator**, and the evaluation of published TSS studies, we identified eight requirements to improve the TSS analysis workflow (see Fig. II.3). These requirements can be classified with respect to the step of the TSS prediction workflow

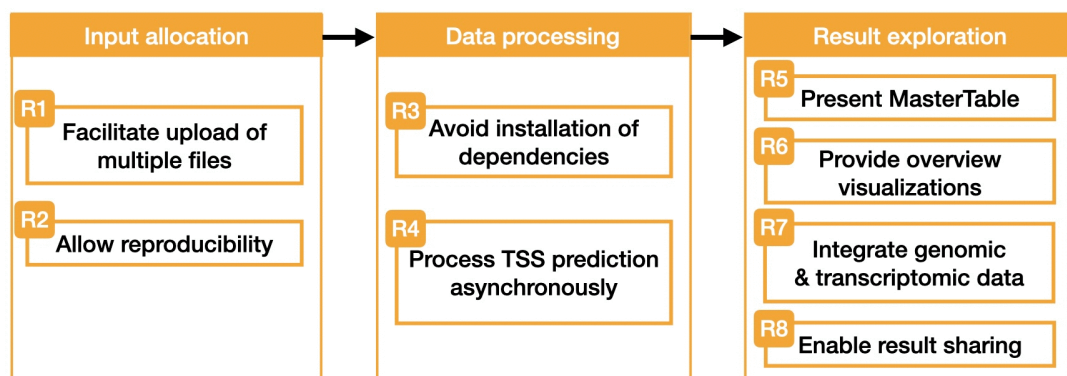


Figure II.3: Requirements identified for the improvement of the TSS prediction workflow. Each of them will be tackled with the implementation of TSSpredator-Web.

they can improve: the input allocation (R1, R2), data processing (R3, R4) and result exploration (R5, R6, R7, R8).

Building on these requirements, we have defined a set of goals to be addressed by TSSpredator-Web:

- G1 Accessibility (R3):** Ensure platform independence by eliminating installation requirements, allowing users to easily access the tool.
- G2 Usability (R1, R4–R7):** Simplify the user experience by providing easy ways of uploading and exploring the data, as well as making long waiting times bearable.
- G3 Exploration (R5–R7):** Support data exploration through interactive visualizations that present TSS predictions across levels of detail.
- G4 Data Integration (R6, R7):** Facilitate exploratory analyses by combining genomic and transcriptomic information into an interactive view, providing genomic context of the results.
- G5 Reproducibility and Data Sharing (R2, R8):** Provide mechanisms to reproduce and share both the prediction results and exploration processes with others.

Back-end of TSSpredator-Web

To facilitate the accessibility for the TSS prediction workflow, TSSpredator-Web has been designed as a web-application. It is freely accessible via the TueVis visualization platform (<https://tsspredator-tuevis.cs.uni-tuebingen.de>). However, we also provide Docker images that facilitate deployment on any server, as well as allow local usage of the tool (<https://github.com/Integrative-Transcriptomics/TSSpredator-Web/releases>). The back-end was designed with **Flask** (v2.3.3) and runs a **JAVA**-compiled version of TSSpredator for the TSS prediction. For multiple simultaneous asynchronous requests, a worker manager based on *Celery* (v5.3.4) and **redis** (v6.2.0) has been integrated in the back-end. The results of the TSS prediction are preprocessed using **Python** scripts and **bedGraphToBigWig** (Kent *et al.* 2010) for the subsequent interactive exploration in the web interface.

Web-interface

The front end of TSSpredator-Web has been developed using **React** and is structured into three distinct pages, each corresponding to one step in the TSS prediction workflow.

II. TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

The first step, *input allocation*, offers an improved file upload experience. Users can upload all necessary files at once and conveniently organize them using a *drag & drop* interface. The predefined sets of parameters described previously can be selected from this interface or adapted by the user. Once all parameters have been set, a configuration file that encompasses all the chosen parameters can be saved for future use, ensuring future reproducibility. If any user is provided with such a configuration file, they only need to upload it along with a folder containing all needed files, for **TSSpredator-Web** to allocate the inputs automatically, enabling a faster start of the analysis.

Upon starting the TSS prediction, users are redirected to the status page, where information on the data processing is regularly updated. This allows an effortless monitoring of the analysis without needing to keep the browser window open, as well as facilitating the sharing of ongoing predictions. Additionally, if any result have already been computed with **TSSpredator-Web**, the users are able to upload a **ZIP**-file of these results. This skips the need to rerun the prediction and still enabling access to the downstream data exploration features of **TSSpredator-Web**.

Data exploration & integration

To support the goals of data exploration & integration, as well as the usability improvement, **TSSpredator-Web** completes the TSS prediction workflow by providing an interactive result analysis. For this, the *MasterTable* as computed by **TSSpredator** is presented as an interactive table, complemented by two visualization approaches: an UpSet plot (Lex *et al.* [2014]) offering a high-level overview of the dataset, and a genomic viewer for contextualization of the TSS predictions.

The interactive *MasterTable* includes common features of exploration, such as searching, filtering, and sorting. Since TSS can be associated with more than one gene and therefore can be classified into more than one of the classes mentioned above, they correspond to one row in the *MasterTable* per classification. The TSS distribution among the different classes or between experiments is visually summarized by an UpSet plot (Lex *et al.* [2014]). By setting which variable should be used for the plot (either TSS classes or different experiments), the users can identify how many TSS positions have multiple TSS classes or how many of them occur in specific combinations of experiments. To get more information on these subsets of TSS, these subsets in the UpSet plot are interactive, such that users can select them, and only their corresponding rows are shown in the *MasterTable*.

For the exploration of TSS in a genomic context, a genome browser has been implemented. Based on the visualization grammar **Gosling** (LYi *et al.* [2022]), the genome browser provides aggregated and detailed views of the data for each strand independently, following Shneiderman's mantra: *Overview first, zoom and filter, details on demand* (Shneiderman [1996]). The aggregated view (Fig. II.4, top) consists of multiple visualization components. The main component shows a stacked bar chart that bins the TSS according to their position in the genome and their assigned class. Depending on the zoom level, the view is aggregated with bins of 50kbp, 10kbp or 5kbp. A further track below the stacked bar charts indicates annotated genes by including gray rectangles at the corresponding position. Such an aggregated view is shown until a full window size resolution of 50,000 bp is reached. From this point on, the detailed view (Fig. II.4, bottom) is displayed, showing each individual TSS and the surrounding annotated genes. Within this view, the main track also includes the normalized expression values for the enriched and control libraries at each genomic position. To provide information on gene regulation, the 50 bp-long sequence upstream of the TSS is visualized on a third track. A tool-tip can provide more detailed information of each TSS, gene, and expression value.

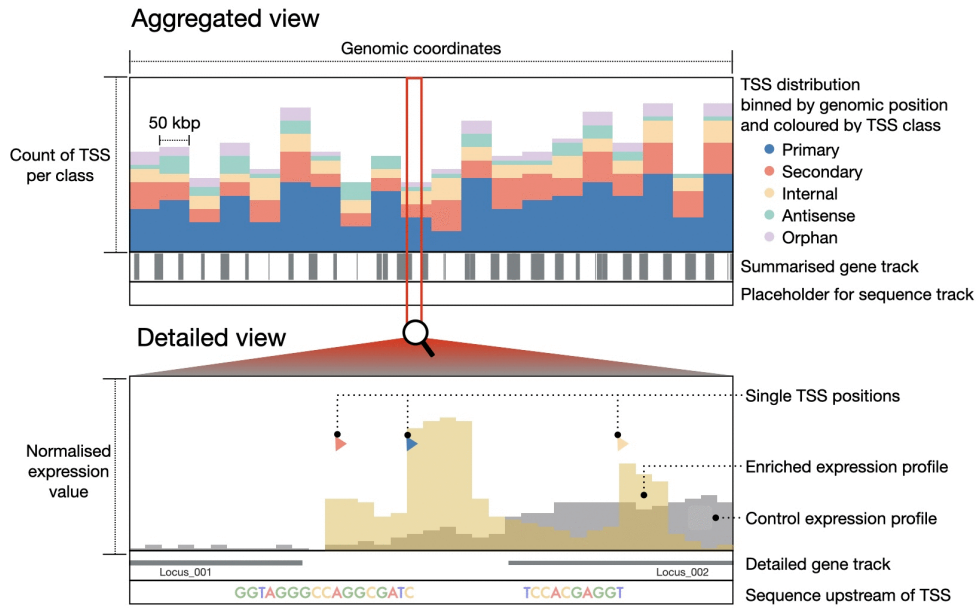


Figure II.4: Sketch of different views of the *genome browser* of TSSpredator-Web. The top section (aggregated view) displays the aggregated stacked bar chart. Each bar shows the count of TSS of a specific class in the respective genomic region bin. Below, the gene track provides a hints for the location of annotated genes. The bottom section illustrates how the plot changes upon falling below the 50,000 bp threshold (detailed view). Individual TSS locations are represented by colored glyphs, and genes are shown in their entirety with their corresponding name or *locus_tag*. The visualization also includes expression data from both control (in gray) and enriched libraries (in yellow), along with a track that illustrates the upstream region of a TSS. For simplicity, only 10 bp of the upstream region are shown on this representation instead of the actual 50 bp.

The described tracks visualize the data separately for each strand, with a difference in the orientation of the plots and glyphs (Fig. II.5). This mimics the visualization method used in other genomic browsers, such as the *Integrated Genome Browser (IGB)* (Freese *et al.* 2016) or the *Integrative Genomics Viewer (IGV)* (Thorvaldsdóttir *et al.* 2013). The *genome browser* of TSSpredator-Web provides two view arrangements to facilitate the exploration of the data within one experiment, and also the comparison across experiments (Fig. II.5). The *single view* mode groups the tracks of each experiment together (see Fig. II.5, top row), thus facilitating the exploration of single experiments, under different conditions, or single strains of one organism. Furthermore, the *aligned view* mode (see Fig. II.5, bottom row) groups the components vertically with respect to their strand, hence facilitating the comparison between conditions or strains. Regardless of the chosen arrangement, all views are synchronized, so that the same zoom level is shown in all instances. Moreover, a synced cross-line appears on hover to identify the position over all views. The *genome browser* can be freely used for exploration at any level of detail of the genomic coordinates. If users require specific TSS positions, these can be searched in the *MasterTable* and then directly visualized in the *genome browser*.

The complete predicted dataset as well as each single visualization can be downloaded from the interface. Moreover, as each prediction of TSSpredator receives a unique URL, the TSS predictions can be easily shared and accessed up to seven days after the corresponding TSS prediction was run.

II. TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

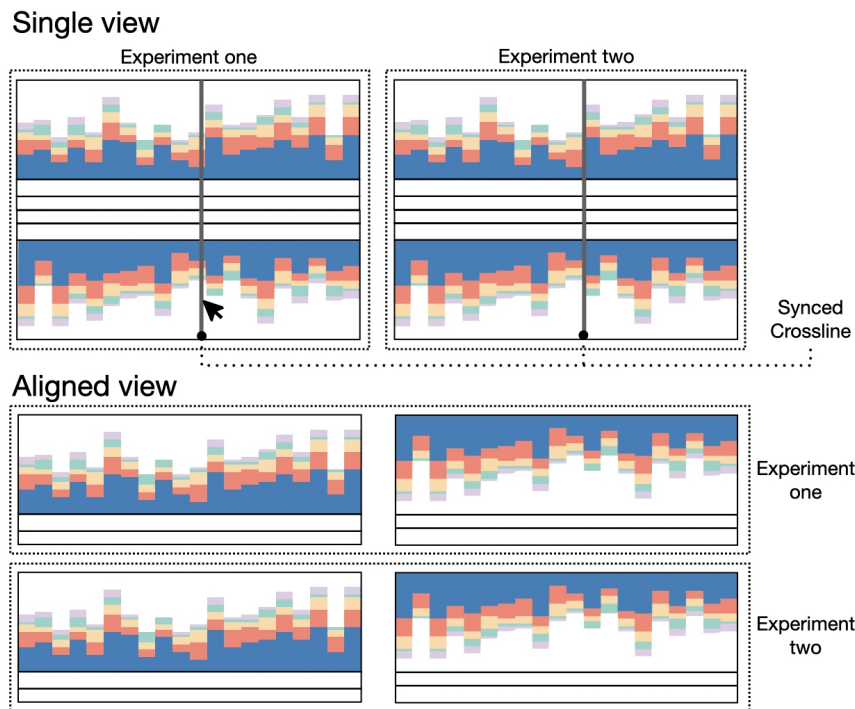


Figure II.5: Visual representation of the two modes of the genome browser to show data of multiple experiments. Each experiment consists of one component per strand, differing on the orientation of the data (for example, the reverse strand is flipped vertically). The single view mode (top section) groups the components based on the experiments for a direct exploration within each experiment. Differently, the aligned view (bottom section) groups the visualization components vertically with respect to the two strands. This allows an easier comparison across experiments. Regardless of the chosen view, a synced crossline is shown on hover. For simplicity, in this figure this line is only shown in the single view.

Use Case

To provide an example of how **TSSpredator-Web** can be used to generate and analyze genome-wide TSS data, a dataset for *Escherichia coli* K-12 MG1655 published by Balkin *et al.* (2025) [GEO Accession No. GSE215300] is used in the following section. In this study, the authors treated the *E. coli* strain with three different antibiotics (novobiocin, rifampicin and tetracycline). All three antibiotics have different modes of action. Novobiocin inhibits the DNA gyrase and, hence reduces DNA replication (Smith and Davis 1967). Rifampicin blocks the bacterial RNA polymerase, reducing the RNA synthesis (Floss and Yu 2005). Lastly, tetracycline binds the 30S ribosomal subunit and inhibits protein synthesis (Chopra and Roberts 2001). For all three treatments and a control condition, the authors measured gene expression using three replicates per condition, following the Cappable-seq protocol (Ettwiller *et al.* 2016) to generate 5'-enriched reads. Non-enriched VCE-capped reads were generated using the *NEBNext Ultra™ II Directional RNA* prep kit are also available under the same GEO accession number. **READemption** (Förstner *et al.* 2014) was used to align the reads and to compute the coverage plots in *wiggle* format. Enriched and non-enriched reads were aligned independently to the *E. coli* reference genome [NCBI Accession No. GCF_000005845.2], taking into account their different protocols for library preparations. The **READemption coverage** command provides different normalized *wiggle* files. To make both independent runs comparable, the `tnoar_mil_normalized` coverage plots were used for further analysis, consisting of 24 *wiggle* files (that is, 4

conditions $\times$ 3 replicates $\times$ 2 strands) per library protocol (non-enriched and enriched). These 48 wiggle files, together with the genome and annotation file obtained from NCBI, are the basis for the prediction of **TSSpredator-Web**.

Since one of the goals of **TSSpredator-Web** is to provide a user-friendly way to predict and explore genome-wide TSS maps, this starts already with the upload of the data. The 50 required files can be easily uploaded to **TSSpredator-Web** using the *drag & drop* functionality (Suppl. Fig. [II.S3](#)). Here, the files were distributed among the four conditions and three replicates. To increase the confidence of the results, the predetermined *very specific* parameters were chosen for the TSS prediction step. For clustering after prediction, a cross-condition shift of 3bp and a cross-replicate shift of 2bp were allowed. Based on these parameters, **TSSpredator** identified a total of 7,194 genomic positions as enriched TSS.

The genome-wide TSS exploration process starts with an overview of the TSS across classes and analyzed conditions. Taking into account only the genomic position of a TSS, the UpSet plot shows that most TSS are classified as internal, directly followed by primary TSS (Fig. [II.6A](#)). However, this distribution does not account for how often a TSS occurs across conditions, meaning that a TSS can be enriched only in one condition. This can be verified by considering both, the genomic position and the condition in which the TSS occur for the UpSet plot (Fig. [II.6B](#)). The results show that primary TSS are, in fact, the predominant class in the results. A similar result can be seen by looking at the *genome browser*, where all TSS positions are aggregated by class and genomic position per condition (Fig. [II.6C](#)), where a predominance of primary TSS can also be observed.

To analyze the distribution of primary TSS even further, one can visualize the occurrence of this class across conditions. This can also be inspected via the UpSet plot (Fig. [II.7A](#)) and provides a glimpse into an interesting subset of TSS: 43 TSS surpass the threshold for detection only in the samples treated with any of the three antibiotics, but not in the control samples (highlighted in Fig. [II.7A](#)). These positions can be analyzed in more detail by interacting with the UpSet plot to filter the *MasterTable* for this specific subset. From here, the *MasterTable* can be sorted by step height (that is, the increase in the enriched library at position i compared to the previous position $i - 1$) to show the most prominent positions (an excerpt of the *MasterTable* shown in Table [II.1](#)). Based on these results, users can search for further information, for example, by searching for more information about genes in known databases, such as the **EcoCyc** database (Keseler *et al.* [2005](#)) for *E. coli*. Some of the genes found in Table [II.1](#) were manually searched in **EcoCyc** to exemplify this exploration workflow. For example, the genes *ugd*, *arnB* and *ais*, are described as responsible for changes in membrane lipopolysaccharides (LPS), indicating a reaction against the hostile environments (Breazeale *et al.* [2002](#); Mainprize *et al.* [2013](#); Nishino *et al.* [2006](#)) as already shown for another antibiotic polymyxin (Breazeale *et al.* [2002](#)). In addition, the gene *dinI* indicates DNA damage, which can be caused by antibiotics, even though it is not part of their active mode of action (Revitt-Mills and Robinson [2020](#)). In summary, some of the TSS with the highest step height and their associated genes reflect how *E. coli* reacts to the high stress caused by antibiotics.

Although providing an overview of the most prominent TSS can be helpful, combining this information with the transcriptomic layer provides even more insight. This can be achieved through the *genome browser*. Here, we inspect the most prominent TSS with respect to the step height: position 2,099,734, the primary TSS of gene *ugd* (locus_tag: b2028, Fig. [II.7B](#)). Here, it can be seen that enrichment libraries show the highest expression for the TSS under treatment with *novobiocin*, in comparison to the other two antibiotics. A recent study identified that among these three antibiotics, the membrane LPS modification

II. TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

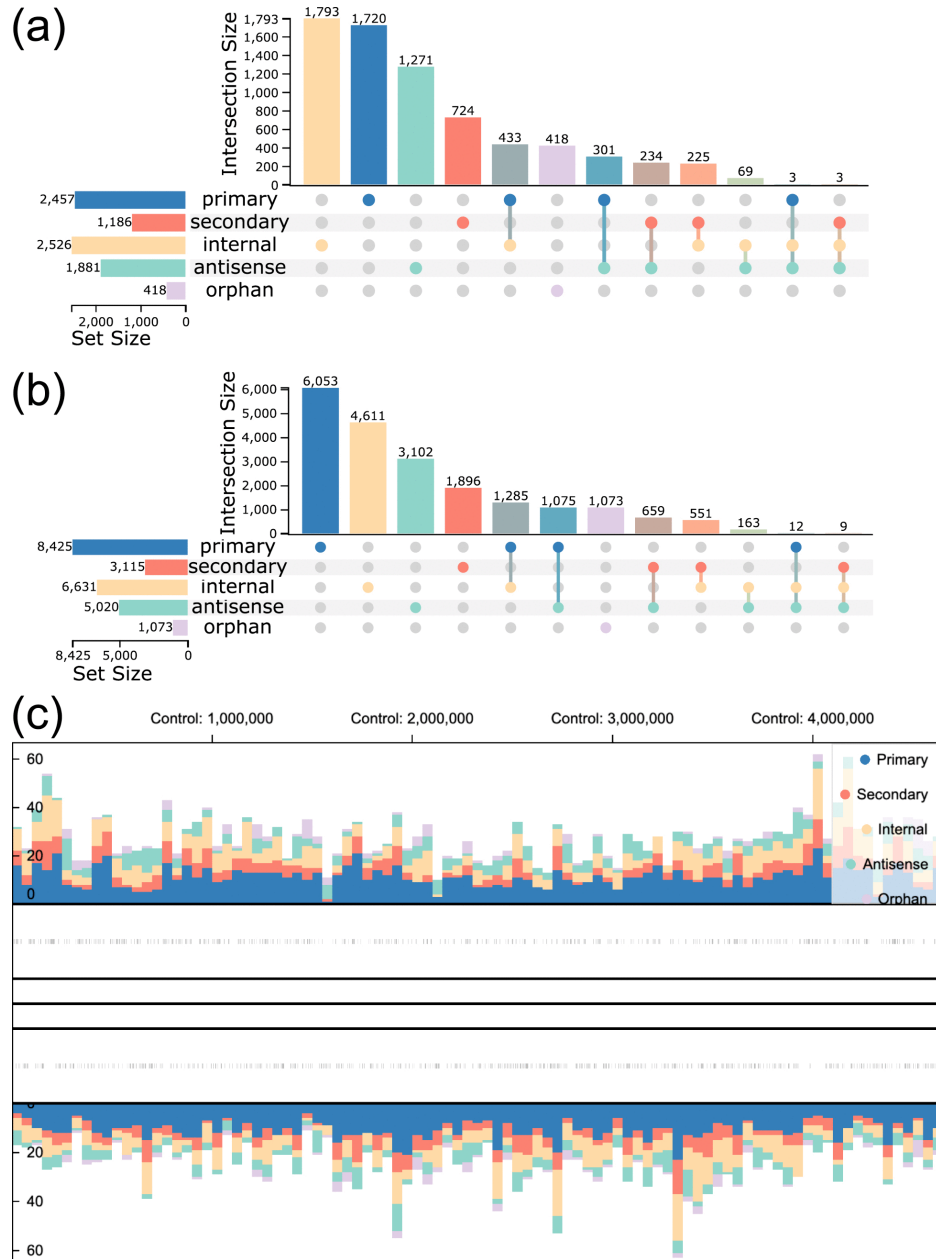


Figure II.6: Analysis of the overall distribution of TSS across conditions, classes and location in the genome for data collected for *E. coli* on four different conditions (one control and three treatment with antibiotics). (a) UpSet plot showing the distribution of enriched TSS aggregated only by their location (i.e., position and strand). (b) UpSet plot showing the distribution of enriched TSS aggregated by their location and the condition they occur. For this UpSet plot, each TSS is counted for each condition separately. (c) Aggregated view of the *genome browser* showing the distribution of TSS colored by class and binned by their position in the genome for the control condition.

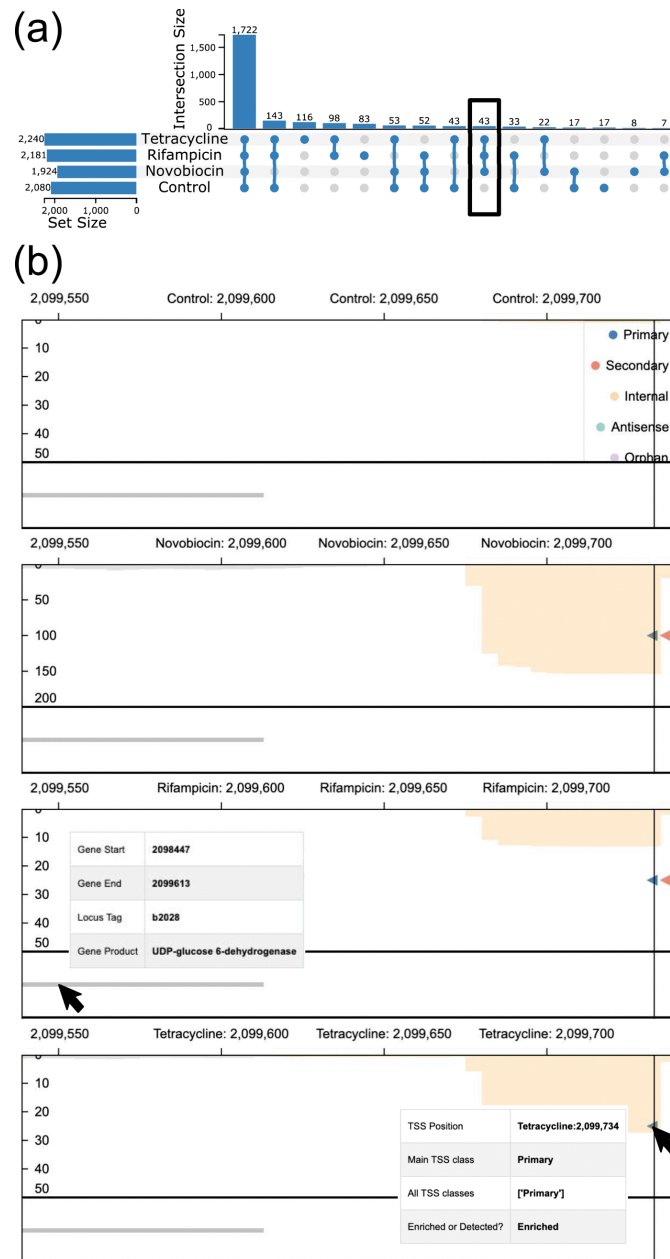


Figure II.7: Analysis of primary TSS for *E. coli* across conditions, especially those TSS occurring only under the treatment with each antibiotic. (a) UpSet plot showing the distribution of primary enriched TSS across conditions, aggregated only by their location. The highlighted set refers to those TSS positions enriched only under the treatment with each antibiotic. (b) Aligned mode of the *genome browser* showing a primary TSS position on the reverse strand (shown via the direction of the glyphs and the bars of the expression profiles) occurring in all conditions with antibiotic treatment, but not in the control condition. The TSS is located upstream of the gene *udp* (locus_tag: b2028). The enriched libraries (orange bars) can be seen increased in all conditions. However, the treatment with *novobiocin* shows the highest expression value. For simplicity, the empty *genome track* has been removed from the figure.

II. TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

Table II.1: Excerpt of the *MasterTable* showing the top 10 enriched primary TSS enriched based on step height and their associated genes. These TSS are enriched in all antibiotic treatment conditions for *E. coli*.

Pos	Strand	stepHeight	Locus_tag	Gene Name	Product
2099734	-	316.07	b2028	ugd	UDP-glucose 6-dehydrogenase
712851	+	158.06	b0687	seqA	negative modulator of initiation of replication
4275169	-	123.71	b4060	yjcB	uncharacterized protein YjcB
3775568	+	75.8	b3601	mtlR	transcriptional repressor MtlR
1056182	-	74.31	b0993	torS	sensor histidine kinase TorS
3210566	-	74.27	b3064	tsaD	N(6)-L-threonylcarbamoyladenine synthase, TsaD subunit
1121509	-	39.35	b1061	dinI	DNA damage-inducible protein I
2365854	+	34.77	b2253	arnB	UDP-4-amino-4-deoxy-L-arabinose aminotransferase
4392704	+	33.23	b4670	yjeV	uncharacterized protein YjeV
2365723	-	28.99	b2252	ais	putative lipopolysaccharide core heptose(II)-phosphate phosphatase

triggered by *ugd*, among other genes, is most effective against *novobiocin* (Deylami *et al.* 2024). A further step to analyze this region beyond TSSpredator-Web would be to extract the promoter and/or the UTR region of this gene to analyze putative regulatory elements in detail. Here, the UTR region of *ugd* was manually extracted using the coordinates provided by TSSpredator-Web, and compared to the RFAM (Ontiveros-Palacios *et al.* 2025) database outside of the presented interface. Though no hit was identified, the secondary structure of the sequence was computed using RNAfold (Gruber *et al.* 2015) and returned a stable secondary structure (see Suppl. Fig. II.S4).

TSSpredator-Web also enables the analysis of another particularly interesting class of TSS: orphan TSS. These positions correspond to transcription start sites that cannot be associated with any annotated gene from the provided file, suggesting the presence of transcriptional activity outside known gene boundaries. Such signals may represent transcriptional units that were overlooked by standard gene annotation pipelines and become detectable only through genome-wide TSS mapping. To investigate such putatively overlooked genes, we analyzed the top 10 orphan TSS present under all conditions, ranked by their step height. A prominent region is close to the orphan TSS at position 2,904,461 on the reverse strand (Fig. II.8A). Upon zooming in on this region in the *genome browser*, a noticeable increase in the enriched library can be observed, with expression levels increasing up to 350. Moreover, the upstream sequence contains a subsequence similar to the Pribnow box (TATAAA) at -9 nt upstream of the TSS (Fig. II.8B), suggesting a binding site for the RNA polymerase. When exploring the nearby regions, a second orphan TSS is identified on the forward strand at position 2,903,986 (Fig. II.8C). This TSS also shows a clear Pribnow box at position -13 nt upstream of the TSS. Together, these two orphan TSS may represent previously unannotated transcriptional units. Their pronounced step heights and well-defined upstream promoter motifs make them strong candidates for further computational and experimental validation, with the potential to improve the genomic annotation of the organism. For example, one could analyze the downstream regions of the TSS to identify possibly overlooked open reading frames.

Discussion & Conclusion

Genome-wide TSS maps provide important information for the analysis of the architecture of the prokaryotic transcriptome (Balkin *et al.* 2025; Ryan *et al.* 2020) or even recent studies on bacteriophages (Putzeys *et al.* 2022; Wolfram-Schauerte *et al.* 2024). These studies facilitate the definition of the regulatory promoter region of genes and provide clear signals for the identification of unannotated genes (Backofen *et al.* 2014; Kazmierczak *et al.* 2005; Witte Paz *et al.* 2024). Due to the large complexity of the underlying data, computational

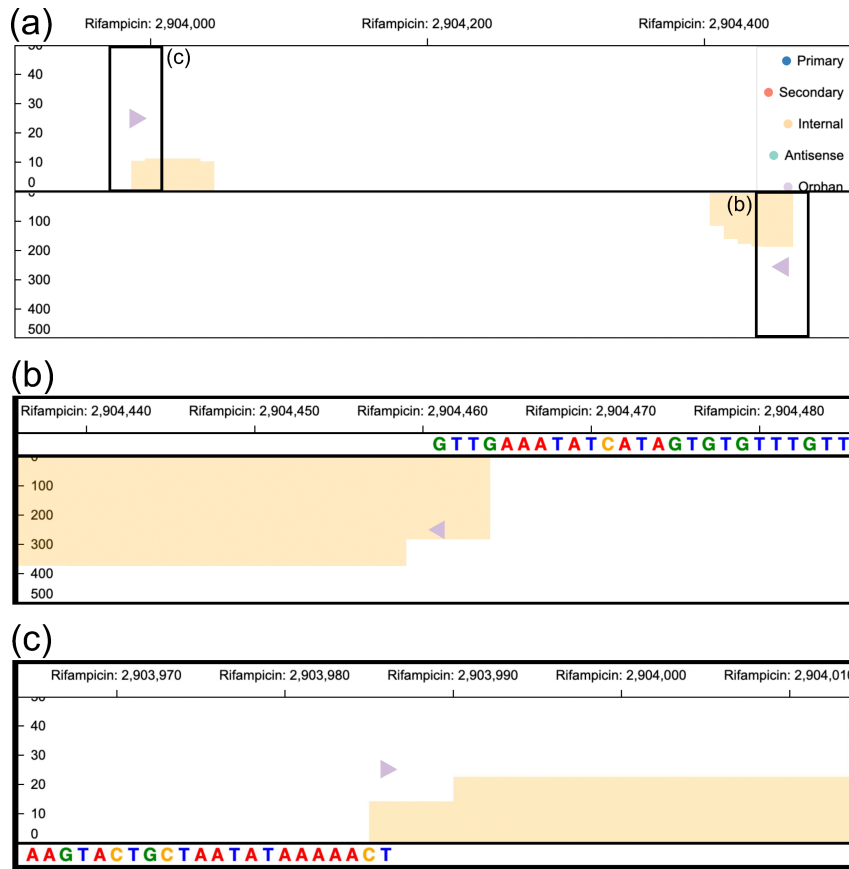


Figure II.8: Usage of the *genome browser* for the exploration and characterization of orphan TSS in *E. coli*. For simplicity, empty visualization tracks have been removed from the figures. (a) *Genome viewer* on *Single view* mode visualizing a region with two orphan TSS that occur in all conditions, shown here only for the rifampicin treatment. On the reverse strand (bottom), the TSS position 2,904,461 shows a prominent *step height* rising up to an expression of around 350. Interestingly, on the forward strand at position 2,903,986 a further orphan TSS position was identified. (b) Zoomed view of TSS at position 2,904,461 of the reverse strand with an expression value of the enriched library (orange bars) around 350. The upstream region shows a putative Pribnow box (TATAAA) starting at -9 nt upstream of the TSS. (c) Zoomed view of TSS at position 2,903,986 with an expression value of the enriched library (orange bars) around 25. The upstream region shows again a clear putative Pribnow box (TAATATAA) starting at -13 nt upstream of the TSS.

methods are required to analyze the data and produce insights. The currently available methods tackle the prediction of TSS with different underlying methodologies (Amman *et al.* [2014]; Ćuklina *et al.* [2016]; Jorjani and Zavolan [2013]; Promworn *et al.* [2017]) and most commonly provide only a command-line tool. Yet, the usage via the command-line demands technical expertise, which should not be expected from researchers with a biological background, especially since they are the individuals who generate insights from the data. Moreover, data exploration and integration are key steps in the generation of insights. However, this step has been neglected by many of the currently existing tools. Therefore, we defined requirements and goals to close the gap in insight generation for TSS prediction workflows at every step through the implementation of TSSpredator-Web.

As a web-application, the platform-independent usage and the user-friendly GUI of TSSpredator-Web enhance the accessibility of the prediction workflow (our defined goal G1). In addition, the web-based application allows for the integration of modern, user-centered approaches that address the remaining goals defined for TSSpredator-Web. For

II. TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

example, it facilitates the reproducibility and data sharing of the workflow (G5) via the upload of a configuration file containing all the input files.

Besides reproducibility, TSSpredator-Web was developed to increase the usability of the TSS prediction workflow (G2). This enhanced usability begins with the improved and intuitive file upload process and continues with the asynchronous TSS prediction executed in the backend. A further major usability improvement is the ability to explore genomic and transcriptomic data directly within the interface. Users can now rely not only on TSSpredator-Web for TSS prediction but also for data exploration (G3), as illustrated in our use case. The exploration can be pursued from different angles. The *MasterTable* provides access to all predicted results of TSSpredator, while the visualizations provide a more comprehensive view of the data. For example, a quick overview of the TSS distribution among classes, experiments, and locations in the genome can be achieved either by using the UpSet plot or by the aggregated view of the *genome browser*. As shown in the use case, users can identify specific TSS sets of interest; for example, those that occur only under specific conditions; such as those enriched only when *E. coli* was treated with antibiotics.

The aim of the *genome browser* is to provide a single view that integrates genomic and transcriptomic data (G4). Prior studies had to rely entirely on external tools, such as MEME, to provide a genomic context for a TSS. Though our implementation of TSSpredator-Web does not provide any statistical information on the occurrence of the sequences upstream of a TSS, it allows a general exploration and contextualization of the data. With this, users are not only able to identify TSS with high confidence values but also to dig deeper into their potential regulation, such as sequences present in their promoter regions.

Moreover, the ability to characterize orphan TSS, those TSS not linked to annotated transcriptional units, opens the door to discovering previously overlooked genomic elements, such as ncRNAs (Barquist *et al.* 2016; Stazic and Voß 2016). Using the genome browser of TSSpredator-Web, the confidence in such sites can be evaluated based on expression, high step height values, or the presence of a Pribnow box. Other tools, such as the recently developed pipeline TSS-CAPTUR (Witte Paz *et al.* 2024), focus on further characterizing these TSS sites to provide a hint about the functionality of the transcript and to close the gap in missing gene annotations. Both TSSpredator-Web and TSS-CAPTUR are part of the Tübingen Visualization Server (TueVis) initiative, a visualization server for user-friendly tools. In the future, it is planned to directly link TSSpredator-Web with TSS-CAPTUR to allow a seamless transition from exploratory TSS prediction to detailed characterization of uncharacterized TSS.

While the current exploration within TSSpredator-Web is primarily visual, users may also be interested in identifying statistically different TSS between conditions. This can be achieved by exporting the normalized expression profiles generated by TSSpredator and extracting the enriched coverage profile surrounding each enriched TSS position based on the *MasterTable*. These profiles are treated as distributions. Our statistical approach uses the Kolmogorov-Smirnov (KS-)test (Kolmogorov 1933; Smirnov 1948) by comparing the coverage distributions around the called TSS between the conditions. For this, we conduct pairwise KS-tests between all replicates of the conditions to be compared and compute a common *p*-value using the Cauchy combination test (Liu and Xie 2020). Finally, all the combined *p*-values of all tested TSS are subject to a Bonferroni *p*-value correction (Bonferroni 1936).

This approach has not yet been integrated into TSSpredator-Web, but we provide a preliminary implementation in the GitHub repository.

Moreover, TSSpredator-Web has so far been thoroughly tested using coverage profiles that account for the complete read length, particularly for defining the predefined

parameter sets (see Suppl. Fig. [II.S1](#)). However, *READemption* (Förstner *et al.* [2014](#)) and other coverage profile generating tools, such as **bedtools** **genomcov** (Quinlan [2014](#)) or **deeptools** (Ramírez *et al.* [2014](#)), also allow for the possibility of accounting for only the first base of the read during profile generation. While **TSSpredator-Web** has not been fully tested on such coverage profiles, we expect it to correctly identify TSS in this data as well. Still, as first-base profiles tend to be less smooth, especially in intergenic regions, we recommend running the prediction with more specific thresholds for the *step height* and *step factor* parameters as the ones used for full-read coverage profiles.

Additionally, future developments of **TSSpredator-Web** may include support for newer sequencing protocols such as Cappable-ONT (Putzeys *et al.* [2022](#)) for long-read-based TSS prediction and Term-seq (Dar *et al.* [2016](#)) for the characterization of transcription termination sites (TTS). This would provide deeper insights into RNA processing, regulation and boundary detection, further extending the depth and applicability of TSS analyses in prokaryotes.

In conclusion, **TSSpredator-Web** provides an accessible and interactive platform for genome-wide TSS exploration, improving the discovery and interpretation of bacterial transcriptomics. Its user-friendly focus, combined with the capability of visual analysis of the data, provides users with a good basis for the analysis of the prokaryotic transcriptome architecture.

Funding

MWP and KN are supported by infrastructural funding from the Cluster of Excellence EXC 2124 ‘Controlling Microbes to Fight Infections’ [project ID 390838134] from the DFG (Deutsche Forschungsgemeinschaft, German Research Foundation). There was no additional external funding received for this study. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author’s Contributions

AH and KN conceived the idea of **TSSpredator**. AH developed **TSSpredator**. MWP and KN conceived the idea of **TSSpredator-Web**. MWP developed **TSSpredator-Web**, together with its deployment and visualizations. MWP and KN wrote the initial manuscript. All authors reviewed the final manuscript.

Data Availability

TSSpredator-Web is available at the TueVis visualization web-server at <https://tsspredator-tuevis.cs.uni-tuebingen.de/>. The source code is available via Zenodo at <https://doi.org/10.5281/zenodo.15525160>. The datasets used as examples for **TSSpredator-Web**, including the data used for the described use case, are also available via Zenodo at <https://doi.org/10.5281/zenodo.17750994>.

Acknowledgements

We thank Valerie Bouillon for her contributions to the prototype of the user-interface for **TSSpredator-Web**. We also thank Dilek Tuncbilek-Dere and Sven Fillinger for further developments of **TSSpredator**. Lastly, we thank Natalia Gogoleva and Yuri Gogolev for answering questions regarding the library preparation of the analyzed dataset.

References

- Amman, F., Wolfinger, M. T., Lorenz, R., Hofacker, I. L., Stadler, P. F., and Findeiß, S. (2014). “TSSAR: TSS Annotation Regime for dRNA-seq Data”. In: *BMC Bioinformatics* vol. 15, no. 1, p. 89. ISSN: 1471-2105. DOI: [10.1186/1471-2105-15-89](https://doi.org/10.1186/1471-2105-15-89).
- Backofen, R., Amman, F., Costa, F., Findeiß, S., Richter, A. S., and Stadler, P. F. (2014). “Bioinformatics of Prokaryotic RNAs”. In: *RNA Biology* vol. 11, no. 5, pp. 470–483. DOI: [10.4161/rna.28647](https://doi.org/10.4161/rna.28647).
- Balkin, A., Plotnikov, A., Konnova, T., Shagimardanova, E., Hamo, H., Gogolev, Y., and Gogoleva, N. (2025). “Cappable-Seq RNA-sequencing Data Sets of *Escherichia coli* K-12 MG1655 Treated with Novobiocin, Tetracycline, and Rifampicin”. In: *Microbiology Resource Announcements* vol. 14, no. 2, e01194–24. DOI: [10.1128/mra.01194-24](https://doi.org/10.1128/mra.01194-24).
- Barquist, L., Burge, S. W., and Gardner, P. P. (2016). “Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families”. In: *Current Protocols in Bioinformatics* vol. 54, no. 1. ISSN: 1934-340X. DOI: [10.1002/cpbi.4](https://doi.org/10.1002/cpbi.4).
- Bischler, T., Tan, H. S., Nieselt, K., and Sharma, C. M. (2015). “Differential RNA-seq (dRNA-seq) for Annotation of Transcriptional Start Sites and Small RNAs in *Helicobacter pylori*”. In: *Methods (San Diego, Calif.)* vol. 86, pp. 89–101. DOI: [10.1016/j.ymeth.2015.06.012](https://doi.org/10.1016/j.ymeth.2015.06.012).
- Bonferroni, C. (1936). “Teoria statistica delle classi e calcolo delle probabilita”. In: *Pubblicazioni del R istituto superiore di scienze economiche e commerciali di firenze* vol. 8, pp. 3–62. DOI: [10.4135/9781412961288.n455](https://doi.org/10.4135/9781412961288.n455).
- Boutard, M., Ettwiller, L., Cerisy, T., Alberti, A., Labadie, K., Salanoubat, M., Schildkraut, I., and Tolonen, A. C. (2016). “Global Repositioning of Transcription Start Sites in a Plant-Fermenting Bacterium”. In: *Nature Communications* vol. 7, no. 1, p. 13783. ISSN: 2041-1723. DOI: [10.1038/ncomms13783](https://doi.org/10.1038/ncomms13783).
- Breazeale, S. D., Ribeiro, A. A., and Raetz, C. R. (2002). “Oxidative Decarboxylation of UDP-glucuronic Acid in Extracts of Polymyxin-Resistant *Escherichia coli*: Origin of Lipid a Species Modified with 4-Amino-4-Deoxy-l-Arabinose”. In: *Journal of Biological Chemistry* vol. 277, no. 4, pp. 2886–2896. DOI: [10.1074/jbc.M109377200](https://doi.org/10.1074/jbc.M109377200).
- Chopra, I. and Roberts, M. (2001). “Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance”. In: *Microbiology and Molecular Biology Reviews* vol. 65, no. 2, pp. 232–260. DOI: [10.1128/MMBR.65.2.232-260.2001](https://doi.org/10.1128/MMBR.65.2.232-260.2001).
- Čuklina, J., Hahn, J., Imakaev, M., Omasits, U., Förstner, K. U., Ljubimov, N., Goebel, M., Pessi, G., Fischer, H.-M., Ahrens, C. H., Gelfand, M. S., and Evguenieva-Hackenberg, E. (2016). “Genome-Wide Transcription Start Site Mapping of *Bradyrhizobium japonicum* Grown Free-Living or in Symbiosis – a Rich Resource to Identify New Transcripts, Proteins and to Study Gene Regulation”. In: *BMC Genomics* vol. 17, no. 1, p. 302. ISSN: 1471-2164. DOI: [10.1186/s12864-016-2602-9](https://doi.org/10.1186/s12864-016-2602-9).
- Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., and Sorek, R. (2016). “Term-Seq Reveals Abundant Ribo-Regulation of Antibiotics Resistance in Bacteria”. In: *Science* vol. 352, no. 6282, aad9822. DOI: [10.1126/science.aad9822](https://doi.org/10.1126/science.aad9822).
- Darling, A. E., Mau, B., and Perna, N. T. (2010). “progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement”. In: *PLoS ONE* vol. 5, no. 6, e11147. DOI: [10.1371/journal.pone.0011147](https://doi.org/10.1371/journal.pone.0011147).
- Deylami, J., Chng, S. S., and Yong, E. H. (2024). “Elucidating Antibiotic Permeation through the *Escherichia coli* Outer Membrane: Insights from Molecular Dynamics”. In:

- Journal of Chemical Information and Modeling* vol. 64, no. 21, pp. 8310–8321. DOI: [10.1021/acs.jcim.4c01249](https://doi.org/10.1021/acs.jcim.4c01249).
- Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., and Sharma, C. M. (2013). “High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple *Campylobacter jejuni* Isolates”. In: *PLOS Genetics* vol. 9, no. 5, e1003495. DOI: [10.1371/journal.pgen.1003495](https://doi.org/10.1371/journal.pgen.1003495).
- Ettwiller, L., Buswell, J., Yigit, E., and Schildkraut, I. (2016). “A Novel Enrichment Strategy Reveals Unprecedented Number of Novel Transcription Start Sites at Single Base Resolution in a Model Prokaryote and the Gut Microbiome”. In: *BMC Genomics* vol. 17, no. 1, pp. 1–14. DOI: [10.1186/s12864-016-2539-z](https://doi.org/10.1186/s12864-016-2539-z).
- Floss, H. G. and Yu, T.-W. (2005). “Rifamycin Mode of Action, Resistance, and Biosynthesis”. In: *Chemical Reviews* vol. 105, no. 2, pp. 621–632. DOI: [10.1021/cr030112j](https://doi.org/10.1021/cr030112j).
- Förstner, K. U., Vogel, J., and Sharma, C. M. (2014). “READemption—a Tool for the Computational Analysis of Deep-Sequencing-Based Transcriptome Data”. In: *Bioinformatics* vol. 30, no. 23, pp. 3421–3423. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btu533](https://doi.org/10.1093/bioinformatics/btu533).
- Freese, N. H., Norris, D. C., and Loraine, A. E. (2016). “Integrated Genome Browser: Visual Analytics Platform for Genomics”. In: *Bioinformatics* vol. 32, no. 14, pp. 2089–2095. DOI: [10.1093/bioinformatics/btw069](https://doi.org/10.1093/bioinformatics/btw069).
- Gruber, A. R., Bernhart, S. H., and Lorenz, R. (2015). “The ViennaRNA Web Services”. In: *RNA Bioinformatics*, pp. 307–326. DOI: [10.1007/978-1-4939-2291-8_19](https://doi.org/10.1007/978-1-4939-2291-8_19).
- Hocq, R., Jagtap, S., Boutard, M., Tolonen, A. C., Duval, L., Pirayre, A., Lopes Ferreira, N., and Wasels, F. (2022). “Genome-Wide TSS Distribution in Three Related Clostridia with Normalized Capp-Switch Sequencing”. In: *Microbiology Spectrum* vol. 10, no. 2, e02288–21. DOI: [10.1128/spectrum.02288-21](https://doi.org/10.1128/spectrum.02288-21).
- Innocenti, N., Golumbeanu, M., d’Hérouël, A. F., Lacoux, C., Bonnin, R. A., Kennedy, S. P., Wessner, F., Serror, P., Boulouc, P., Repoila, F., *et al.* (2015). “Whole-Genome Mapping of 5’ RNA Ends in Bacteria by Tagged Sequencing: A Comprehensive View in *Enterococcus faecalis*”. In: *RNA* vol. 21, no. 5, pp. 1018–1030. DOI: [10.1261/rna.048470.114](https://doi.org/10.1261/rna.048470.114).
- Jorjani, H. and Zavolan, M. (2013). “TSSer: An Automated Method to Identify Transcription Start Sites in Prokaryotic Genomes from Differential RNA Sequencing Data”. In: *Bioinformatics* vol. 30, no. 7, pp. 971–974. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btt752](https://doi.org/10.1093/bioinformatics/btt752).
- Kazmierczak, M. J., Wiedmann, M., and Boor, K. J. (2005). “Alternative Sigma Factors and Their Roles in Bacterial Virulence”. In: *Microbiology and Molecular Biology Reviews* vol. 69, no. 4, pp. 527–543. DOI: [10.1128/MMBR.69.4.527-543.2005](https://doi.org/10.1128/MMBR.69.4.527-543.2005).
- Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S., and Karolchik, D. (2010). “BigWig and BigBed: Enabling Browsing of Large Distributed Datasets”. In: *Bioinformatics* vol. 26, no. 17, pp. 2204–2207. DOI: [10.1093/bioinformatics/btq351](https://doi.org/10.1093/bioinformatics/btq351).
- Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M., and Karp, P. D. (2005). “EcoCyc: A Comprehensive Database Resource for *Escherichia coli*”. In: *Nucleic Acids Research* vol. 33, no. suppl_1, pp. D334–D337. DOI: [10.1093/nar/gki108](https://doi.org/10.1093/nar/gki108).
- Kolmogorov, A. (1933). “Sulla determinazione empirica di una legge di distribuzione”. In: *Giorn Dell’inst Ital Degli Att* vol. 4, pp. 89–91.
- Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. (2014). “UpSet: Visualization of Intersecting Sets”. In: *IEEE Transactions on Visualization and Computer graphics* vol. 20, no. 12, pp. 1983–1992. ISSN: 1077-2626. DOI: [10.1109/TVCG.2014.2346248](https://doi.org/10.1109/TVCG.2014.2346248).

- Liu, Y. and Xie, J. (2020). “Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures”. In: *Journal of the American Statistical Association* vol. 115, no. 529, pp. 393–402. DOI: [10.1080/01621459.2018.1554485](https://doi.org/10.1080/01621459.2018.1554485).
- LYi, S., Wang, Q., Lekschas, F., and Gehlenborg, N. (2022). “Gosling: A Grammar-Based Toolkit for Scalable and Interactive Genomics Data Visualization”. In: *IEEE Transactions on Visualization and Computer Graphics* vol. 28, no. 1, pp. 140–150. ISSN: 1077-2626, 1941-0506, 2160-9306. DOI: [10.1109/TVCG.2021.3114876](https://doi.org/10.1109/TVCG.2021.3114876).
- Mainprize, I. L., Bean, J. D., Bouwman, C., Kimber, M. S., and Whitfield, C. (2013). “The UDP-glucose Dehydrogenase of *Escherichia coli* K-12 Displays Substrate Inhibition by NAD That Is Relieved by Nucleotide Triphosphates”. In: *Journal of Biological Chemistry* vol. 288, no. 32, pp. 23064–23074. DOI: [10.1074/jbc.M113.486613](https://doi.org/10.1074/jbc.M113.486613).
- Nishino, K., Hsu, F.-F., Turk, J., Cromie, M. J., Wösten, M. M., and Groisman, E. A. (2006). “Identification of the Lipopolysaccharide Modifications Controlled by the *Salmonella* PmrA/PmrB System Mediating Resistance to Fe (III) and Al (III)”. In: *Molecular Microbiology* vol. 61, no. 3, pp. 645–654. DOI: [10.1111/j.1365-2958.2006.05273.x](https://doi.org/10.1111/j.1365-2958.2006.05273.x).
- Ontiveros-Palacios, N., Cooke, E., Nawrocki, E. P., Triebel, S., Marz, M., Rivas, E., Griffiths-Jones, S., Petrov, A. I., Bateman, A., and Sweeney, B. (2025). “RFAM 15: RNA Families Database in 2025”. In: *Nucleic Acids Research* vol. 53, no. D1, pp. D258–D267. DOI: [10.1093/nar/gkac1023](https://doi.org/10.1093/nar/gkac1023).
- Paget, M. S. (2015). “Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution”. In: *Biomolecules* vol. 5, no. 3, pp. 1245–1265. DOI: [10.3390/biom5031245](https://doi.org/10.3390/biom5031245).
- Promworn, Y., Kaewprommal, P., Shaw, P. J., Intarapanich, A., Tongsima, S., and Piriyaongsa, J. (2017). “ToNER: A Tool for Identifying Nucleotide Enrichment Signals in Feature-Enriched RNA-seq Data”. In: *PLOS ONE* vol. 12, no. 5, pp. 1–18. DOI: [10.1371/journal.pone.0178483](https://doi.org/10.1371/journal.pone.0178483).
- Putzeys, L., Boon, M., Lammens, E.-M., Kuznedelov, K., Severinov, K., and Lavigne, R. (2022). “Development of ONT-cappable-seq to Unravel the Transcriptional Landscape of *Pseudomonas* Phages”. In: *Computational and Structural Biotechnology Journal* vol. 20, pp. 2624–2638. ISSN: 2001-0370. DOI: [10.1016/j.csbj.2022.05.034](https://doi.org/10.1016/j.csbj.2022.05.034).
- Quinlan, A. R. (2014). “BEDTools: the Swiss-army tool for genome feature analysis”. In: *Current protocols in bioinformatics* vol. 47, no. 1, pp. 11–12. DOI: [10.1002/0471250953.bi1112s47](https://doi.org/10.1002/0471250953.bi1112s47).
- Ramírez, F., Dünder, F., Diehl, S., Grüning, B. A., and Manke, T. (2014). “deepTools: a flexible platform for exploring deep-sequencing data”. In: *Nucleic acids research* vol. 42, no. W1, W187–W191. DOI: [10.1093/nar/gku365](https://doi.org/10.1093/nar/gku365).
- Revitt-Mills, S. A. and Robinson, A. (2020). “Antibiotic-Induced Mutagenesis: Under the Microscope”. In: *Frontiers in Microbiology* vol. 11, p. 585175. DOI: [10.3389/fmicb.2020.585175](https://doi.org/10.3389/fmicb.2020.585175).
- Rohmer, C., Dobritz, R., Tuncbilek-Dere, D., Lehmann, E., Gerlach, D., George, S. E., Bae, T., Nieselt, K., and Wolz, C. (2022). “Influence of *Staphylococcus aureus* Strain Background on Sa3int Phage Life Cycle Switches”. In: *Viruses* vol. 14, no. 2471. ISSN: 1999-4915. DOI: [10.3390/v14112471](https://doi.org/10.3390/v14112471).
- Ryan, D., Jenniches, L., Reichardt, S., Barquist, L., and Westermann, A. J. (2020). “A High-Resolution Transcriptome Map Identifies Small RNA Regulation of Metabolism in the Gut Microbe *Bacteroides thetaiotaomicron*”. In: *Nature communications* vol. 11, no. 1, p. 3557. DOI: [10.1038/s41467-020-17348-5](https://doi.org/10.1038/s41467-020-17348-5).

- Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010). “The Primary Transcriptome of the Major Human Pathogen *Helicobacter pylori*”. In: *Nature* vol. 464, no. 7286, pp. 250–255. ISSN: 1476-4687. DOI: [10.1038/nature08756](https://doi.org/10.1038/nature08756).
- Shneiderman, B. (1996). “The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations”. In: *Proceedings 1996 IEEE Symposium on Visual Languages*, pp. 336–343. DOI: [10.1109/VL.1996.545307](https://doi.org/10.1109/VL.1996.545307).
- Smirnov, N. (1948). “Table for estimating the goodness of fit of empirical distributions”. In: *The annals of mathematical statistics* vol. 19, no. 2, pp. 279–281. DOI: [10.1214/aoms/1177730256](https://doi.org/10.1214/aoms/1177730256).
- Smith, D. H. and Davis, B. D. (1967). “Mode of Action of Novobiocin in *Escherichia coli*”. In: *Journal of Bacteriology* vol. 93, no. 1, pp. 71–79. DOI: [10.1128/jb.93.1.71-79.1967](https://doi.org/10.1128/jb.93.1.71-79.1967).
- Sorek, R. and Cossart, P. (2010). “Prokaryotic Transcriptomics: A New View on Regulation, Physiology and Pathogenicity”. In: *Nature Reviews Genetics* vol. 11, no. 1, pp. 9–16. DOI: [10.1038/nrg2695](https://doi.org/10.1038/nrg2695).
- Stazic, D. and Voß, B. (2016). “The Complexity of Bacterial Transcriptomes”. In: *Journal of Biotechnology* vol. 232, pp. 69–78. ISSN: 0168-1656. DOI: [10.1016/j.jbiotec.2015.09.041](https://doi.org/10.1016/j.jbiotec.2015.09.041).
- Storz, G., Vogel, J., and Wassarman, K. M. (2011). “Regulation by Small RNAs in Bacteria: Expanding Frontiers”. In: *Molecular Cell* vol. 43, no. 6, pp. 880–891. DOI: [10.1016/j.molcel.2011.08.022](https://doi.org/10.1016/j.molcel.2011.08.022).
- Thorvaldsdóttir, H., Robinson, J. T., and Mesirov, J. P. (2013). “Integrative Genomics Viewer (IGV): High-performance Genomics Data Visualization and Exploration”. In: *Briefings in Bioinformatics* vol. 14, no. 2, pp. 178–192. ISSN: 1467-5463. DOI: [10.1093/bib/bbs017](https://doi.org/10.1093/bib/bbs017).
- Westermann, A. J. and Vogel, J. (2021). “Cross-species RNA-seq for deciphering host–microbe interactions”. In: *Nature Reviews Genetics* vol. 22, no. 6, pp. 361–378. DOI: [10.1038/s41576-021-00326-y](https://doi.org/10.1038/s41576-021-00326-y).
- Witte Paz, M., Vogel, T., and Nieselt, K. (2024). “TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA Transcripts”. In: *NAR Genomics and Bioinformatics* vol. 6, no. 4, lqae168. ISSN: 2631-9268. DOI: [10.1093/nargab/lqae168](https://doi.org/10.1093/nargab/lqae168).
- Wolfram-Schauerte, M., Moskalchuk, A., Pozhydaieva, N., Rojas, A. A. R., Schindler, D., Kaiser, S., Pazcia, N., and Höfer, K. (2024). “T4 Phage RNA Is NAD-capped and Alters the NAD-cap Epitranscriptome of *Escherichia coli* during Infection through a Phage-Encoded Decapping Enzyme”. In: *bioRxiv*, pp. 2024–04. DOI: [10.1101/2024.04.04.588121](https://doi.org/10.1101/2024.04.04.588121).

Supplementary Material

The TSSpredator algorithm

Algorithm II.S1 Library normalization of TSSpredator

Input: $r \in N$ with N is the number of coverage profile pairs (samples)

C_{non}^r : Normalized non-enriched coverage profiles r

C_{enr}^r : Normalized enriched coverage profiles r

PL : Percentile for library normalization, default $PL = 90$

PE : Percentile for enrichment normalization, default $PE = 50$

Output: Library normalized coverage profiles e_{non}^r and e_{enr}^r

$Q \leftarrow \emptyset$

$EF \leftarrow \emptyset$

for all $r \in N$ **do**

$Q^r \leftarrow \text{GETPERCENTILE}(C_{enr}^r, PL)$

$Q \leftarrow \{Q^i, Q\}$

for $s \in \{+, -\}$ **do**

$N_{enr}^{r,s} \leftarrow C_{enr}^{r,s} / Q^i$

$N_{non}^{r,s} \leftarrow C_{non}^{r,s} / Q^i$

end for

end for

$Q_{min} \leftarrow \min(Q)$

for all $s \in \{+, -\}$ and $i \in N$ **do**

$N_{enr}^{*r,s} \leftarrow N_{enr}^{r,s} * Q_{min}$

$N_{non}^{*r,s} \leftarrow N_{non}^{r,s} * Q_{min}$

$\mathcal{T} \leftarrow \text{PREDICTTSS}(N_{enr}^{*r,s}, N_{non}^{*r,s})$

$F_r \leftarrow \text{GETENRICHMENTFACTORS}(\mathcal{T})$

$EF_r \leftarrow \text{GETPERCENTILE}(F_r, PE)$

$EF \leftarrow \{EF_r, EF\}$

end for

$EF_{max} \leftarrow \max(EF)$

for all $s \in \{+, -\}$ and $r \in N$ **do**

$e_{non}^{r,s} \leftarrow N_{non}^{*r,s} * \frac{EF_i}{EF_{max}}$

$e_{enr}^{r,s} \leftarrow N_{enr}^{*r,s}$

end for

Algorithm II.S2 Helper functions for TSS prediction

```
function DETECTTSSPERREPLICATE( $e_{enr}^r, Q_{min}, T_h, T_f, W$ )  
   $T_h^{abs} \leftarrow T_h \cdot Q_{min}, \mathcal{D} \leftarrow \emptyset$   
  for  $i = 2$  to  $L$  do  
     $\Delta h \leftarrow e_{enr}^r(i) - e_{enr}^r(i-1), F \leftarrow e_{enr}^r(i)/e_{enr}^r(i-1)$   
    if  $\Delta h \geq T_h^{abs}$  and  $F \geq T_f$  then  
       $\mathcal{D} \leftarrow \{\mathcal{D}, i\}$   
    end if  
  end for  
  return COLLAPSECLOSEPOSITIONS( $\mathcal{C}, W$ )  
end function  
  
function GETENRICHMENTFACTORACROSSREPLICATES( $\mathcal{T}_k^{detected}(j), e_{k,non}^r, e_{k,enr}^r$ )  
   $EF_{max} \leftarrow 0$   
  for all replicates  $r$  of condition  $k$  do  
     $EF \leftarrow e_{enr}^r(j)/e_{non}^r(j)$   
     $\max(EF_{max}, EF)$   
  end for  
  return  $EF_{max}$   
end function
```

Algorithm II.S3 TSS prediction with TSSpredator, for each strand individually

Input: Library normalized coverage profiles e_{non}^r and e_{enr}^r for all conditions K
 Minimal percentile Q_{min} , Thresholds step height T_h , step factor T_f and enrichment factor T_{EF}
 Allowed shifts per replicate Δ_r , and conditions Δ_k , Reduction values ρ_h, ρ_f
 Minimal replicate support R_{min} , Collapse window size W
for all replicates r **do**
 $\mathcal{T}_r \leftarrow \text{DETECTTSSPERREPLICATE}(e_{enr}^r, Q_{min}, T_h, T_f, W)$
end for

▷ Cross-replicate grouping and threshold relaxation
 ▷ *If TSS have a position difference of at most Δ_r across replicates of a condition, account as equal*
 $\mathcal{T}_k \leftarrow \text{GROUPWITHINCONDITIONS}(\mathcal{T}_r, \Delta_r)$
for all clusters $\mathcal{T}_k$ with position j **do**
 for all replicates r not represented in $\mathcal{T}_k$ **do**
 $T'_f \leftarrow \max(1, T_f - \rho_f)$, $T'_h \leftarrow \max(0, T_h - \rho_h)$
 $T_h^{\text{abs}} \leftarrow T'_h \cdot Q_{min}$
 for all $x \in [j - \Delta_r, j + \Delta_r]$ **do**
 $\Delta h \leftarrow e_{enr}^r(x) - e_{enr}^r(x - 1)$, $F \leftarrow e_{enr}^r(x) / e_{enr}^r(x - 1)$
 if $\Delta h \geq T_h^{\text{abs}}$ **and** $F \geq T'_f$ **then**
 $\mathcal{T}_k \leftarrow \{\mathcal{K}_k, x\}$
 break
 end if
 end for
 end for
end for

▷ *A TSS needs to occur at least in R_{min} replicated to be accounted*
 $\mathcal{T}_k^{\text{detected}} \leftarrow \text{FILTERFORMINIMALOCCURRENCE}(\mathcal{T}_k, R_{min})$

▷ Enrichment factor evaluation
 $\mathcal{T}_k^{\text{enriched}} \leftarrow \emptyset$
for all TSS in $\mathcal{T}_k^{\text{detected}}$ with position j **do**
 $EF \leftarrow \text{GETENRICHMENTFACTORACROSSREPLICATES}(\mathcal{T}_k(j), e_{k,non}^r, e_{k,enr}^r)$
 if $EF \geq EF_{thr}$ **then**
 $\mathcal{T}_k^{\text{enriched}} \leftarrow \{\mathcal{T}_k^{\text{enriched}}, j\}$
 end if
end for

▷ Cross-condition merging
 ▷ *If TSS have a position difference of at most Δ_k across conditions, account as equal*
 $\mathcal{T}^{\text{enriched,detected}} \leftarrow \text{GROUPACROSSCONDITIONS}(\mathcal{T}_k^{\text{enriched}}, \mathcal{T}_k^{\text{detected}}, \Delta_k)$
return $\mathcal{T}^{\text{enriched,detected}}$

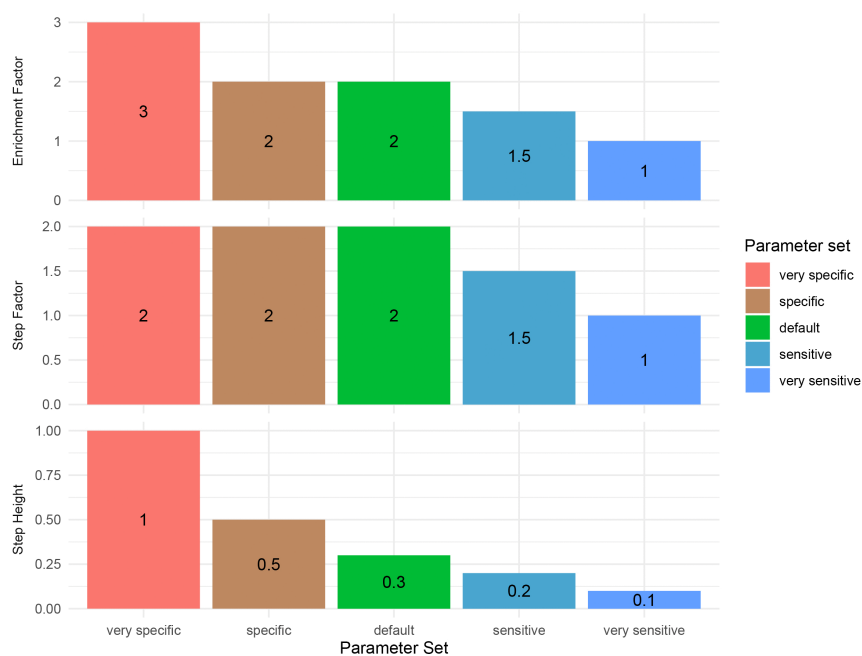


Figure II.S1: Overview of the five predefined sets of thresholds for *Step height*, *Step factor* and *Enrichment factor* for the TSS prediction with TSSpredator.

II. TSSpredator-Web: A Web-Application for Transcription Start Site Prediction and Exploration

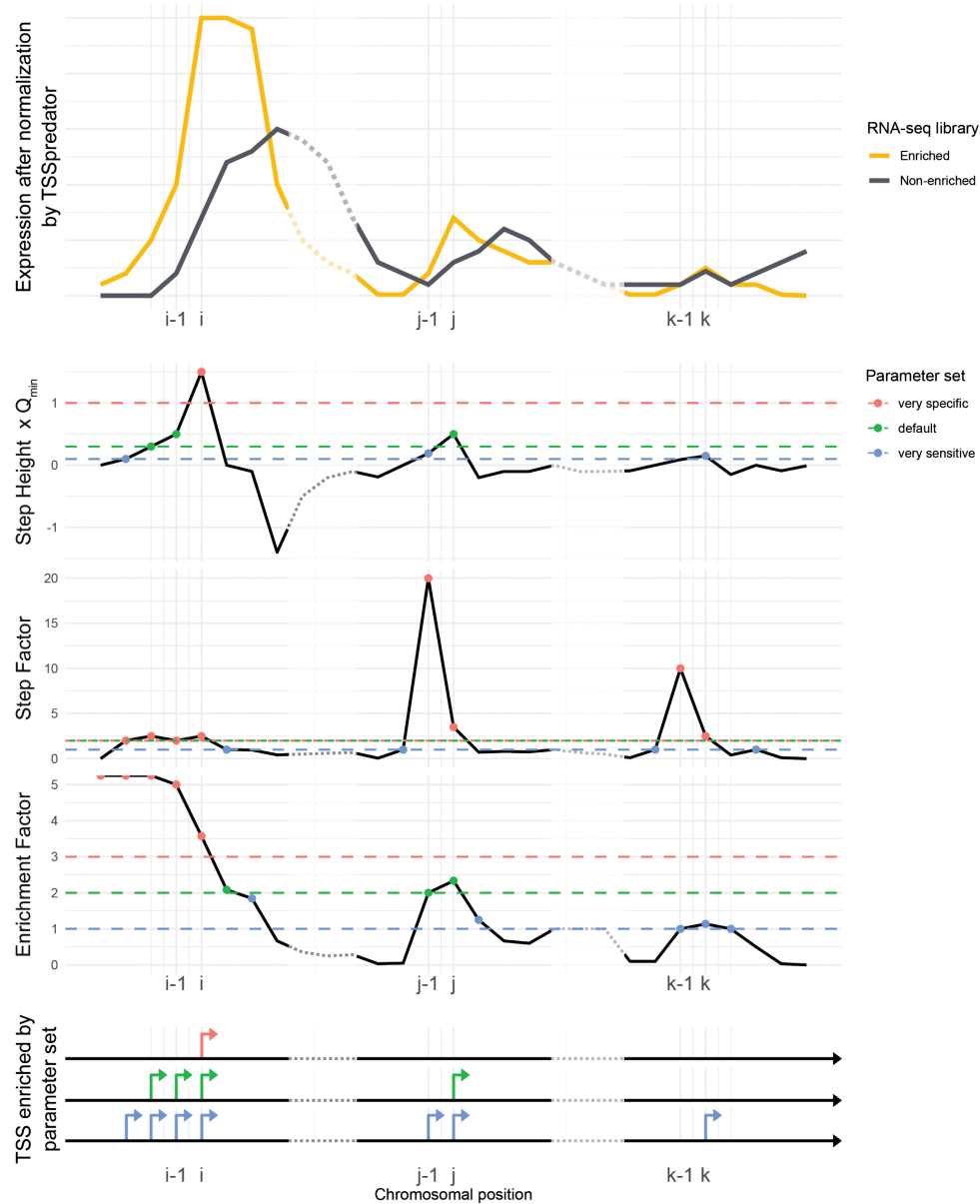


Figure II.S2: Overview of how different parameter sets influence TSS calling in TSSpredator. Three discontinuous genomic positions (i , j , k) show loci with putative TSS. The top plot shows the normalized expression profiles of the enriched and non-enriched RNA-seq libraries for one condition and one replicate. The middle plots display the derived metrics *Step Height*, *Step Factor*, and *Enrichment Factor*, together with the corresponding thresholds for the *very specific*, *default*, and *very sensitive* parameter sets. Positions exceeding a given threshold are marked with a colored point according to that parameter set. Note that for *Step Factor*, the thresholds for the *very specific* and *default* sets coincide. In the bottom plot, a TSS is called at positions where all three metrics surpass the thresholds of a given parameter set. The color of each TSS indicates the parameter set that would identify it.

Interface of TSSpredator-Web

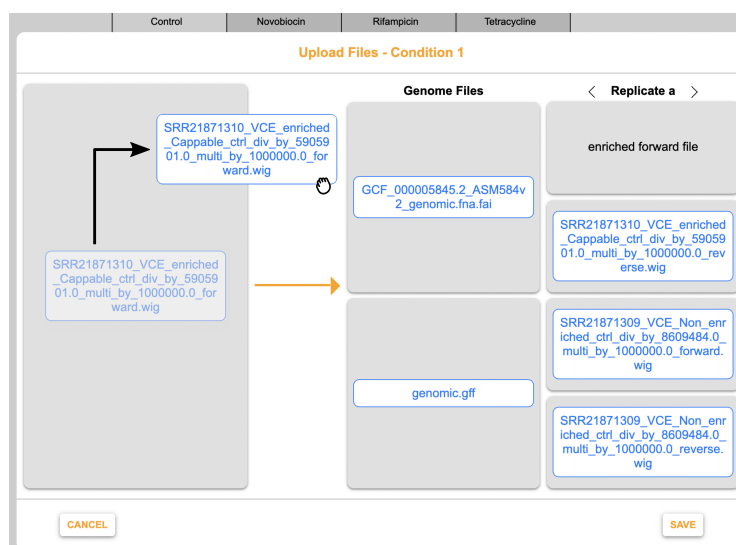


Figure II.S3: Screenshot of *drag & drop* process for file distribution. Instead of selecting files individually, they can be uploaded at once and distributed across the different conditions and replicates.

Use case

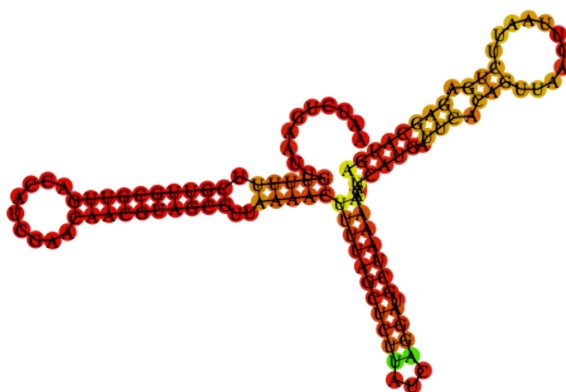


Figure II.S4: Secondary structure of the UTR region of the gene *ugd* as predicted by RNAfold.

Appendix III

TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts

Mathias Witte Paz, Thomas Vogel, Kay Nieselt

Published in *NAR Genomics and Bioinformatics*, Volume 6, Issue 4, December 2024, lqae168, DOI: [10.1093/nargab/lqae168](https://doi.org/10.1093/nargab/lqae168)

Abstract

RNA-seq and its 5'-enrichment methods for prokaryotes have enabled the precise identification of transcription start sites (TSSs), improving gene expression analysis. Computational methods are applied to this data to identify TSSs and classify them based on proximal annotated genes. While some TSSs cannot be classified at all (orphan TSSs), other TSSs are found on the reverse strand of known genes (antisense TSSs), but are not associated with the direct transcription of any known gene. Here, we introduce **TSS-Captur**, a novel pipeline, that uses computational approaches to characterize genomic regions starting from experimentally confirmed, but unclassified TSSs. By analyzing TSS data, **TSS-Captur** characterizes unclassified signals, complementing prokaryotic genome annotation tools. TSS-Captur categorizes extracted transcripts as either messenger RNA (mRNA) for genes with coding potential, or non-coding RNA (ncRNA) for non-translated genes. Additionally, it predicts the transcription termination site for each putative transcript. For ncRNA genes, the secondary structure is computed. Moreover, all putative promoter regions are analyzed to identify enriched motifs. An interactive report allows seamless data exploration. We validated **TSS-Captur** with a *Campylobacter jejuni* dataset and characterized unlabeled non-coding RNAs in *Streptomyces coelicolor*. **TSS-Captur** is available both as a web-application and as a command-line tool.

Keywords: Transcriptomics, Genomics, Transcriptome Architecture, Transcript Characterization



Introduction

In molecular biology and genetics, the identification of transcription start sites (TSSs) can provide valuable insights into the regulation of gene expression in prokaryotic organisms. For example, it can aid researchers to understand the mechanisms controlling transcription initiation by allowing the identification of nearby transcription factor binding sites (Sorek and Cossart [2010]) and hence provide new insights into gene regulation. Differential RNA-seq (dRNA-seq) (Sharma *et al.* [2010]) and Cappable-seq (Ettwiller *et al.* [2016]) are enrichment methods for RNA-seq that allow the identification of TSSs in prokaryotes. These methods modify the RNA libraries prior to sequencing, allowing for the base-specific identification of the 5'-end of expressed genes. Such experiments typically result in hundreds to thousands of TSSs, making a manual determination cumbersome. To facilitate the process of TSS identification, computational methods have been developed in the past years.

TSSpredator (Dugar *et al.* [2013]) is one such tool that determines TSS signals from experimental data. It then associates the identified TSSs with annotated genes based on their proximity and strand. Studies that have generated genome-wide TSS maps across various bacterial genomes have led to important insights into gene regulation and expression. One study, a genome-wide transcriptome analysis of *Helicobacter pylori*, uncovered regulatory mechanisms within the 5' untranslated regions (5'-UTR) of annotated genes (Bischler *et al.* [2015]). A more recent study reported a genome-wide TSS map for *Bacteroides thetaiotaomicron* (Ryan *et al.* [2020]). Besides providing insight into the vast transcription regulation mechanisms of the organism, this study reported around 269 potential non-coding RNA candidates, including *GibS*, a *trans*-acting RNA gene involved in bacterial metabolism regulation. In these and many other studies, numerous TSS signals were not found close to any annotated gene. Hence, such TSSs were typically reported as orphans (oTSSs). Moreover, other TSSs were located on the antisense strand within the coordinates of an annotated gene (thus typically called antisense TSSs, aTSSs), but have no annotated gene in the direct downstream region.

Genome annotation pipelines, such as **Bakta** (Schwengers *et al.* [2021]), **Prokka** (Seemann [2014]), and **PGAP** (Tatusova *et al.* [2016]), primarily focus on identifying sequences with coding potential, located within messenger RNA (mRNA) transcripts, as well as housekeeping non-coding RNA (ncRNA) genes. Other ncRNA genes are often overlooked due to their diverse structural patterns. In addition, the complexity also increases, since ncRNA genes with low sequence similarity may fold into the same or very similar structures. This structural and sequence heterogeneity of ncRNA genes makes their computational prediction challenging, especially in prokaryotes (Barquist *et al.* [2016]; Seemann [2014]; Stazic and Voß [2016]). Nevertheless, prokaryotic ncRNA genes play important functional roles, such as the *cis* or *trans* action of small RNAs (Carvalho Barbosa *et al.* [2020]). Orphan and antisense TSSs are promising signals that may indicate the presence of overlooked genes, such as these ncRNAs. Moreover, regions starting at these sites may also transcribe previously unidentified protein-coding genes, such as small open reading frames (sORFs), which have proven equally challenging to characterize (Gray *et al.* [2022]; Orr *et al.* [2020]). Thus, an exploratory tool designed to automatically characterize the transcripts starting at these sites as well as their respective upstream regions would offer valuable insights into the prokaryotic transcriptome.

To enable the characterization of these TSS classes and their downstream regions in an explorative manner, we introduce **TSS-Captur**, a TSS-characterization pipeline for unclassified RNA-transcripts. **TSS-Captur** combines the results of transcriptomic

experiments with the underlying genomic sequence, allowing the characterization of putative transcripts, starting with orphan and antisense TSS signals predicted using **TSSpredator**. For the overall characterization, **TSS-Captur** integrates different tools. **TSS-Captur** determines whether a transcript beginning at an oTSS or aTSS transcribes into a messenger RNA (protein-coding) or into a non-coding RNA (non-coding gene). For this, it combines two well-established prediction methods: a comparative genomics approach (**QRNA** (Rivas and Eddy 2001)) and an *ab initio* method (**CNIT** (Guo *et al.* 2019)). Furthermore, **TSS-Captur** associates TSS signals with predicted transcription termination sites to compute the length of the transcript. It also provides the visualization of the computed secondary structure of ncRNA genes, enabling the user to explore the thermodynamic characteristics (i.e., folding potential) of such genes. Lastly, **TSS-Captur** facilitates the analysis of motifs within the regions upstream of each TSS signal for the identification of transcription factor binding sites or other regulatory elements. All these results are presented in an interactive report that offers both an overview of the predicted results and detailed information for each transcript. Implemented as a stand-alone command tool and as an interactive web tool, **TSS-Captur** offers a user-friendly complement to existing annotation methods, aiming to enhance understanding of the bacterial transcriptome.

Related Work

The accuracy and completeness of a genome annotation is a key part for multiple computational analyses, since it provides a baseline for the functional potential of an organism (Schwengers *et al.* 2021). Typical prokaryotic genome annotation pipelines, such as **Bakta**, **Prokka** or **PGAP**, not only lack the annotation of 5' and 3' UTRs, but also of small ncRNAs (Schwengers *et al.* 2021; Seemann 2014; Tatusova *et al.* 2016). Other approaches, such as **Promotech** (Chevez-Guardado and Peña-Castillo 2021) or **G4promfinder** (Di Salvo *et al.* 2019), complement these predictions by characterizing promoter regions. However, the diverse regulatory mechanisms in prokaryotes present challenges in generalizing these methods. Adding the transcriptomic layer of information can provide more precise or even potentially new results, such as refining a transcript's structure or identifying previously unannotated expressed regions. Tools such as **APERO** (Leonard *et al.* 2019) and **baerhunter** (Ozuna *et al.* 2020) integrate transcriptomic data with genome data by using mapped RNA-seq reads to identify expressed regions that are not proximal to annotated genes, thereby identifying potential small non-coding genes. However, RNA-seq data alone has been proven to be insufficient for an accurate base-exact prediction of gene boundaries (Bischler *et al.* 2015). Instead of relying on RNA-seq data alone, **ANNOgesic** (Yu *et al.* 2018) uses dRNA-seq data to identify TSSs and their corresponding transcripts. **ANNOgesic** offers a plethora of different methods for enhancing bacterial genome annotations, including the prediction of small RNAs (sRNA), sORF, circular RNAs, CRISPR related RNAs, riboswitches and RNA-thermometers. However, **ANNOgesic** lacks a user-friendly interface for executing and exploring results, making it difficult to confirm and further investigate its predictions. Our approach, **TSS-Captur**, complements **ANNOgesic** by providing an intuitive web-based platform for exploration.

Material & Methods

TSS-Captur is a pipeline that characterizes genomic regions starting with TSS signals that cannot be associated with any known annotated gene (orphan or antisense TSSs). It is based on **Nextflow DSL2** (Di Tommaso *et al.* 2017) for the stage management and

III. TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts

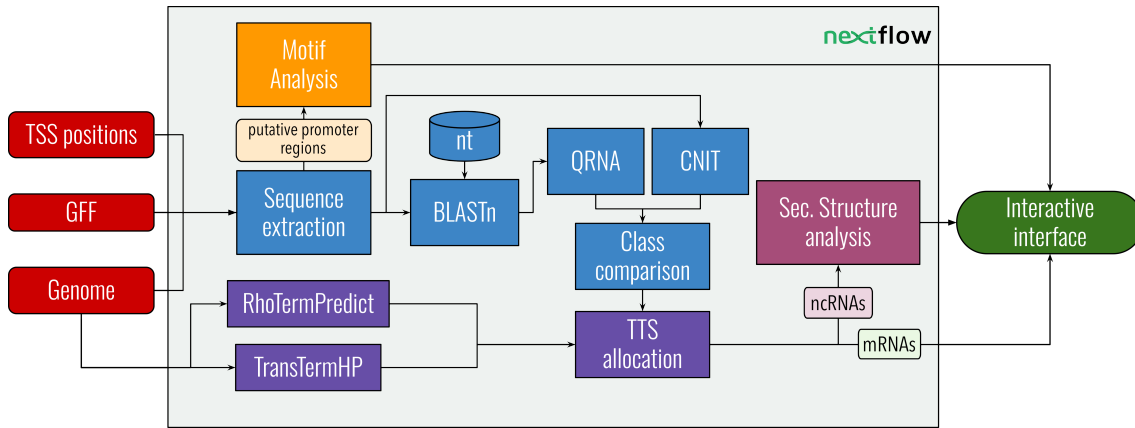


Figure III.1: Overview of **TSS-Captur**. Different shapes have been used to represent data (ovals), processes (rectangles) or databases (cylinders). The input files (red) are provided to **Nextflow** for the data preparation. This provides the input to the RNA classification (blue) and the motif analysis (yellow). The transcription termination site (TTS, purple) prediction works directly on the genome and the results are combined with the classified transcripts. Here, transcripts classified as ncRNAs are analyzed for their secondary structure (magenta). Lastly, all results are presented in the interactive interface (green).

Docker for the containerization of all required tools. To facilitate the user experience, **TSS-Captur** can be used as a web-application (Supplementary Figure III.S1), and results can be explored through an interactive web-interface (Supplementary Figure III.S2). The web-application is based on **NextJS**, while the report is created by **Frozen-Flask** and the JavaScript library **DataTables**.

We also provide a stand-alone command-line version of the software. In the ensuing sections, we provide a short overview of the **TSS-Captur** pipeline (see Fig. III.1), followed by detailed explanations of each step.

TSS-Captur requires three different input files: the **MasterTable** from the TSS determination and classification process run by **TSSpredator**, as well as the genome and annotation files used for the computation of the TSSs. **TSS-Captur** first filters all TSSs to retain those classified as oTSSs or aTSSs from the **MasterTable**, extracts the respective downstream genomic region and classifies it as representing either an mRNA or ncRNA. Two different methods are used for this step: **CNIT** (Guo *et al.* 2019), a method based on support-vector machines (SVMs), as well as **QRNA** (Rivas and Eddy 2001), a comparative sequence analysis requiring a pairwise alignment. Furthermore, the tools **TransTermHP** (Kingsford *et al.* 2007) and **RhoTermPredict** (Di Salvo *et al.* 2019) are used to predict genome-wide transcription termination sites (TTS). While **TransTermHP** predicts intrinsic terminators via the identification of hairpin loops (Kingsford *et al.* 2007), **RhoTermPredict** identifies Rho-dependent sites via the identification of Rho-utilization (RUT) sites and RNA polymerase (RNAP) pause sites (Di Salvo *et al.* 2019). Based on these predictions, a putative TTS (either intrinsic or Rho-dependent) is allocated for each potential transcript. In the case of transcripts classified as ncRNA genes, **TSS-Captur** computes their secondary structure using **RNAfold** (Gruber *et al.* 2008). Furthermore, the putative promoter regions of all transcripts are analyzed using **MEME** (Bailey and Elkan 1994) to identify enriched motifs. All results are accessible via an interactive interface.

Data Preparation

After obtaining only orphan and antisense TSSs from the **MasterTable**, sequences downstream of these TSSs are extracted from the genome. The mean gene length (μ_G) is calculated for all entries in the annotation file and used as the transcript length's upper limit, capped at a maximum of 1000bp to fulfill the limitations of **QRNA**. In addition to using μ_G as a transcript limit, **TSS-Captur** also stops transcript extraction if any specified genomic feature is found within the extraction range. Such features include any TSS listed in the **MasterTable** (either primary, secondary, antisense or orphan) or the start of an entry in the annotation file. By stopping extraction at any TSS, **TSS-Captur** minimizes the impact of regions related to expressed genes, such as their 5'-UTR regions, on the characterization of the overlooked transcripts. Moreover, this step also prevents the transcripts from overlapping with known but unexpressed genes. Consonant with the recommendation of Rivas and Eddy (2001), all extracted regions shorter than 30 nt are removed from the analysis together with their TSS (Rivas and Eddy [2001]; Sorek and Cossart [2010]). After extraction, each transcript is saved in a separate **FASTA**-file, along with its TSS class (oTSS or aTSS). To ensure consistency in reading direction (5'-3'), the complement of reverse strand transcripts is computed.

Classification of transcripts

For the characterization of transcripts, **TSS-Captur** uses the comparative genomic tool **QRNA** and the SVM-based *ab initio* tool **CNIT** (Guo *et al.* [2019]; Rivas and Eddy [2001]). Running both classifiers strengthens **TSS-Captur**'s prediction by accounting for different sequence characteristics and evolutionary aspects of the classification. **CNIT** requires only the transcript sequence as input and offers two pre-trained models: one trained with animal data and the other with plant data. To our knowledge, no similar tool has been developed for prokaryotes. To ensure the appropriate application of **CNIT** to prokaryotic data, both models were tested on a subset of annotated genes (see *Verification and Use Case* below for details). Unlike **CNIT**, **QRNA** relies on a pairwise alignment. To determine the most appropriate alignment for each sequence, **TSS-Captur** uses **BLAST** and the **nt** database (Altschul *et al.* [1990]). A recommended optimal evolutionary distance between sequences, which conserves structure (for ncRNAs) and still returns plausible alignments, is within the range 65-85% (Barquist *et al.* [2016]; Rivas and Eddy [2001]). To achieve such a pairwise alignment, the sensitive discontinuous **dc-Megablast** search strategy of **BLAST** is employed, since it mainly returns hits with 80% similarity (McGinnis and Madden [2004]). To address the long runtime and the high sensitivity of this search approach, **TSS-Captur** reduces the search space in the database to consider only species within the same genus as the analyzed organism. This is achieved by using the organism's accession code and the **ETE toolkit** (Huerta-Cepas *et al.* [2016]) library to query the **NCBI Taxonomy**. The hits reported by **BLAST** are evaluated according to the following criteria: (1) the sequence similarity is within the optimal range and (2) the coverage of the transcript is maximized, prioritizing conservation at the 5'-region of the query (see Supplementary Section B for details). For each transcript, the best alignment is stored in a **FASTA**-file for usage in **QRNA**.

The outputs of both tools are compared to compute a final call for each transcript. Both **CNIT** and **QRNA** provide a class prediction, a prediction score, and the best functional genomic subregion (e.g., the predicted coding-sequence region for an mRNA transcript). **TSS-Captur** computes a class-specific *z*-score normalization to enable a comparison of the two methods' prediction scores. In case of agreement, the output of the best-scoring program is used. In the case of class disagreement of the two tools, **TSS-Captur** proceeds

III. TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts

as follows:

- If either of the tools predicts a coding gene (i.e. mRNA), and the other an ncRNA, the mRNA class is chosen (see *Verification* for details).
- If QRNA returns no prediction due to a missing alignment or due to choosing the null model (*OTH*), CNIT's prediction is used.

Since both tools also return the coordinates of the most probable functional subsequence, a putative 3'-end for each transcript is computed during this step. However, this is refined by identifying a nearby putative TTS in the following step.

Transcription termination sites

TSS-Captur makes its termination site prediction in a genome-wide manner using TransTermHP (Kingsford *et al.* 2007) in the case of intrinsic termination and RhoTermPredict (Di Salvo *et al.* 2019) for Rho-dependent termination. TransTermHP is run using nocoRNAC (Herbig and Nieselt 2011) as a wrapper, since it is optimized for whole-genome analysis and directly filters terminators with low confidence scores (below 75). Both tools report annotation files (GFF-files) with predicted genome-wide termination regions and their corresponding confidence score.

After the aforementioned transcript classification, TSS-Captur associates transcripts with potential terminators. For each transcript, all predicted TTS starting within the coordinates defined by QRNA or CNIT are considered. Each putative TTS is evaluated by considering its prediction score and its distance to the 3'-end predicted during classification (see Supplementary Section C for details). The best-scoring TTS is taken for each transcript.

The transcript's end is then adjusted according to the type of terminator identified. For intrinsic terminators, the end of the hairpin loop marks the TTS. In Rho-dependent cases, the RNAP pause site (150 nt downstream of the RUT site) is analyzed, as it may contain either a hairpin or only a RNAP-pause site (Henkin 1996). If a hairpin structure is detected, both the RUT site and hairpin are included in the termination region, with the TTS placed directly behind the hairpin. If no hairpin is found, the end of the RNAP pause site is considered as the transcript's end. This process defines a TTS and a transcript length (TTS position – TSS position) for each transcript.

Secondary structure prediction

For each transcript classified as an ncRNA, RNAFold (Lorenz *et al.* 2011) is used to compute the secondary structure with the minimum free energy (MFE). This RNA structure is saved as a JPG-file and linked in the interactive report.

Promoter region analyses

The putative promoter region of each accounted TSS is analyzed using MEME (Bailey and Elkan 1994; Bailey, Boden, *et al.* 2009) to identify over-represented motifs. Such enriched motifs could be either novel regulatory elements or also known transcription factor binding sites, such as the Pribnow-Box or the -35 box. Identifying such motifs can strengthen the validity of the TSS signals. For this analysis, TSS-Captur extracts the 50bp upstream region of each TSS to be passed to MEME, which identifies a user-defined number of motifs ranging from 5 to 20bp in length. As long as no more than 1,000 TSSs are analyzed in one batch, MEME will also provide the start of each motif in each sequence, which is essential for identifying common regulation motifs. Results are provided by MEME in a user-friendly

Table III.1: Evaluation of **CNIT**’s performance on *C. jejuni*’s subset. Each model was evaluated by computing the sensitivity, specificity and precision concerning the annotated RNA class.

Model	Type	Sen.	Spec.	Prec.
Plant	mRNA	0.92	1	1
	ncRNA	1	0.92	0.926
Animal	mRNA	0.82	1	1
	ncRNA	1	0.82	0.847

HTML report and as an XML-file. The XML output is further converted into a TSV-file for easier data manipulation. Both formats are included in **TSS-Captur**’s final output.

TSS-Captur Result Interface

To streamline exploration of **TSS-Captur**’s output, a comprehensive HTML report summarizes the results in one interface (see Supplementary Figure III.S2), via interactive tables for each characterization step. The *overview* page offers a summary of analyzed and discarded TSSs, including their classification as either orphan or antisense. The central table on this page merges important data: TSS position, strand, predicted class (mRNA/ncRNA), promoter motifs, among other information. Moreover, the page integrates the visualization of the secondary structure for each transcript predicted as ncRNA. The pages on the *terminator allocation* and *classification* provide in-depth information on the corresponding processes. For example, the *classification* page allows users to trace the rationale behind each transcript’s assigned class. The **MEME** HTML report is directly integrated in **TSS-Captur**’s report, allowing a detailed inspection of promoter motifs and leveraging all of **MEME**’s existing visualization advantages, as well as the other tools from the **MEME** Suite (e.g., **TOMTOM** (Bailey, Boden, *et al.* 2009)). Finally, the *ignored TSS* page lists TSSs excluded by **TSS-Captur** due to their short length.

Results

Verification of TSS-Captur

To assess the overall performance of **TSS-Captur**, we used the expression data for *Campylobacter jejuni* str. NCTC 11169 published by Dugar *et al.* (Dugar *et al.* 2013) to compute a *MasterTable* using **TSSpredator**, with default parameters. The annotation file (RefSeq ID NC_002163) was used to classify the TSSs with respect to annotated genes. From this set, we selected a subset of 100 primary TSSs (pTSS) and their corresponding annotations, equally balancing both RNA classes (ncRNA and mRNA).

First, we extracted the transcripts using the exact coordinates of each annotated gene of the subset to assess the two models of **CNIT** (animal-based and plant-based) with respect to sensitivity, specificity and precision. Both models demonstrated high sensitivity, specificity, and precision (all ≥ 0.82 ; see Table III.1), indicating that the use of **CNIT** is suitable for prokaryotes. However, the *plant* model delivered better results (≥ 0.92 ; Table III.1) and is therefore used subsequently.

Afterwards, the influence of the extracted transcript’s length on the overall RNA classification implemented in **TSS-Captur** was assessed. For each TSS and its annotated transcript end l , multiple regions with varying end l' were extracted, with $l' = l + t$ and $t = 5i, i \in [-20, 20]$. To achieve a larger dataset, we lowered the 30 nt extraction threshold to 5 nt. Based on the RNA class (mRNA or ncRNA) from the annotation

III. TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts

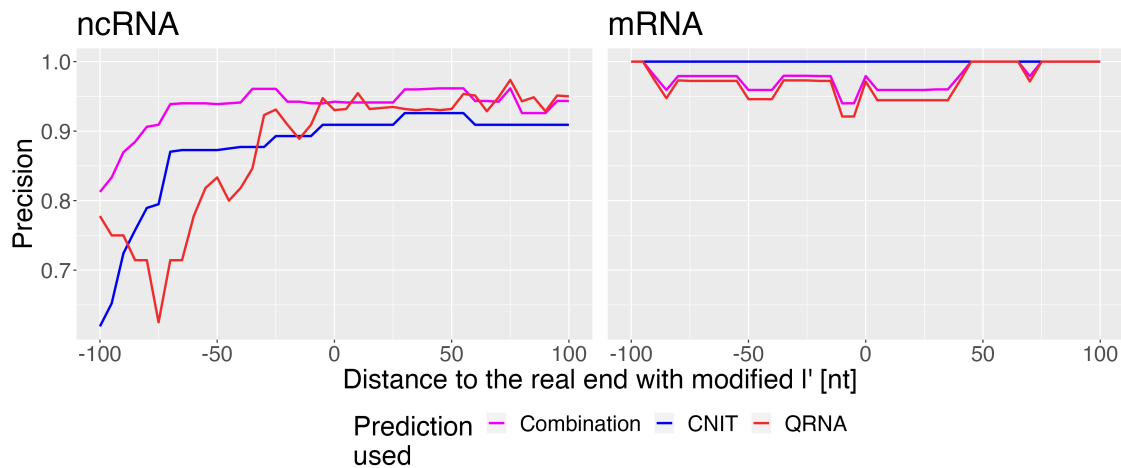


Figure III.2: Evaluation of RNA classification. The three different types of predictions are shown: CNIT, QRNA and the combination implemented in TSS-Captur. For each prediction type and for each modified l' , the precision was computed for each type of transcript (ncRNA and mRNA).

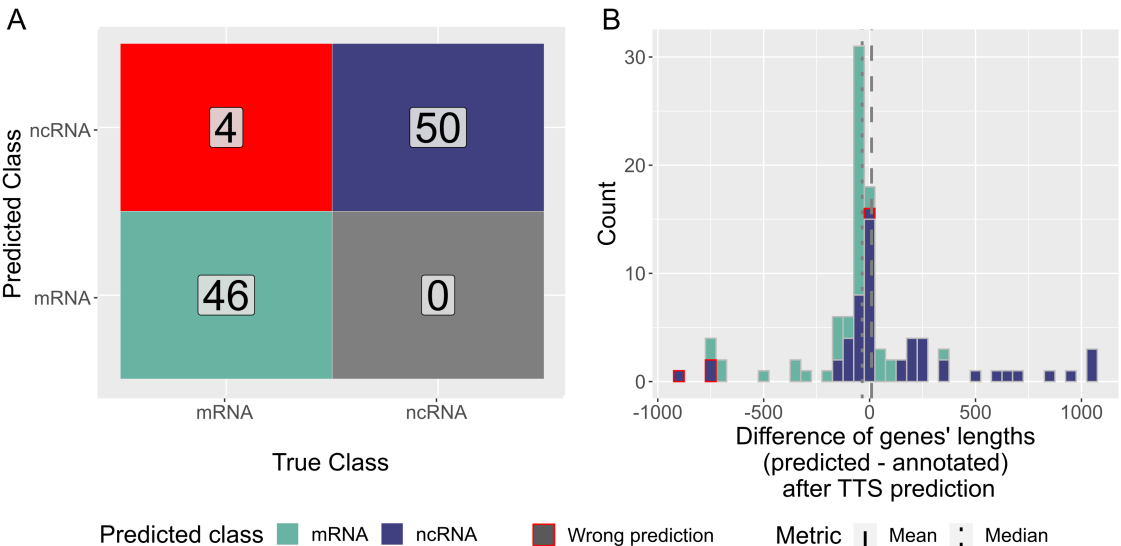


Figure III.3: Evaluation of TSS-Captur by treating pTSSs as oTSSs. (A): Confusion matrix on the real and predicted RNA classes for the transcripts. (B): Distributions of the genes' differences (predicted - annotated) by predicted RNA class (bin width = 50 nt). The mean and median differences are also shown. Wrongly predicted RNAs are highlighted (red).

file, we computed the precision for the three classification approaches (CNIT, QRNA and TSS-Captur's resulting decision of the combination of both methods). This analysis revealed that by combining both approaches as implemented in TSS-Captur, the highest precision (within the range 0.8-0.95) across most of the transcripts for ncRNA genes was achieved, while CNIT's precision was the highest for mRNA genes across all lengths (Fig. III.2). Still, the combination implemented in TSS-Captur returned also precise results for mRNAs (> 0.95) across all transcripts' length.

Lastly, we modified the *MasterTable* to assess the whole process behind TSS-Captur. We removed all oTSSs or aTSSs from the *MasterTable* to avoid their analysis in TSS-Captur. The entries in the *MasterTable* and the annotation file for the 100 pTSSs of the subset

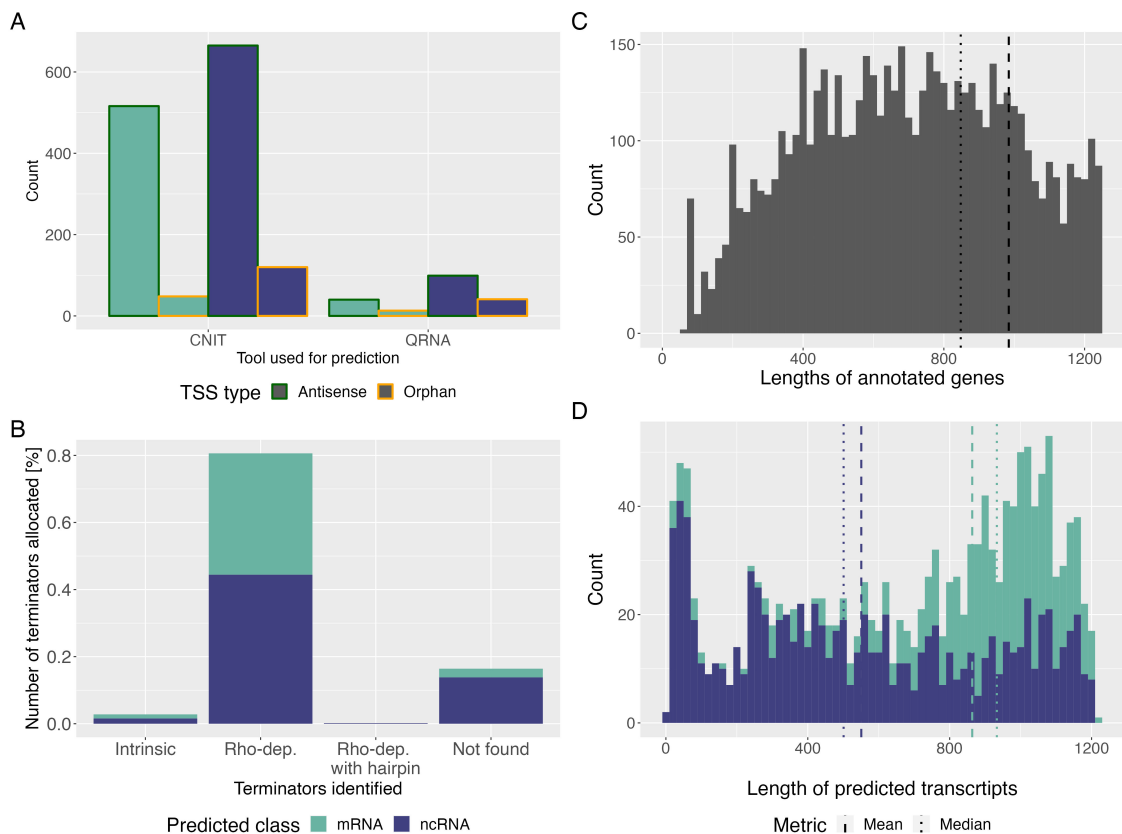


Figure III.4: Visual summary of TSS-Captur output for *S. coelicolor*'s data. (A): Summary of the predicted RNA classes for each transcript by the tool used for prediction and the original TSS type. (B): Distribution of the allocated terminator regions, wrt. their type. (C & D): Histogram (bin width = 20) of the length for annotated genes (C) and for predicted transcripts (D). Only annotated genes shorter than 1250 nt are shown for a better comparison with the predictions.

were modified to simulate orphan TSSs. The remaining TSSs were left unmodified to have a feasible *MasterTable*. The output of TSS-Captur was compared with the original classes and lengths. All ncRNA genes were correctly classified, and of the 50 mRNA genes, only four were incorrectly classified as ncRNAs (Fig. III.3A). For the prediction of the gene's length, the difference between the predicted and the annotated length was computed, returning a mean difference of 9.2 nt and a median of -35 nt (Fig. III.3B). 57 genes had an absolute length difference ≤ 100 nt, while 74 genes had an absolute length difference ≤ 250 nt. For those 18 genes with an absolute length difference ≥ 500 nt, a clear difference in the classes was observed. On the one hand, eight coding genes were strongly underestimated in their length. On the other hand, nine ncRNAs were greatly overestimated in their length.

Use Case: *Streptomyces coelicolor*

To demonstrate the applicability of TSS-Captur, we analyzed a genome-wide TSS dataset of *Streptomyces coelicolor* (Jeong *et al.* 2016). For this, the expression data was analyzed using TSSpredator with sensitive parameters and an annotation file for the organism (RefSeq ID NC_003888.3). This resulted in a total of 1,660 antisense and orphan TSSs that were passed to TSS-Captur as input, together with the genome sequence and annotation file.

118 of these TSSs were filtered out, since they would have produced transcripts of a

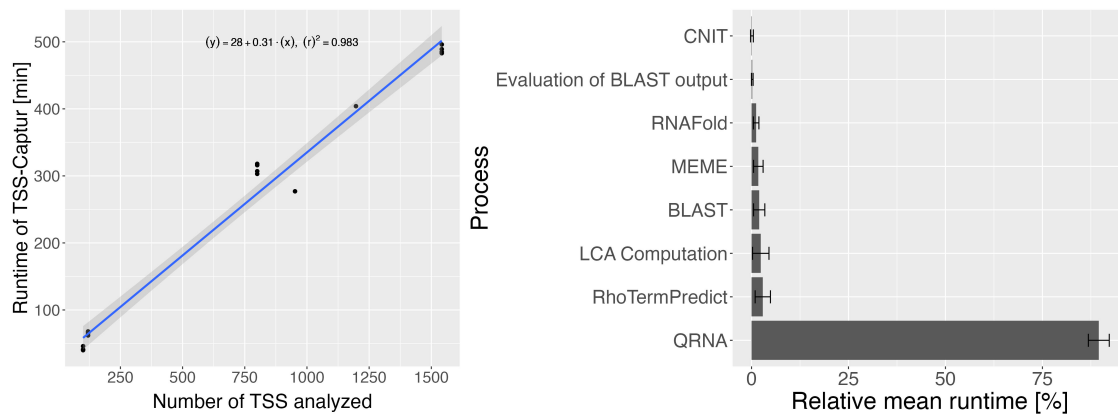


Figure III.6: Runtime analysis of **TSS-Captur**. The runtime of 18 analyses is visualized together with the computed linear regression line (left). The mean of the runtime as well as the standard deviation of each process was computed and shown here in % relative to the overall runtime (right). Processes that were faster than one minute (e.g., **TransTermHP**) are not shown.

Runtime analysis

The runtime of **TSS-Captur** was assessed using multiple datasets (in total, 18 different runs were conducted) including the conduct of all genome-wide analyses (termination site predictions, promoter identification, *etc.* It was performed on a server with an Intel(R) Xeon(R) E5-4610 v2 @ 2.30GHz CPU using at most 4 cores. The results show a clear linear dependency on the number of TSSs (Figure III.6, left) with an increase in the runtime by 18.6s per TSS. In all runs, the most time-consuming step is the **QRNA** analysis (89.5% of total; see Figure III.6, right).

Discussion

Genome-wide base-resolved transcriptome maps offer insights into regulation regions of annotated genes (e.g., 5'-UTRs, promoters) (Bischler *et al.* 2015; Ryan *et al.* 2020). However, they also reveal strong transcription signals in unannotated regions – a rich source of potentially undiscovered genes. Our exploration tool, **TSS-Captur**, unlocks the potential of these sites, enabling user-friendly analysis that combines genomic and transcriptomic data for their characterization.

The main back-end component of **TSS-Captur** is a pipeline implemented using **Nextflow**. Among the advantages of **Nextflow** are its declarative syntax, as well as modularity and extensibility, which facilitate multiple aspects of the identification and characterization of TSSs.

For the classification of previously uncharacterized transcript sequences, both intrinsic genomic features (**CNIT**) and evolutionary conservation (**QRNA**) are considered. Similarly, the transcription termination site prediction combines complementary methods (**TransTermHP** & **RhoTermPredict**) for a broad analysis. The analysis of secondary structures and promoter regions offers insights into the function of the transcript as well as its regulation mechanisms. A unified report facilitates exploration of the results.

To validate our approach, we used a dataset based on primary TSSs in *Campylobacter jejuni*. We analyzed the precision of the RNA classification across varying transcript lengths, demonstrating that our combined approach provides a balanced performance for both gene classes. By simulating orphan TSSs, we evaluated the overall performance of

III. TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts

TSS-Captur. The results showed accurate classification for all ncRNAs and almost all protein-coding genes, while lengths' predictions vary in accuracy across genes.

Our analysis of the *S. coelicolor* TSS dataset showcases the utility of **TSS-Captur** in exploring unannotated genomic regions. Of particular interest is the identification of a highly enriched G/C motif upstream of many of the unannotated transcripts. While no direct matches were found using **TOMTOM**, this motif might be associated with G-quadruplex regulation mechanisms known to operate in *S. coelicolor* (Di Salvo *et al.* 2018). This finding demonstrates how **TSS-Captur** facilitates hypothesis generation. Via the report of **TSS-Captur**, other tools from the **MEME Suite** (Bailey, Boden, *et al.* 2009) are available to the user to further investigate identified enriched motifs within the promoter regions.

TSS-Captur's predictions exhibited some overlap with the experimentally validated ncRNAs from Vockenhuber *et al.* 2011. Here, it is important to remark that the experimental conditions between the analyzed TSSs with **TSS-Captur** and the study of Vockenhuber *et al.* differed, hence not all transcripts might be present. This, together with our choice of using the TSS prediction of **TSSpredator** with the most sensitive parameters, explains the rather low rate of commonly called TSSs. Nonetheless, the confirmation with the study, along with the **RFAM** validation of some predicted ncRNAs, emphasizes how **TSS-Captur** can be used to detect novel ncRNAs.

Both **TSS-Captur** and **ANNogesic** use a number of different backend tools for transcript characterization. While **ANNogesic** currently offers a larger range of methods than **TSS-Captur**, addressing additional aspects of transcriptomic data annotation, it also has some limitations. **ANNogesic's** reliance on existing databases and known mechanisms (like the Shine-Dalgarno sequence) for transcript characterization may limit the discovery of novel genes. In contrast, **TSS-Captur's** classification relies on more abstract characteristics of the transcripts, with the potential for uncovering previously unknown mechanisms of transcription. Furthermore, by including **RhoTermPredict**, **TSS-Captur** accounts for Rho-dependent termination, a well-known mechanism in prokaryotes (Henkin 1996). Moreover, **ANNogesic's** command-line interface and lengthy documentation pose challenges for users lacking extensive computational skills. One motivation to develop **TSS-Captur** was to provide users with a user-friendly and accessible tool. Our web-interface simplifies file parsing, eliminating the need for inexperienced users to engage in direct command-line interaction. Also, **TSS-Captur** unifies results within a single interface, offering data exploration alongside convenient access to secondary structure visualizations and external analysis options, such as **MEME**.

Due to the flexibility of **Nextflow**, new modules, tools or even other data sources can be easily included in **TSS-Captur**. One issue that we identified in our two use cases is the prediction of the 3'-end of transcripts. In particular, the lengths of predicted ncRNAs often seem to be overestimated. This was revealed by the validation study, as well as our observation that **TSS-Captur's** predicted transcripts often aligned only with subsequences within the **RFAM** entries. In the future, we may consider integrating more recent tools or other data sources for the prediction of the 3'-end, such as **TermNN** (Brandenburg *et al.* 2022) or Term-seq data (Dar *et al.* 2016).

For the classification step of uncharacterized transcript regions, we used two complementary methods and thereby provide a general idea of the transcript's function (either mRNA or ncRNA). Nonetheless, there remains an opportunity to advance the classification of prokaryotic transcripts. For instance, new methods for intrinsic classification, specifically trained with prokaryotic data, could be developed. These approaches could yield even more precise predictions about transcript classes, though one must account for the fact that sequence homology-based classification typically offers

higher precision, especially as the number of sequences increases (Barquist *et al.* [2016]; Fu *et al.* [2013]). This is particularly true when intrinsic features alone are insufficient for accurate classification, as it is often the case with short-length sORFs (Samayoa *et al.* [2011]). Moreover, improvements are possible not only in classification but also in the identification of the most likely secondary structures for ncRNA genes. By considering homologs with similar, though not identical, primary structures, tools like **Infernal** can incorporate this possibility of structural prediction (Nawrocki and Eddy [2013]). However, the identification of the relevant set of homologs is not trivial and often requires manual intervention, making such an approach difficult to automate (Barquist *et al.* [2016]).

With respect to the classification, it needs to be acknowledged that **TSS-Captur** defines two distinctive classes (mRNA and ncRNA) for a transcript. This strict classification might not be adequate for all transcripts, such as *dual-function* small RNAs (Raina *et al.* [2018]; Schnoor *et al.* [2024]). This lesser known class of genes consists of two components: a small non-coding RNA and a small open-reading frame component. The limited number of relevant prokaryotic examples hinders the generalization of identification methods for this class of genes (Schnoor *et al.* [2024]). Nevertheless, any future development that provides a clearer classification for such transcripts can be easily integrated in **TSS-Captur** due to the **Nextflow**-based structure.

In summary, starting from experimental TSS data, **TSS-Captur** predicts the characterization of unclassified signals and allows a user-friendly exploration to complement prokaryotic annotation tools, contributing to the understanding of bacterial transcriptomes. By combining a user-centric design, **TSS-Captur** empowers researchers with diverse technical backgrounds to characterize experimentally detected transcription signals and eventually accelerate the discovery of novel genomic elements and their regulatory mechanisms.

Availability

TSS-Captur can be accessed at <https://tsscaptor-tuevis.cs.uni-tuebingen.de/>. The source code is available via **Zenodo** at <https://zenodo.org/doi/10.5281/zenodo.12527008>. The datasets used for this study are also available online at <https://zenodo.org/doi/10.5281/zenodo.12526907>. All tools required for the pipeline have been containerized into a Docker image and are publicly available at <https://hub.docker.com/r/mwitte/tsscaptor>

Competing interests

No competing interest is declared.

Author contributions statement

MWP and KN conceived the idea for **TSS-Captur**. MWP implemented the first pipeline version and conducted the experiments. TV implemented the **DSL2** version and the web application. All authors wrote and reviewed the manuscript.

Acknowledgements and Funding

We thank Peter Wills for his valuable feedback on the manuscript. We acknowledge infrastructural funding from the Cluster of Excellence EXC 2124 ‘Controlling Microbes to Fight Infections’ [project ID 390838134] from the German Research Foundation (DFG). We acknowledge support from the Open Access Publishing Fund of the University of Tübingen.

References

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). “Basic local alignment search tool”. In: *Journal of Molecular Biology* vol. 215, no. 3, pp. 403–410. ISSN: 0022-2836. DOI: [10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2). URL: <https://www.sciencedirect.com/science/article/pii/S0022283605803602>.
- Bailey, T. L. and Elkan, C. (1994). “Fitting a mixture model by expectation maximization to discover motifs in biopolymers”. In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology* vol. 2, pp. 28–36. ISSN: 1553-0833.
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). “MEME Suite: tools for motif discovery and searching”. In: *Nucleic Acids Research* vol. 37, no. suppl_2, W202–W208. ISSN: 0305-1048. DOI: [10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335).
- Barquist, L., Burge, S. W., and Gardner, P. P. (2016). “Studying RNA Homology and Conservation with Infernal: From Single Sequences to RNA Families”. In: *Current Protocols in Bioinformatics* vol. 54, no. 1. ISSN: 1934-340X. DOI: [10.1002/cpbi.4](https://doi.org/10.1002/cpbi.4).
- Bischler, T., Tan, H. S., Nieselt, K., and Sharma, C. M. (2015). “Differential RNA-seq (dRNA-seq) for annotation of transcriptional start sites and small RNAs in *Helicobacter pylori*”. In: *Methods* vol. 86, pp. 89–101. ISSN: 1046-2023. DOI: [10.1016/j.ymeth.2015.06.012](https://doi.org/10.1016/j.ymeth.2015.06.012). URL: <https://www.sciencedirect.com/science/article/pii/S1046202315002546>.
- Brandenburg, V. B., Narberhaus, F., and Mosig, A. (July 2022). “Inverse folding based pre-training for the reliable identification of intrinsic transcription terminators”. en. In: *PLOS Computational Biology* vol. 18, no. 7, e1010240. ISSN: 1553-7358. DOI: [10.1371/journal.pcbi.1010240](https://doi.org/10.1371/journal.pcbi.1010240).
- Carvalho Barbosa, C., Calhoun, S. H., and Wieden, H.-J. (2020). “Non-coding RNAs: what are we missing?” In: *Biochemistry and Cell Biology* vol. 98, no. 1, pp. 23–30. ISSN: 0829-8211. DOI: [10.1139/bcb-2019-0037](https://doi.org/10.1139/bcb-2019-0037).
- Chevez-Guardado, R. and Peña-Castillo, L. (Nov. 2021). “Promotech: a general tool for bacterial promoter recognition”. en. In: *Genome Biology* vol. 22, no. 1, p. 318. ISSN: 1474-760X. DOI: [10.1186/s13059-021-02514-9](https://doi.org/10.1186/s13059-021-02514-9).
- Dar, D., Shamir, M., Mellin, J. R., Koutero, M., Stern-Ginossar, N., Cossart, P., and Sorek, R. (2016). “Term-seq reveals abundant ribo-regulation of antibiotics resistance in bacteria”. In: *Science* vol. 352, no. 6282, aad9822. DOI: [10.1126/science.aad9822](https://doi.org/10.1126/science.aad9822).
- Di Salvo, M., Pinatel, E., Talà, A., Fondi, M., Peano, C., and Alifano, P. (Feb. 2018). “G4PromFinder: an algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs”. en. In: *BMC Bioinformatics* vol. 19, no. 1, p. 36. ISSN: 1471-2105. DOI: [10.1186/s12859-018-2049-x](https://doi.org/10.1186/s12859-018-2049-x).
- Di Salvo, M., Puccio, S., Peano, C., Lacour, S., and Alifano, P. (2019). “RhoTermPredict: an algorithm for predicting Rho-dependent transcription terminators based on *Escherichia coli*, *Bacillus subtilis* and *Salmonella enterica* databases”. In: *BMC Bioinformatics* vol. 20, no. 1, p. 117. ISSN: 1471-2105. DOI: [10.1186/s12859-019-2704-x](https://doi.org/10.1186/s12859-019-2704-x). URL: <https://doi.org/10.1186/s12859-019-2704-x>.
- Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). “Nextflow enables reproducible computational workflows”. In: *Nature Biotechnology* vol. 35, no. 4, pp. 316–319. ISSN: 1546-1696. DOI: [10.1038/nbt.3820](https://doi.org/10.1038/nbt.3820). URL: <https://doi.org/10.1038/nbt.3820>.
- Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., and Sharma, C. M. (2013). “High-Resolution Transcriptome Maps Reveal Strain-Specific Regulatory Features of Multiple *Campylobacter jejuni* Isolates”. In: *PLOS Genetics*

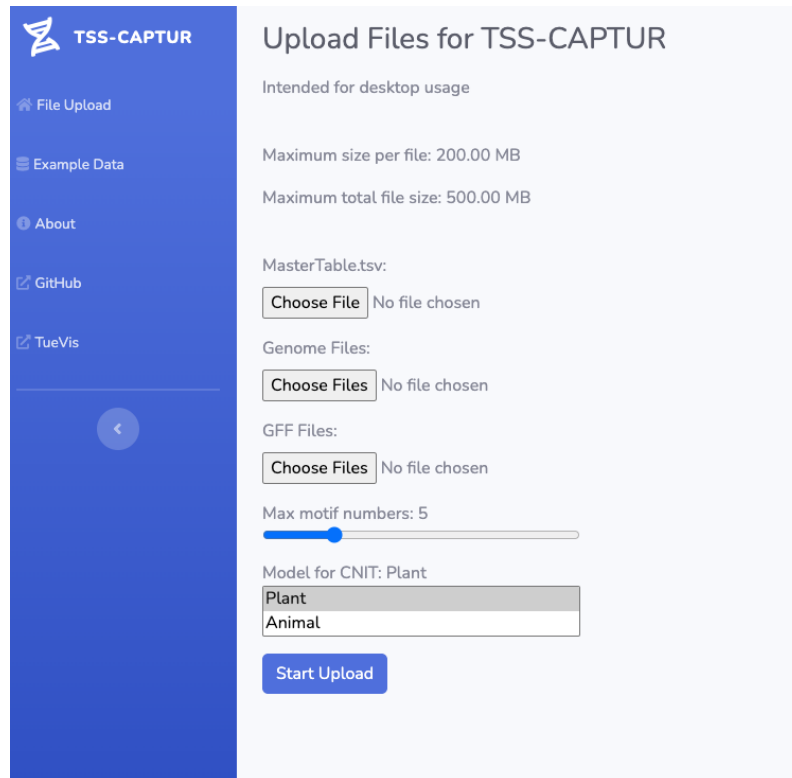
- vol. 9, no. 5, e1003495. DOI: [10.1371/journal.pgen.1003495](https://doi.org/10.1371/journal.pgen.1003495), URL: <https://doi.org/10.1371/journal.pgen.1003495>.
- Ettwiller, L., Buswell, J., Yigit, E., and Schildkraut, I. (2016). “A Novel Enrichment Strategy Reveals Unprecedented Number of Novel Transcription Start Sites at Single Base Resolution in a Model Prokaryote and the Gut Microbiome”. In: *BMC Genomics* vol. 17, no. 1, pp. 1–14. DOI: [10.1186/s12864-016-2539-z](https://doi.org/10.1186/s12864-016-2539-z).
- Fu, Y., Deiorio-Haggar, K., Anthony, J., and Meyer, M. M. (2013). “Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria”. In: *Nucleic Acids Research* vol. 41, no. 6, pp. 3491–3503. ISSN: 0305-1048. DOI: [10.1093/nar/gkt055](https://doi.org/10.1093/nar/gkt055).
- Gray, T., Storz, G., and Papenfort, K. (Jan. 2022). “Small Proteins; Big Questions”. In: *Journal of Bacteriology* vol. 204, no. 1, e00341–21. DOI: [10.1128/JB.00341-21](https://doi.org/10.1128/JB.00341-21).
- Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R., and Hofacker, I. L. (2008). “The Vienna RNA Websuite”. In: *Nucleic Acids Research* vol. 36, no. suppl_2, W70–W74. ISSN: 0305-1048. DOI: [10.1093/nar/gkn188](https://doi.org/10.1093/nar/gkn188).
- Guo, J.-C., Fang, S.-S., Wu, Y., Zhang, J.-H., Chen, Y., Liu, J., Wu, B., Wu, J.-R., Li, E.-M., Xu, L.-Y., Sun, L., and Zhao, Y. (2019). “CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition”. In: *Nucleic Acids Research* vol. 47, no. W1, W516–W522. ISSN: 0305-1048. DOI: [10.1093/nar/gkz400](https://doi.org/10.1093/nar/gkz400).
- Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., and Noble, W. S. (Feb. 2007). “Quantifying similarity between motifs”. In: *Genome Biology* vol. 8, no. 2, R24. ISSN: 1474-760X. DOI: [10.1186/gb-2007-8-2-r24](https://doi.org/10.1186/gb-2007-8-2-r24).
- Henkin, T. M. (1996). “Control of transcription termination in prokaryotes”. In: *Annual review of genetics* vol. 30, no. 1, pp. 35–57. ISSN: 0066-4197.
- Herbig, A. and Nieselt, K. (2011). “nocoRNAC: Characterization of non-coding RNAs in prokaryotes”. In: *BMC Bioinformatics* vol. 12, no. 1, p. 40. ISSN: 1471-2105. DOI: [10.1186/1471-2105-12-40](https://doi.org/10.1186/1471-2105-12-40), URL: <https://doi.org/10.1186/1471-2105-12-40>.
- Huerta-Cepas, J., Serra, F., and Bork, P. (2016). “ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data”. In: *Molecular Biology and Evolution* vol. 33, no. 6, pp. 1635–1638. ISSN: 0737-4038. DOI: [10.1093/molbev/msw046](https://doi.org/10.1093/molbev/msw046).
- Jeong, Y., Kim, J.-N., Kim, M. W., Bucca, G., Cho, S., Yoon, Y. J., Kim, B.-G., Roe, J.-H., Kim, S. C., Smith, C. P., and Cho, B.-K. (2016). “The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2)”. In: *Nature Communications* vol. 7, no. 1, p. 11605. ISSN: 2041-1723. DOI: [10.1038/ncomms11605](https://doi.org/10.1038/ncomms11605).
- Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., Griffiths-Jones, S., Toffano-Nioche, C., Gautheret, D., Weinberg, Z., Rivas, E., Eddy, S. R., Finn, R. D., Bateman, A., and Petrov, A. I. (2021). “RFAM 14: expanded coverage of metagenomic, viral and microRNA families”. In: *Nucleic Acids Research* vol. 49, no. D1. ISSN: 0305-1048. DOI: [10.1093/nar/gkaa1047](https://doi.org/10.1093/nar/gkaa1047).
- Kingsford, C. L., Ayanbule, K., and Salzberg, S. L. (2007). “Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake”. In: *Genome Biology* vol. 8, no. 2, R22. ISSN: 1474-760X. DOI: [10.1186/gb-2007-8-2-r22](https://doi.org/10.1186/gb-2007-8-2-r22), URL: <https://doi.org/10.1186/gb-2007-8-2-r22>.
- Leonard, S., Meyer, S., Lacour, S., Nasser, W., Hommais, F., and Reverchon, S. (Sept. 2019). “APER0: a genome-wide approach for identifying bacterial small RNAs from RNA-Seq data”. eng. In: *Nucleic Acids Research* vol. 47, no. 15, e88. ISSN: 1362-4962. DOI: [10.1093/nar/gkz485](https://doi.org/10.1093/nar/gkz485).

- Lorenz, R., Bernhart, S. H., Höner zu Siederdisen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). “ViennaRNA Package 2.0”. In: *Algorithms for Molecular Biology* vol. 6, no. 1, p. 26. ISSN: 1748-7188. DOI: [10.1186/1748-7188-6-26](https://doi.org/10.1186/1748-7188-6-26).
- McGinnis, S. and Madden, T. L. (2004). “BLAST: at the core of a powerful and diverse set of sequence analysis tools”. In: *Nucleic Acids Research* vol. 32, no. suppl_2, W20–W25. ISSN: 0305-1048. DOI: [10.1093/nar/gkh435](https://doi.org/10.1093/nar/gkh435).
- Nawrocki, E. P. and Eddy, S. R. (2013). “Infernal 1.1: 100-fold faster RNA homology searches”. In: *Bioinformatics* vol. 29, no. 22, pp. 2933–2935.
- Orr, M. W., Mao, Y., Storz, G., and Qian, S.-B. (2020). “Alternative ORFs and small ORFs: shedding light on the dark proteome”. In: *Nucleic Acids Research* vol. 48, no. 3. ISSN: 0305-1048. DOI: [10.1093/nar/gkz734](https://doi.org/10.1093/nar/gkz734).
- Ozuna, A., Liberto, D., Joyce, R. M., Arnvig, K. B., and Nobeli, I. (Feb. 2020). “baerhunter: an R package for the discovery and analysis of expressed non-coding regions in bacterial RNA-seq data”. In: *Bioinformatics* vol. 36, no. 3, pp. 966–969. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btz643](https://doi.org/10.1093/bioinformatics/btz643).
- Raina, M., King, A., Bianco, C., and Vanderpool, C. K. (2018). “Dual-Function RNAs”. In: *Microbiology Spectrum* vol. 6, no. 5. DOI: [10.1128/microbiolspec.rwr-0032-2018](https://doi.org/10.1128/microbiolspec.rwr-0032-2018).
- Rivas, E. and Eddy, S. R. (2001). “Noncoding RNA gene detection using comparative sequence analysis”. In: *BMC Bioinformatics* vol. 2, no. 1, p. 8. ISSN: 1471-2105. DOI: [10.1186/1471-2105-2-8](https://doi.org/10.1186/1471-2105-2-8). URL: <https://doi.org/10.1186/1471-2105-2-8>.
- Ryan, D., Jenniches, L., Reichardt, S., Barquist, L., and Westermann, A. J. (July 2020). “A high-resolution transcriptome map identifies small RNA regulation of metabolism in the gut microbe *Bacteroides thetaiotaomicron*”. en. In: *Nature Communications* vol. 11, no. 1, p. 3557. ISSN: 2041-1723. DOI: [10.1038/s41467-020-17348-5](https://doi.org/10.1038/s41467-020-17348-5).
- Samayoa, J., Yildiz, F. H., and Karplus, K. (2011). “Identification of prokaryotic small proteins using a comparative genomic approach”. In: *Bioinformatics* vol. 27, no. 13, pp. 1765–1771. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btr275](https://doi.org/10.1093/bioinformatics/btr275).
- Schnoor, S. B., Neubauer, P., and Gimpel, M. (2024). “Recent insights into the world of dual-function bacterial sRNAs”. en. In: *WIREs RNA* vol. 15, no. 1. ISSN: 1757-7012. DOI: [10.1002/wrna.1824](https://doi.org/10.1002/wrna.1824).
- Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J., and Goesmann, A. (2021). “Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification”. In: *Microbial Genomics* vol. 7, no. 11, p. 000685. ISSN: 2057-5858. DOI: [10.1099/mgen.0.000685](https://doi.org/10.1099/mgen.0.000685).
- Seemann, T. (2014). “Prokka: rapid prokaryotic genome annotation”. In: *Bioinformatics* vol. 30, no. 14, pp. 2068–2069. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btu153](https://doi.org/10.1093/bioinformatics/btu153).
- Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. (2010). “The Primary Transcriptome of the Major Human Pathogen *Helicobacter pylori*”. In: *Nature* vol. 464, no. 7286, pp. 250–255. ISSN: 1476-4687. DOI: [10.1038/nature08756](https://doi.org/10.1038/nature08756).
- Sorek, R. and Cossart, P. (2010). “Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity”. In: *Nature Reviews Genetics* vol. 11, no. 1, pp. 9–16. ISSN: 1471-0064. DOI: [10.1038/nrg2695](https://doi.org/10.1038/nrg2695).
- Stazic, D. and Voß, B. (2016). “The Complexity of Bacterial Transcriptomes”. In: *Journal of Biotechnology* vol. 232, pp. 69–78. ISSN: 0168-1656. DOI: [10.1016/j.jbiotec.2015.09.041](https://doi.org/10.1016/j.jbiotec.2015.09.041).
- Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., Lomsadze, A., Pruitt, K. D., Borodovsky, M., and Ostell, J. (Aug. 2016). “NCBI prokaryotic genome annotation pipeline”. In: *Nucleic Acids Research* vol. 44, no. 14, pp. 6614–6624. ISSN: 0305-1048. DOI: [10.1093/nar/gkw569](https://doi.org/10.1093/nar/gkw569).

- Vockenhuber, M.-P., Sharma, C. M., Statt, M. G., Schmidt, D., Xu, Z., Dietrich, S., Liesegang, H., Mathews, D. H., and Suess, B. (2011). “Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*”. In: *RNA biology* vol. 8, no. 3, pp. 468–477. ISSN: 1547-6286. DOI: [10.4161/rna.8.3.14421](https://doi.org/10.4161/rna.8.3.14421).
- Yu, S.-H., Vogel, J., and Förstner, K. U. (2018). “ANNOgesic: a Swiss army knife for the RNA-seq based annotation of bacterial/archaeal genomes”. In: *GigaScience* vol. 7, no. 9. ISSN: 2047-217X. DOI: [10.1093/gigascience/giy096](https://doi.org/10.1093/gigascience/giy096).

Supplementary Material

A. The TSS-Captur pipeline



The screenshot displays the TSS-Captur web interface. On the left is a blue sidebar with navigation links: 'File Upload' (active), 'Example Data', 'About', 'GitHub', and 'TueVis'. The main content area is titled 'Upload Files for TSS-CAPTUR' and includes the following elements:

- Intended for desktop usage
- Maximum size per file: 200.00 MB
- Maximum total file size: 500.00 MB
- MasterTable.tsv: A 'Choose File' button and the text 'No file chosen'.
- Genome Files: A 'Choose Files' button and the text 'No file chosen'.
- GFF Files: A 'Choose Files' button and the text 'No file chosen'.
- Max motif numbers: 5, with a slider bar.
- Model for CNIT: Plant, with a dropdown menu showing 'Plant' and 'Animal' options.
- A 'Start Upload' button at the bottom.

Figure III.S1: Snapshot of the web-interface of **TSS-Captur**. It allows the user to upload their files, as well as selecting a few parameters for the pipeline. Two other pages provide more information wrt. **TSS-Captur**. The *Example Data* page links different datasets that can be used directly in **TSS-Captur**, as well as their precomputed reports. The *About* page provides an overview of the pipeline, such as the tools used for characterization.

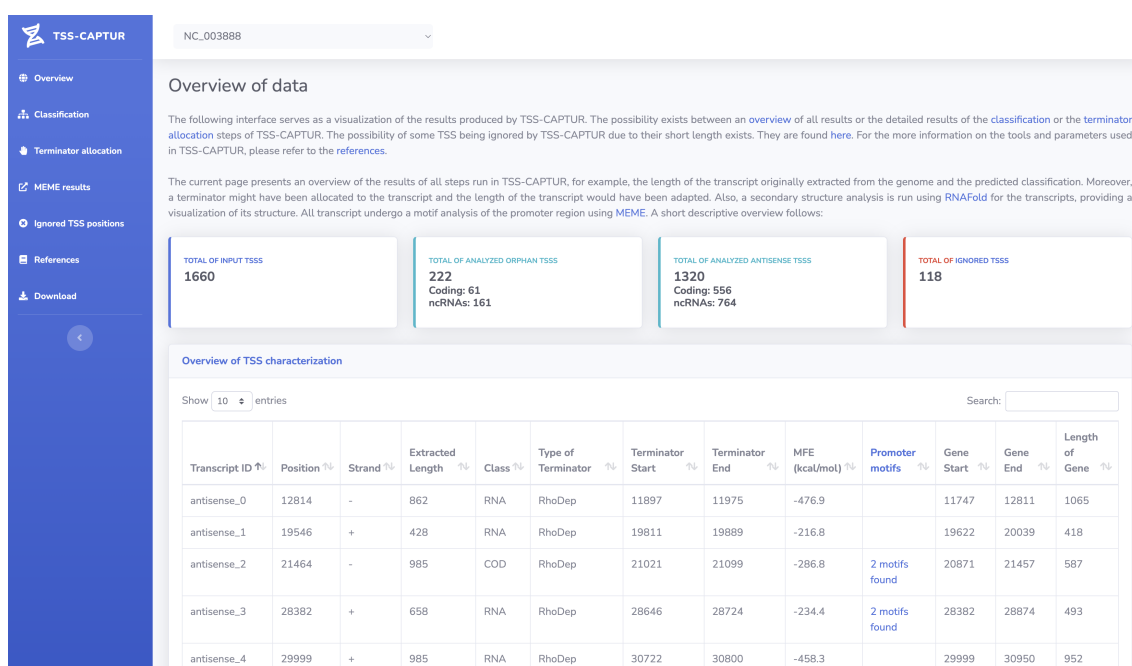


Figure III.S2: Snapshot of the interactive report of **TSS-Captur** visualizing *S. coelicolor*'s data. The interface provides different levels of details for each analysis step run in **TSS-Captur** that can be accessed via the navigation menu (left). An overview of the data is provided by summarizing main results of the analyzed TSSs. The information regarding each transcript is listed in a tabular view that allows sorting and filtering.

B. Evaluation of BLAST's hits

After **BLAST** returns a set of local alignments A for each transcript, the optimal hit H^* must be selected. Instead of relying solely on the best E -value or bit-score, other characteristics are crucial. The ideal pairwise-alignment hit should fulfill the following criteria:

- The distance to the TSS (D_{TSS}) should be minimal.
- The similarity Sim should be within the range of 65 – 85%.
- The coverage of the query (Cov_q) in the alignment should be maximized.

To evaluate each hit ($H_i \in A$), the scoring function [Eq. 1](#) is used. **qcovhsp**, **qstart** and **pident** are values returned by **BLAST** that represent the coverage of the query, the start of coordinate of the alignment in the query and the overall identity of the alignment, respectively. A weighted average S combines the three scores with weights 0.45, 0.35, and 0.2, reflecting the relative importance of TSS proximity, similarity, and query coverage. For each transcript, the alignment H^* is taken, with $H^* = \arg \max S(H_i), \forall H_i \in A$, and stored in a **FASTA**-file for usage in **QRNA**.

$$\begin{aligned}
 D_{TSS}(S_q, t) &= \frac{\min(\max(0, S_q - t), t + 20)}{t + 20} \\
 D_{Sim}(I_A) &= \frac{\mathcal{N}_{\mu=75, \sigma=8}(I_A)}{\mathcal{N}_{\mu=75, \sigma=8}(75)} \\
 Cov_q(C_q) &= \min(1, \frac{C_q - 50}{25}) \\
 S(H_i) &= 0.45 \cdot D_{TSS}(\text{qstart}_{H_i}, 10) \\
 &\quad + 0.35 \cdot D_{Sim}(\text{pident}_{H_i}) \\
 &\quad + 0.2 \cdot Cov_q(\text{qhspcov}_{H_i})
 \end{aligned} \tag{Eq. 1}$$

C. Evaluation of termination sites

During classification, **QRNA** and **CNIT** return the most-probable functional subsequence and hence hint for a first 3'-end of the transcripts. After this, **TSS-Captur** continues with the allocation of a predicted termination site. For each transcript, any termination regions predicted by **RhoTermPredict** and **TransTermHP** starting within the subsequence defined by the two classification tools are considered as putative termination sites. Since the used tools might return false-positives, the predictions are scored to identify the most-plausible termination site. We consider two aspects for this evaluation: the confidence value TTS_{score} returned by the corresponding prediction tool, and the location of the termination region wrt. the coordinates reported by the classification step (T_{start} and T_{end} , respectively, for the start and end of the predicted transcript). Note that T_{start} is the same as the TSS position and T_{end} is the putative end predicted by either **QRNA** or **CNIT**. The scores of **TransTermHP** and **RhoTermPredict** have different ranges, hence they are normalized, computing TTS'_{score} . A weighted average is computed using the normalized score and the distances to the TSS and the 3'-end of the transcript, as seen in [Eq. 2](#).

$$\begin{aligned}
 D_{start} &= \frac{|TTS_{start} - T_{start}|}{length(T)} \\
 D_{end} &= \frac{|TTS_{start} - T_{end}|}{length(T)} \\
 score &= \frac{TTS'_{score}}{2} + \frac{D_{start}}{4} + \left(\frac{1 - D_{end}}{4} \right)
 \end{aligned} \tag{Eq. 2}$$

III. TSS-Captur: A User-Friendly Pipeline for Characterizing Unclassified RNA transcripts

D. Use case: *S. coelicolor*

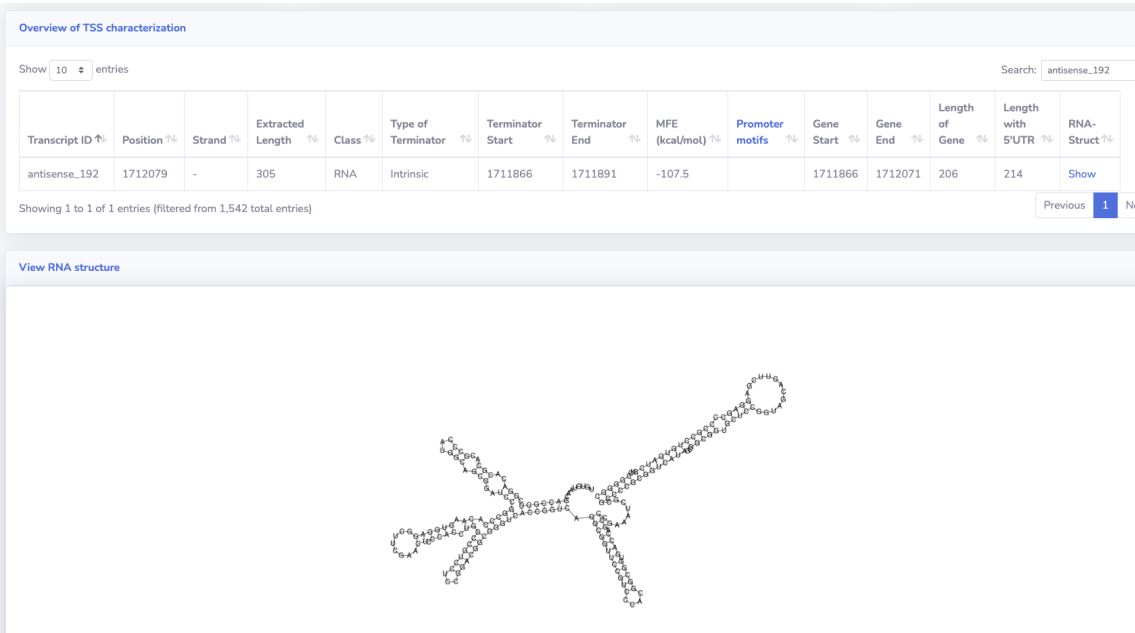


Figure III.S3: Interactive report of TSS-Captur visualizing *S. coelicolor*'s transcript *antisense_192*. The predicted secondary structure is directly visualized in the interface after clicking the corresponding element in the table.

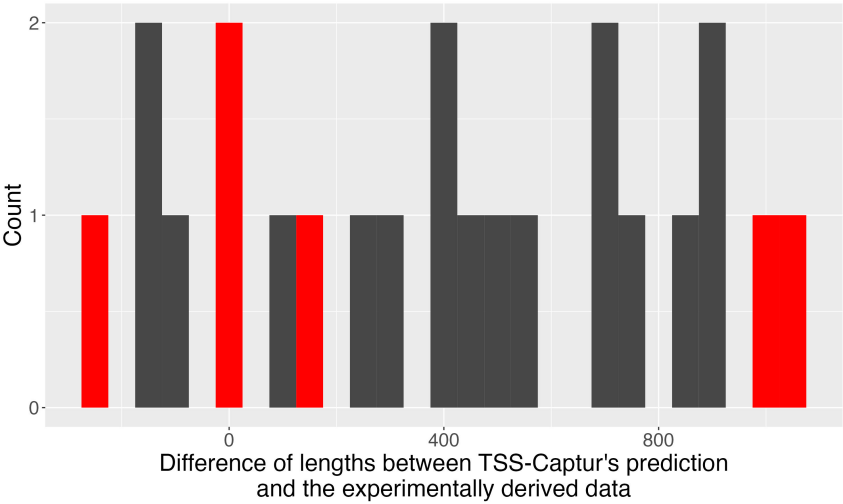


Figure III.S4: Histogram (bin width = 50) showing the length differences between experimental results presented by Vockenhuber *et al.* (2011) and predicted transcripts with TSS-Captur for *S. coelicolor*. Highlighted bars (red) represent transcripts confirmed using Northern Blot.

Table III.S1: Comparison of results of TSS-Captur with the study by Vockenhuber *et al.* (2011). The first columns represent the name, start, end and length reported by the study. Transcripts confirmed via Northern Blot (NB) by Vockenhuber *et al.* are indicated. For TSS-Captur's results, the position of the TSS, the start, end, class and length reported by the characterization are shown.

Study Vockenhuber <i>et al.</i>					Results of TSS-Captur			
Name	Start	End	Length	NB	ID	TSS	Length	Class
scr0991	1045891	1046090	200	Yes	orphan_20	1045891	351	RNA
scr1601	1711967	1712074	214	Yes	antisense_192	1712079	214	RNA
scr1821	1950867	1950948	82	Yes	orphan_32	1950867	1077	RNA
scr2736-1	2982224	2982505	282	Yes	orphan_48	2982224	34	RNA
as2780	3034099	3034209	111	Yes	antisense_376	3034209	1142	RNA
scr3035	3321216	3321353	138	Yes	orphan_59	3321353	140	RNA
as3287	3636488	3636569	82	No	antisense_488	3636567	527	RNA
as3317	3669433	3669657	225	No	antisense_491	3669433	1104	COD
as3321	3673385	3673581	197	No	antisense_494	3673385	709	COD
scr3323	3675098	3675329	232	No	orphan_77	3675098	331	RNA
scr3558	3933527	3933668	142	No	antisense_599	3933677	32	RNA
scr3559	3934693	3934927	235	No	orphan_88	3934687	71	RNA
scr3920	4315358	4315484	127	No	antisense_682	4315355	848	COD
scr3928	4323842	4324023	182	No	orphan_109	4323842	584	RNA
as4261	4674583	4674770	188	No	antisense_744	4674582	36	RNA
as4567	4984635	4984793	159	No	antisense_804	4984793	1059	COD
scr4632	5055055	5055172	118	No	orphan_142	5055173	533	RNA
as4675	5106760	5106941	182	No	antisense_860	5106761	417	RNA
as4699	5124039	5124366	328	No	antisense_870	5124354	621	RNA
as5028	5464169	5464418	250	No	antisense_942	5464419	806	RNA
scr6106	6706584	6706667	84	No	antisense_1148	6706584	932	COD
scr6908	7672451	7672670	220	No	orphan_230	7672451	916	RNA
scr6925	7688336	7688466	131	No	orphan_233	7688336	877	RNA

Table III.S2: Hits of *S. coelicolor*'s predicted ncRNA genes in the **RFAM** database. The **accession code** column corresponds to the identification used within **RFAM**, while the column **ID** refers to the label given to the transcript by **TSS-Captur**. **Start** and **End** refer to the local coordinates within each transcript that were aligned. The **predicted length** refers to the length predicted by **TSS-Captur**, while the aligned length refers to how much of the transcript was actually aligned. Hits marked with an asterisk (*) represent transcripts that come from the data of Vockenhuber *et al.*.

ID	Target name	Acc. code	Start	End	<i>E</i>-value	Pred. length	Aligned length
orphan_87	6C	RF01066	25	100	1.70×10^{-05}	167	76
antisense_111	Cobalamin Riboswitch	RF00174	229	331	4.40×10^{-14}	330	103
orphan_20	*Cobalamin Riboswitch	RF00174	17	235	4.90×10^{-27}	153	219
orphan_233	*Scr6925	RF02826	1	130	2.10×10^{-25}	560	130
antisense_192	*Scr1601	RF02830	6	213	3.00×10^{-40}	214	208
antisense_840	tRNA	RF00005	401	468	7.10×10^{-12}	467	68
orphan_127	ydaO-yuaA	RF00379	182	338	1.50×10^{-26}	337	157
orphan_128	Scr4115	RF02829	80	188	2.40×10^{-16}	644	109
orphan_157	Scr5239	RF02605	66	224	3.90×10^{-32}	543	159
orphan_167	che1	RF02935	370	471	4.70×10^{-15}	486	102
orphan_78	Bacterial ssrRNA	RF00177	602	743	6.80×10^{-32}	739	142

Appendix IV

A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

Mathias Witte Paz, Alina Bitzer, Kay Nieselt, Simon Heilbronner

Published in *NAR Genomics and Bioinformatics*, Volume 7, Issue 4, December 2025, lqaf183, DOI: [10.1093/nargab/lqaf183](https://doi.org/10.1093/nargab/lqaf183).

Abstract

Metallophores are secondary metabolites that enable bacterial growth in metal-limited environments such as the human nasal microbiome. While synthesis and uptake of metallophores in *Staphylococcus aureus* is well characterized, the diversity across the *Staphylococcus* genus remains unclear. We performed a comprehensive bioinformatic analysis of 77 representative species, as well as over 1800 strains, to map metallophore biosynthetic gene clusters (BGCs) and uptake systems. Staphyloferrin A (SF-A) biosynthesis was widely conserved, though disrupted loci were found in some species, with some of them appearing to have replaced SF-A with a newly discovered, still uncharacterized, BGC. In contrast, staphyloferrin B (SF-B) and staphylopine (STP) production were restricted to select species. Uptake systems were more broadly distributed, showing evidence of "cheating" species that lack biosynthesis, but retain the required lipoproteins for metallophore usage. *S. lugdunensis* exemplifies this, encoding multiple uptake systems without producing known metallophores. Strain-level variation was also observed, particularly with specific cases of SF-A truncation, but also for the diversity of lipoprotein receptors. These findings highlight the diversity of metallophore systems, suggesting diverse metallophore-dependent cooperation and competition within the *Staphylococcus* genus. This work provides a foundation for future experimental studies to identify the role of metallophores in microbial community interactions.

Keywords: Metallophores, Genomics, Biosynthetic Gene Clusters, Staphylococci

Introduction

Staphylococci are common members of the human and animal microbiota. The pathogen *Staphylococcus aureus* colonizes the anterior nares of approximately one-third of the human population, and this colonization represents a major risk factor for subsequent infections (Wertheim, Melles, *et al.* 2005; Wertheim, Vos, *et al.* 2004). Other staphylococcal species generally considered less virulent—such as *Staphylococcus epidermidis*—are ubiquitously present on human skin and nasal mucosa (Bomar *et al.* 2018).

Skin and mucosal surfaces are nutrient-limited environments, particularly with respect to trace metal ions such as iron (Fe), manganese (Mn), copper (Co), and zinc (Zn). These transition metals are essential cofactors for key metabolic enzymes, and although required only in trace amounts, bacterial growth is fully dependent on their availability. Among them, iron has been studied most extensively due to its critical role in enzymes involved in DNA replication, glycolysis, and respiration (Schaible and Kaufmann 2004; Weinberg 2000). However, Mn and Zn are also essential for bacterial physiology (Lonergan and Skaar 2019). On epithelial surfaces, host-secreted metal-binding proteins such as lactoferrin (Fe) and calprotectin (Mn/Zn) restrict the availability of these ions (Zackular *et al.* 2015). This defense mechanism is known as nutritional immunity and limits bacterial proliferation (Murdoch and Skaar 2022).

To overcome metal starvation, many microorganisms possess dedicated biosynthetic gene clusters (BGCs) allowing the production of metallophores, small, high-affinity secondary metabolites, that scavenge trace metals from the environment. In some cases, metallophores even extract metals directly from host proteins like lactoferrin or calprotectin (Neumann *et al.* 2017). Among the best-studied classes of metallophores are siderophores, a subclass specialized in iron scavenging (Kraemer *et al.* 2015). Once metal-saturated, the chelators are recognized by specific receptors on the bacterial surface (lipoproteins in Gram-positive bacteria) and imported into the cell, where the metal is released via reduction or degradation of the metal chelator. Metallophores can function as public goods within bacterial communities. Once secreted, they can be utilized by neighboring bacteria that express compatible receptors. This enables diverse ecological interactions, ranging from cooperative “shared labor” to exploitative “cheating”, where non-metallophore producing bacteria steal the benefits without contributing to the metabolic costs of their production (Kramer *et al.* 2020).

Staphylococcus aureus is known to produce metallophores, and both its metallophores and their physiological relevance are well characterized (Ghssein and Ezzeddine 2022) (Fig. IV.1). The pathogen synthesizes the iron-binding siderophores staphyloferrin A (SF-A) and staphyloferrin B (SF-B). SF-A is a carboxylate-type siderophore produced via a nonribosomal peptide synthesis (NRPS) independent system: The BGC consists of *sfaABCD* genes with *sfaB/D* representing *iucA/C* homologs, making staphyloferrin A a classical NRPS independent siderophore (NIS) (Beasley, Vinés, *et al.* 2009). Iron-loaded SF-A is recognized by the substrate-binding lipoprotein HtsA and imported via the HtsABC system (Cotton *et al.* 2009). SF-B is also a carboxylate-type siderophore, synthesized by SbnABCDEFGHI in an NRPS-independent fashion but without *IucA/C* homologs (Dale, Doherty-Kirby, *et al.* 2004). The siderophore is then imported via the SirABC system, following recognition by the SirA lipoprotein (Dale, Sebelsky, *et al.* 2004).

In addition to iron acquisition, *S. aureus* produces the Mn/Zn-binding metallophore staphylopin (STP) which is also synthesized in an NRPS-independent manner by CntKLME (Ghssein, Brutusco, *et al.* 2016). Metal-saturated staphylopin is recognized by the lipoprotein CntA and imported through the CntABCDF system.

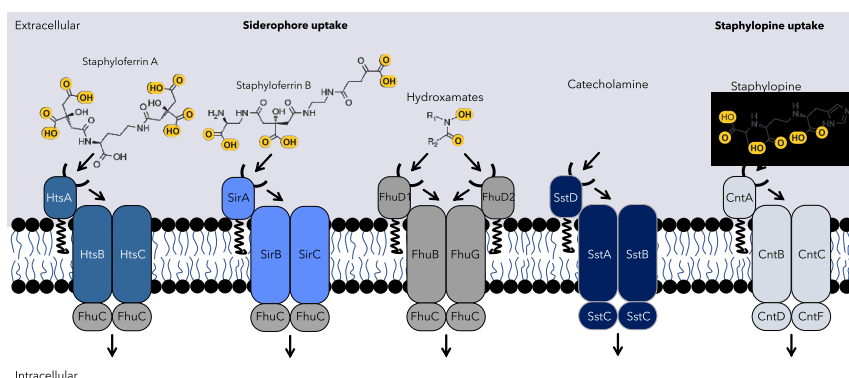


Figure IV.1: Schematic representation of *S. aureus* siderophores and metallophores, as well as their acquisition systems.

S. aureus is also capable of utilizing siderophores produced by other organisms (xenosiderophores). The lipoproteins FhuD1 and FhuD2 mediate uptake of hydroxamate-type siderophores (Sebulsky, Hohnstein, *et al.* [2000]; Sebulsky, Speziali, *et al.* [2004]). In contrast to HtsA, SirA and CntA the xenosiderophore receptors are more promiscuous and more than one dedicated molecule can be recognized. The precise structural mechanism of hydroxamate binding to the FhuD1/FhuD2 receptors is unclear, but it has been shown that the proteins bind different hydroxamate type siderophores with different affinities, suggesting specialization and multiple alleles might allow an increased spectrum of hydroxamates to be acquired (Sebulsky and Heinrichs [2001]; Sebulsky, Speziali, *et al.* [2004]). Similarly, SstA can recognize catecholate type iron chelators from host and bacterial origin (Beasley, Marolda, *et al.* [2011]) and, as identified in *S. lugdunensis*, different alleles are proposed to expand the range of molecules that can be acquired (Flannagan *et al.* [2022]).

While SF-A is thought to be widely produced across the staphylococcal genus, SF-B production appears to be restricted to *S. aureus* (Dale, Doherty-Kirby, *et al.* [2004]). However, the capacity of other staphylococci to produce metallophores or to use those of other species remains poorly understood. Additional insight is needed to enhance our understanding of metallophore-driven interspecies interactions within the human microbiome.

In this study, we conducted a comprehensive bioinformatic analysis of metallophore biosynthesis and acquisition genes across the *Staphylococcus* genus. We found that while SF-A biosynthetic genes are broadly conserved, several species possess truncated loci, likely rendering them non-functional. Interestingly, one distinct phylogenetic clade harbors a metallophore biosynthesis operon that remains to be characterized, possibly compensating for the loss of SF-A production. SF-B biosynthetic genes were confined to *S. aureus* and closely related, animal-associated species. However, several coagulase-negative staphylococci (CoNS) encode the SF-B acquisition gene *sirA*, suggesting an adaptation to exploit SF-B produced by pathogens. Likewise, xenosiderophore uptake genes (*fhuD*, *sstA*) are widespread, and many species carry multiple paralogs, indicating functional diversification and expanded substrate specificity. Finally, we observed low intra-species variation in metallophore biosynthesis and acquisition genes, suggesting that horizontal gene transfer of these systems is limited within this genus.

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

Table IV.1: Number of available strains from the GTDB and number of remaining assemblies after filtering for four selected species and quality ($N50 > 50$ kb, # contigs < 250).

Species	Number of strains retrieved from the GTDB	Used assemblies
<i>S. capitis</i>	215	180
<i>S. epidermidis</i>	1,556	1,292
<i>S. hominis</i>	274	235
<i>S. lugdunensis</i>	144	128

Material and Methods

To characterize the metallophore synthesis and uptake mechanisms in the staphylococcal genes, we conducted an analysis on the species- and on the strain-level. For the analysis on the *Staphylococcus* species level, the accession numbers of representative samples for each *Staphylococcus* species were collected from the Genome Taxonomy Database (GTDB) (Parks *et al.* 2022). This resulted in 84 genomes that span the available species. For some species, the GTDB returned more than one representative. In these cases, only one representative per species was selected for a more consistent comparison, resulting in 77 genomes (Supp. Table S1).

For the analysis on the strain level, four human-related species were selected. The accession numbers of all available strains from the GTDB were retrieved and the corresponding genomes and protein sequences were downloaded from the NCBI database. Some strains' assemblies were removed due to their low quality ($N50 < 50kb$ or they contained a total of more than 250 contigs). In total, approximately 1,800 strains were analysed. The final number of strains per species for this analysis can be found in Table IV.1. The corresponding accession codes can be found as supplementary data.

The following analyses were first computed for the set of species and applied subsequently on the selected strains. Based on the retrieved genomes, biosynthetic gene clusters (BGCs) were identified using **antiSMASH** (version 7.1.0) (Blin *et al.* 2023). **antiSMASH** classifies the BGCs into different cluster types with a rough functional description. Due to our focus on siderophore synthesis potential, we focused only on BGCs labeled as NRPS-independent siderophores and Non-ribosomal peptide (NRP) metallophores. Additional modules from **antiSMASH** were activated to annotate the identified BGCs. The module **ClusterCompare** annotates and provides a similarity score to BGCs found in the **MIBiG** database (Terlouw *et al.* 2023). If a BGC was similar to any known metallophore-related BGC in *S. aureus* (*sfaABCD*, *sbmABCDEFGHI*, or *cntKLME*), we quantified its completeness. That is, we assessed whether all required synthesis genes were present. The BGCs were classified as complete if they had 100% similarity to the annotated counterpart. Otherwise, they were considered incomplete as long as the similarity was not zero. If the analyzed representative for a species showed an incomplete BGC, other available strains from the GTDB were analyzed to assess whether this was a species-wide observation.

Lastly, while some BGCs were classified as metallophores or siderophores, they reflected no similarity to any BGC in the **MIBiG** database via **ClusterCompare**. These BGCs were then compared to each other and to the reference BGCs of *S. aureus* using the tool **clinker** (version 0.0.31) (Gilchrist and Chooi 2021).

Since **antiSMASH** focuses solely on the identification of biosynthetic genes, it cannot assess a species' capacity to uptake metallophores. In *Staphylococci*, lipoproteins play a key role in the uptake of siderophore-iron complexes. Hence, we expect that gene homologs are present in the investigated genomes and can uptake the corresponding siderophore.

Considering the evolutionary diversity across species and strains, our objective was to establish a threshold that would yield significant outcomes in terms of sequence diversity for homolog identification.

To analyze the uptake potential of the species, a sequence homology analysis with MMSEQS2 (version 15, commit: ad6dfc) was performed (Steinegger and Söding [2017]). The amino acid sequence of the six lipoproteins of *S. aureus* was used as queries in MMSEQS2 to search within all staphylococcal proteomes. To identify the best threshold, the results for the *S. aureus* search were analyzed in detail. After removing the self-hits (100% similarity), the highest off-target similarity was observed between HtsA and a heme-binding protein, at 49.2% similarity. Based on this, we set a 50% sequence similarity threshold to define potential homologs of the protein references across the remaining species. If a protein shows a similarity higher than 50% to more than one of the reference lipoproteins of *S. aureus*, the lipoprotein was called homolog only to the reference gene with the highest similarity. Moreover, multiple genes of the same species can be identified as potential homologs of one reference lipoprotein. In such cases, the number of potential homologs was counted and the similarity value of the best hit per lipoprotein per species is retained.

To identify whether specific phylogenetic clusters had similar metallophore synthesis and uptake potential, a phylogenetic tree was calculated based on a core-genome multilocus sequence typing (cgMLST). The cgMLST was identified using chewBBACA (version 3.3.10) (Silva *et al.* [2018]) to then create a distance matrix using the script `mlst2dist` (<https://github.com/tripitakit/mlst2dist>) based on the algorithm proposed by Allelematch (Galpern *et al.* [2012]). From the resulting distance matrix, a phylogenetic tree was reconstructed using the neighbor joining algorithm implemented in MEGA11 (version 11.0.11) (Tamura *et al.* [2021]).

For the synthesis potential at the strain level, we applied the same classification as before, categorizing strains as complete or incomplete based on the presence of the full biosynthetic gene cluster. In the case of *S. lugdunensis*, a previous study identified an NRPS system with metallophore potential (Heilbronner *et al.* [2011]). To analyze the occurrence of this BGC across *S. lugdunensis* strains, we queried the BGCs identified by antiSMASH using BiG-SCAPE (version 2) (Navarro-Muñoz *et al.* [2020]) and included its presence in the summarized results. For the uptake potential, a strain was considered positive for uptake if it harbored at least one potential homolog of the reference lipoproteins, using the previously defined sequence similarity threshold of 50%. The results for metallophore synthesis and uptake potential were summarized using an UpSet plot (Lex *et al.* [2014]). The UpSet plot visualizes the distribution and overlap of these traits across strains for each species individually.

Results & Discussion

Metallophore Biosynthesis Genes Vary Between Staphylococcal Species

To investigate the genus *Staphylococcus*, we collected a non-redundant set consisting of 77 species representing all currently available species at the Genome Taxonomy Database (GTDB). A phylogenetic tree was computed based on the core genome MLST, and the presence of the *sfa*, *sbm*, and *cmt* gene clusters—associated with the production of staphyloferrin A (SF-A), staphyloferrin B (SF-B), and staphylopin (STP), respectively—was assessed (Fig. [IV.2]).

Though described as representatives for the corresponding species, the assemblies show a high diversity in their assembly levels (Fig. [IV.2], leftmost heatmap column). Moreover, despite being classified as *Staphylococcus* in the NCBI taxonomy, our phylogenetic analysis

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

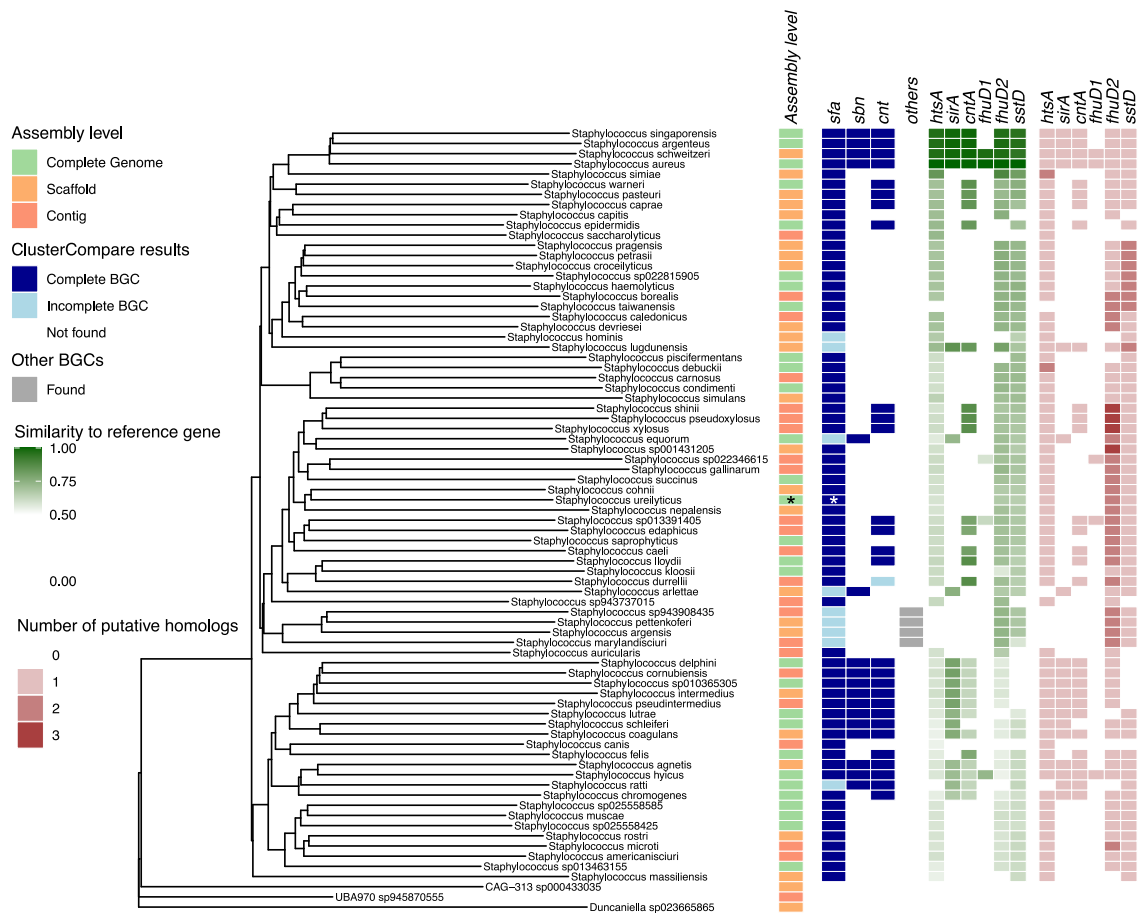


Figure IV.2: Overview of metallophore synthesis and uptake potential in the genus *Staphylococcus*. The dendrogram shows the phylogenetic relationships of staphylococcal species, alongside the results summarizing their metallophore-related biosynthetic and uptake capacities. The assembly level of each used reference is visualized in the first heatmap. The synthesis potential was assessed with *antiSMASH*. Results annotated as the three known *S. aureus* metallophores obtained with *ClusterCompare* are shown in the second to the left heatmap column. Blue rectangles indicate biosynthetic gene clusters (BGCs), with dark blue representing complete BGCs (all synthesis genes present) and light blue indicating incomplete BGCs. Gray rectangles mark additional BGCs predicted as potential metallophore synthesizers but without specific annotation. The last two columns represent the uptake potential: green shading indicates the sequence similarity of the proteome to *S. aureus* metallophore uptake lipoproteins, while red shading shows the number of putative homologs identified. Strains that were reassessed and that have shown different results have been marked with an asterisk (*).

identified three species as outliers, consistent with alternative genus-level classification proposed by the GTDB. Still, these species were included in our analysis and used as outgroups for the phylogenetic tree. It is important to note that none of these three species was identified as a potential metallophore producer nor cheater by our analysis.

Our data show that SF-A biosynthesis is a core trait across the genus, with *sfa* genes present in all analyzed staphylococcal species (Fig. IV.2, leftmost blue heatmap column). However, 10 species carry truncated *sfa* loci, suggesting abrogation of SF-A biosynthesis. We used *clinker* to investigate structurally distinct *sfa*-loci in more detail (Fig. IV.3). This analysis confirmed that *S. lugdunensis* harbors a truncated locus lacking *sfaDA* genes. This has been described previously and *in vitro* evidence is available showing that the truncation hinders SF-A biosynthesis (Brozyna *et al.* 2014). Interestingly, *S. ureilyticus*

(GCF_002902235.1) appears to carry a similar truncation although the location of the sequence at the very end of a contig raises some doubt regarding the accuracy or quality of this assembly. To address this, we repeated the analysis using the reference assembly from the RefSeq database (GCF_025558845.1) (Pruitt *et al.* 2007), which revealed a complete siderophore biosynthesis cluster. Nevertheless, since siderophore production is known to be strain-specific (Zhao *et al.* 2024), the current assembly data do not allow for generalization of these findings to the entire species.

Additional eight species including a clade containing *S. pettenkoferi*, *S. argensis*, *S. marylandensis*, and *S. sp943908435*, as well as individual species in distinct clades such as *S. equorum* and *S. hominis* showed internal deletion of *sfaA* (Fig. IV.3). Analysis of additional genomes from *S. pettenkoferi*, *S. argensis*, *S. equorum* and *S. hominis* with **antiSMASH** confirmed that *sfa* inactivation is conserved at the species level (Fig. IV.6B and Supp. Table S2). Our results are in general agreement with a previous study, showing *S. chromogenes* but not *S. equorum* to produce SF-A (Reydams *et al.* 2024).

Together, these data show that *sfa* genes are omnipresent within staphylococci, but recombinations of the locus are observed in 10 out of 74 (13.5%) species, most likely hindering the production of SF-A.

In contrast, only 23% of the species (17 genomes out of 74) harbor the *sbm* genes for SF-B biosynthesis. These include *S. aureus* and its close relatives (*S. schweitzeri*, *S. argenteus*, *S. singaporensis*), but also more distantly related, non-human-associated species such as *S. equorum*, *S. delphini*, *S. pseudintermedius*, and *S. ratti*. This is in agreement with studies suggesting occasional SF-B production by non-*S. aureus* staphylococci (Lindsay *et al.* 1994).

Staphylopine biosynthesis genes (*cnt*) are found in 39% of species (29 out of 74) across various clades. One potentially non-functional *cnt* locus was identified in *S. durrellii*. Since no further genomes for this species are currently available at GTDB nor RefSeq, this hinders a further confirmation analysis for this species. However, we performed a detailed comparison of the *cnt* BGCs of *S. aureus* and *S. durrellii* (Fig. IV.4). This analysis revealed that *S. durrellii* lacks the gene *cntK*, most likely preventing STP production.

antiSMASH identified biosynthetic gene clusters (BGCs) potentially involved in metallophore production that have not yet been annotated in the **MIBiG** database (Fig. IV.2, grey column). These novel NIS-BGCs were discovered in a clade comprising *S. pettenkoferi*, *S. argensis*, *S. marylandensis*, and *S. sp943908435*. For further characterization, the BGCs were compared to each other and to known metallophore-producing BGCs of *S. aureus* using **clinker** (Fig. IV.5). Moreover, two other BGCs previously associated with metallophore production were included in the comparison: an NRPS system in *S. lugdunensis* (Heilbronner *et al.* 2011) and an NRPS-independent siderophore of *Mammaliococcus sciuri* (GCA_037849205.1, previously named *Staphylococcus sciuri*) (Zhao *et al.* 2024).

The analysis showed a high degree of conservation between the uncharacterized clusters regarding both operon organization and sequence similarity, strongly suggesting that the strains produce the same or at least very similar compounds. Among these, the *S. marylandsciuri* BGC showed the highest sequence divergence, despite preserving the overall operon structure. Interestingly, the BGC identified in *M. sciuri* shows conserved synthesis and transporter genes, suggesting a resemblance to the uncharacterized BGCs of the mentioned clade. Importantly, no significant similarity to *sfa*, *sbm*, or *cnt* gene clusters was identified, suggesting that these BGCs likely encode the biosynthesis of a distinct, still uncharacterized metallophore. Notably, the phylogenetic clade of staphylococci carrying these BGCs corresponds to the *sfa*-deficient clade described above (Fig. IV.2, Fig. IV.3).

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

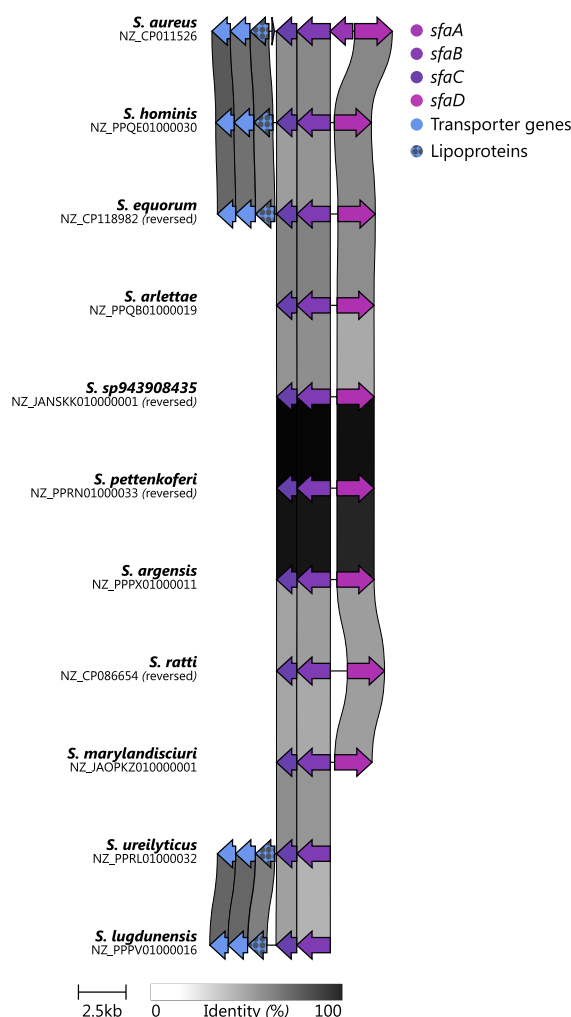


Figure IV.3: Comparison of all incomplete BGCs related to the synthesis of staphyloferrin A and the reference BGC of *S. aureus* using the tool *clinker*. The core biosynthetic genes (*sfaABCD*) are colored individually. Transporter genes are colored with respect to the annotation provided by *antiSMASH*. Other genes have been removed from the figure for clarity.

Accordingly, it is tempting to speculate that this clade replaced SF-A by a novel siderophore, most likely to produce a privatized compound that is not shared with other staphylococcal species (Kramer *et al.* 2020).

Furthermore, our analysis of *S. lugdunensis* also identified the previously reported NRPS-system of unknown function (Heilbronner *et al.* 2011). This system shows no similarity to any other staphylococcal metallophore BGCs (Fig. IV.5), but *antiSMASH* shows a sequence similarity to the streptococcal NRPS system responsible for the production of the siderophore equibactin (Heather *et al.* 2008). Nevertheless, it is important to highlight that the BGC is not annotated as a metallophore BGC, but only as an NRPS system; hence, it was not identified by our first analysis using *antiSMASH*. As discussed in the original publication, it is unclear if this system produces an antibacterial compound similar to lugdunin (Zipperer *et al.* 2016) or if it produces a metallophore. However, the similarity to equibactin along with the presence of a dedicated import system of the energy coupling factor type (ECF-transporter)—which are known as trace nutrient acquisition systems (Rempel *et al.* 2019)—strongly suggests the production of a nutrient scavenging molecule. Nevertheless, the observation that *S. lugdunensis* does not produce

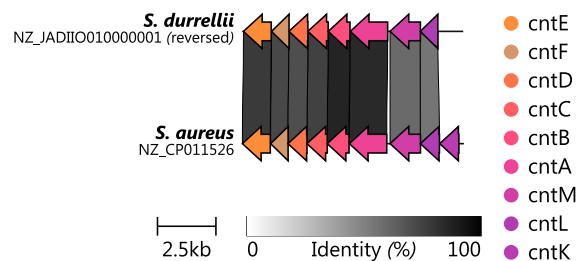


Figure IV.4: Comparison of incomplete *cnt* BGCs of *S. durrellii* to the corresponding loci of *S. aureus* using *clinker*. Other genes found within the BGC have been removed for clarity.

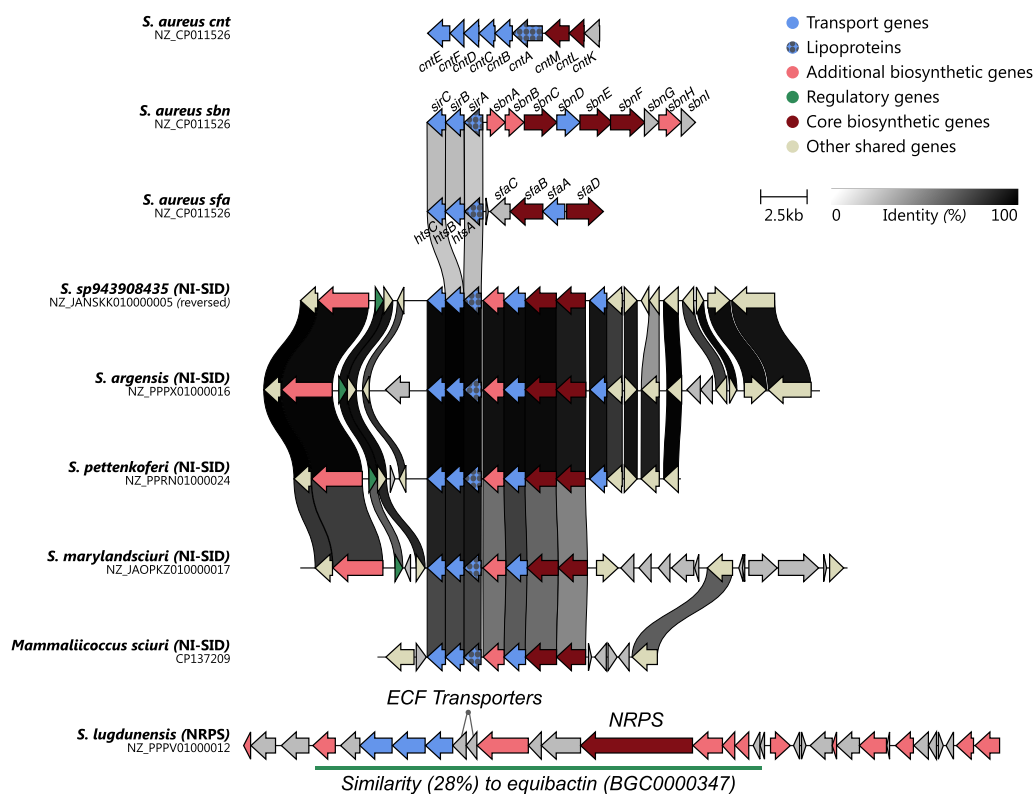


Figure IV.5: Comparison of novel BGCs identified in this study, known metallophore BGCs of *S. aureus*, and two uncharacterized BGCs from other studies using the tool *clinker*. For all BGCs, the genes have been colored with respect to the annotations provided by *antiSMASH*. For the last BGC, a subset of genes shows a similarity to the BGC producing equibactin (BGC0000347).

iron-scavenging siderophores *in vitro* (Brozyna *et al.* 2014; Zhao *et al.* 2024) might either suggest inactivity of the system or that the molecule does not target iron but a different metal such as Zn, Mn, or Cu. Additional laboratory-based experiments are needed to answer these questions.

Metallophore Acquisition Systems Vary Between Staphylococcal Species

Metallophores, once secreted, act as public goods and can be utilized by other community members. To assess the potential for metallophore acquisition across the genus, we analyzed the distribution of genes encoding known siderophore receptors (Fig. IV.2, green columns).

The *htsA* gene, encoding the SF-A receptor, is broadly conserved, consistent with the widespread presence of *sfa* biosynthetic genes. In species lacking intact *sfa* loci, *htsA* was

often also absent, supporting the idea of coordinated loss of synthesis and uptake systems. However, some species retained *htsA* despite non-functional *sfa* loci, suggesting that they may scavenge SF-A produced by other species. For example, SF-A usage of *S. lugdunensis* has been demonstrated experimentally (Brozyna *et al.* 2014). Furthermore, it has been shown that SF-A is a widely accessible siderophore within the nasal microbiome and its usage by “cheaters” that do not produce siderophores themselves reduces the fitness of the producers (Zhao *et al.* 2024). Intriguingly, *S. taiwanensis* was the only species with an intact *sfa* operon but no detectable *htsA*. Since no additional genome sequences are currently available, the results derived with MMSEQS2 were analyzed in more detail. The most similar gene of *S. taiwanensis* (locus tag: HYI43_00815) to *htsA* has a similarity of around 40%, hence falling below our defined homology threshold of 50%. Interestingly, this gene is not found within the borders of the *sfa*-related BGC identified by antiSMASH, but with a distance of around 680 kb to the core biosynthesis genes. Moreover, we evaluated the presence of the remaining components of the *htsABC* and identified homolog genes to *htsB* (HYI43_04060, 72.5%) and *htsC* (HYI43_04065, 73.8%). This led to the identification of a neighbouring coding sequence region (HYI43_04055) with a substrate-binding protein potential found. However, this coding region is labeled as a pseudogene, explaining why the corresponding amino acid sequence was missing from the species’ proteome, and hence from the MMSEQS2 analysis. Moreover, the sequence of this pseudogene translates to a 34 aa long protein sequence, highlighting the difference to the complete sequence of *htsA* of 327 aa. This raises questions about the functionality of this putative protein and therefore about *S. taiwanensis*’ ability to re-import SF-A.

All SF-B-producing species encode the corresponding receptor SirA (Fig. IV.2, second-left green column). Only two species, *S. lugdunensis* and *S. chromogenes*, were identified that encode *sirA* without an *sbn* cluster. Again, it has been experimentally demonstrated that *S. lugdunensis* can utilize SF-B and a similar trait can be expected for *S. chromogenes* (Brozyna *et al.* 2014). Interestingly, the presence of *sirA* in *S. chromogenes* was shown to be strain- rather than species-specific (Reydams *et al.* 2024) and the phenomenon of SF-B cheating appears to be a rare trait among staphylococcal species.

A similar picture was apparent for STP production and acquisition (Fig. IV.2, middle green column). Almost all species with STP biosynthesis genes harbor the corresponding receptor *cntA*, or if they have an inactive locus, they retain the receptor. Interesting exceptions are *S. schleiferi*, which shows an intact *cntKLME* operon but no *cntA* receptor, and *S. lugdunensis* encoding *cntA* without corresponding biosynthesis genes. To verify the results of *S. schleiferi*, we proceeded similarly as with *S. taiwanensis*. Here, no other protein was identified as similar to *cntA* by our MMSEQS2 search. For the ABC transporters *cntBC*, two potential homologs were identified: JM183_RS11640 for *cntB* (sequence similarity: 64.4%) and JM183_RS11635 for *cntC* (69.1%). A neighboring gene (JM183_RS11645) described as *nickel ABC transporter substrate-binding protein* was also identified in this case. However, a premature stop codon leads to a protein sequence of 390 aa, hence differing with respect to the expected length of 532 aa for *cntA*, suggesting that *S. schleiferi* might not be able to uptake STP.

Siderophores are costly metabolic molecules and act as public goods (Kramer *et al.* 2020). Cheaters arise within bacterial communities that consume but do not produce siderophores, which exhibit a metabolic burden to producers. Accordingly, siderophore production genes along with associated receptors are known to be under evolutionary pressure. For *Pseudomonas aeruginosa* it is known that the siderophore pyoverdine is produced with many subtle modifications, requiring modified receptors for acquisition (Gu, Shao, Qu, *et al.* 2025). This allows production of privatized siderophores preventing their

usage by cheaters. It is widely unclear if this phenomenon is also occurring in *Staphylococci*.

To address this, we investigated the conservation of the metallophore-binding sites within the receptors HtsA, SirA and CntA across species. A multiple sequence alignment was computed using ClustalO (v1.2.4) (Sievers *et al.* [2011]) for the sets of homologs across all species and the conservation of amino acid residues known to be engaged in binding of the metallophores (Grigg, Cheung, *et al.* [2010]; Grigg, Cooper, *et al.* [2010]; Song *et al.* [2018]) was analysed (Suppl. Table IV.S3). For HstA we found 6 out of 9 important amino acids to be strictly conserved amongst the HstA alleles of all species. However, 21 species (33%) showed conserved divergence of three amino acids (R86N, K203Q, R304K) and a single species showed a unique profile (R86N, K203E, R304-K) (Suppl. Figure IV.S1). It is tempting to speculate that these receptor variants might possess altered affinity towards SF-A or even recognize a modified version of the siderophore as observed for *P. aeruginosa* (Gu, Shao, Qu, *et al.* [2025]). However, additional biological and biochemical evidence is needed to support this idea. For SirA divergence was observed in two species at two binding sites (Suppl. Fig. IV.S2). An exchange of T144N was found in *S. lugdunensis*, and R206C was identified in *S. chromogenes* (Suppl. Figure IV.S2). The first species *S. lugdunensis* was demonstrated to consume SF-B, strongly suggesting that the receptor variant is able to recognize SF-B, putative effects on affinity remain however speculative. The effect of R206C on *S. chromogenes* remains unclear. Lastly, 55% of the identified CntA homologs differ from the canonical *S. aureus* receptor by carrying an N448T exchange (Suppl. Figure IV.S3). Again it remains unclear if these exchanges affect specificity or affinity.

S. aureus is known to import hydroxamate-type siderophores via the Fhu system (Cabrera *et al.* [2001]), using two receptors, FhuD1 and FhuD2 (Sebulsky and Heinrichs [2001]). Homologs of *fhuD1* were rare (5 out of 77 species), while *fhuD2* was widely conserved across the genus (89% of species, 69 out of 77, Fig. IV.2, second-right green column). Interestingly, several species harbored up to three distinct *fhuD2* homologs (Fig. IV.2, red columns), possibly expanding their hydroxamate siderophore uptake range.

The Sst system enables *S. aureus* to import catecholate-type siderophores (Beasley, Marolda, *et al.* [2011]). Homologs of *sst* were widespread (67 out of 77, 87%, Fig. IV.2, rightmost green column). However, a clade of seven species, closely related to *S. lugdunensis*, harbored multiple *sst* copies. This duplication, previously observed in *S. lugdunensis* (Flannagan *et al.* [2022]), may broaden the catecholate substrate range available to these species.

Taken together, *S. lugdunensis* presents a staphylococcal species with special characteristics. It appears to produce a single metallophore of so far unknown specificity. It produces neither SF-A nor SF-B nor STP but encodes the dedicated uptake systems for all these metallophores. Finally, the species encodes FhuD and multiple homologues of SstD. Altogether this suggests that *S. lugdunensis* is a highly adapted “cheater” that depends strongly on the biosynthetic potential of other species to thrive under metal-limited conditions.

Strain-Level Variation

Recent studies have demonstrated that metallophore production and uptake may vary at the strain level (Zhao *et al.* [2024]). To investigate this, we analyzed metallophore-related loci at the strain level for four selected species (*S. hominis*, *S. epidermidis*, *S. capitis*, and *S. lugdunensis*).

For *S. epidermidis*, the majority of strains (1,142 of 1,292 analysed genomes) mirrored the profile of the GTDB reference strain, encoding both production and uptake systems for

SF-A and STP and the presence of *sstD* (Fig. IV.6A). However, certain groups of isolates showed different profiles. 97 isolates lacked *sstD*, suggesting the inability of these strains to use catecholate-type siderophores. Additionally, individual isolates showed single or multiple genes being compromised, suggesting strain-level differences in their abilities to produce or consume metallophores.

Additionally, we looked for evidence for the development of divergent receptor alleles within *S. epidermidis*. Comparison of the full length HstA protein sequences showed a high degree of conservation (69.6% to 99.2% identity to the reference allele of *S. aureus*) (Suppl. Figure IV.S4a). However, when considering only the residues known to be involved in SF-A binding, the conservation increased to 100% in 99.9% of isolates (Suppl. Figure IV.S4b). The strains carrying alterations showed truncated alleles, suggesting inactivated proteins or inaccurate assemblies. Altogether, this analysis suggests that the receptors bind staphyloferrin A in a conserved manner, and direct evidence that the receptors develop to accommodate putative staphyloferrin A variants is not apparent.

The *S. hominis* strains showed high conservation regarding their ability to produce or secrete metallophores with only 2 out of 235 sequenced isolates diverging from the GTDB reference strain by lacking individual systems (Fig. IV.6B).

In *S. capitis*, only 29 of the 180 strains showed the gene combination of the GTDB reference strain, while the majority of strains (140) appear proficient in staphylopine production and acquisition. This suggests that the representative strain defined by the GTDB (GCF_002902325.1) is poorly representing the species in this regard (Fig. IV.6C).

Two almost equal groups of strains were identified in *S. lugdunensis*. While 62 of 128 analysed strains showed identical traits as the reference strain, 52 strains encoded an additional FhuD1 homologue, suggesting an increased ability to acquire hydroxamate-type siderophores (Fig. IV.6D). In addition, ten strains did not encode any *FhuD* protein and three strains showed individual combinations of gene inactivations. More interestingly, the NRPS system putatively encoding a metallophore of unknown function was identified in 127 strains, showing a conserved trait across multiple *S. lugdunensis* strains.

Conclusion

This study provides a systematic overview of metallophore synthesis and uptake potential across the *Staphylococcus* genus, revealing the diversity at both species and strain level, based on similar approaches applied to other species (Cordero *et al.* 2012; Gu, Shao, Rehm, *et al.* 2024; Koedooder *et al.* 2023). Our analysis shows that staphyloferrin A (SF-A) biosynthesis is a widespread trait among staphylococci. However, multiple species harbor truncations within the *sfa* locus, likely reflecting adaptations that favor metallophore scavenging rather than the metabolic cost of siderophore production. Furthermore, it also suggests the functional replacement of SF-A, as seen in the phylogenetic clade with a novel but conserved NRPS-independent metallophore-related BGC of yet unknown function. Future analyses of this yet uncharacterized BGC could focus on its experimental validation, as well as on the application of novel bioinformatic methods that focus on the characterization of the chemical structure of the produced metabolite (Butler *et al.* 2023).

In contrast, staphyloferrin B (SF-B) and staphylopine (STP) production is restricted to a subset of species, with examples of gene loss or inactivation (e.g., in *S. durrellii*) driving both species- and strain-level variation. Our data also confirm some occurrences of metallophore cheating, where species retained uptake systems without maintaining the corresponding biosynthesis genes. As a clear representative of the cheater species, *S. lugdunensis* displays an extensive repertoire of metallophore receptors despite lacking

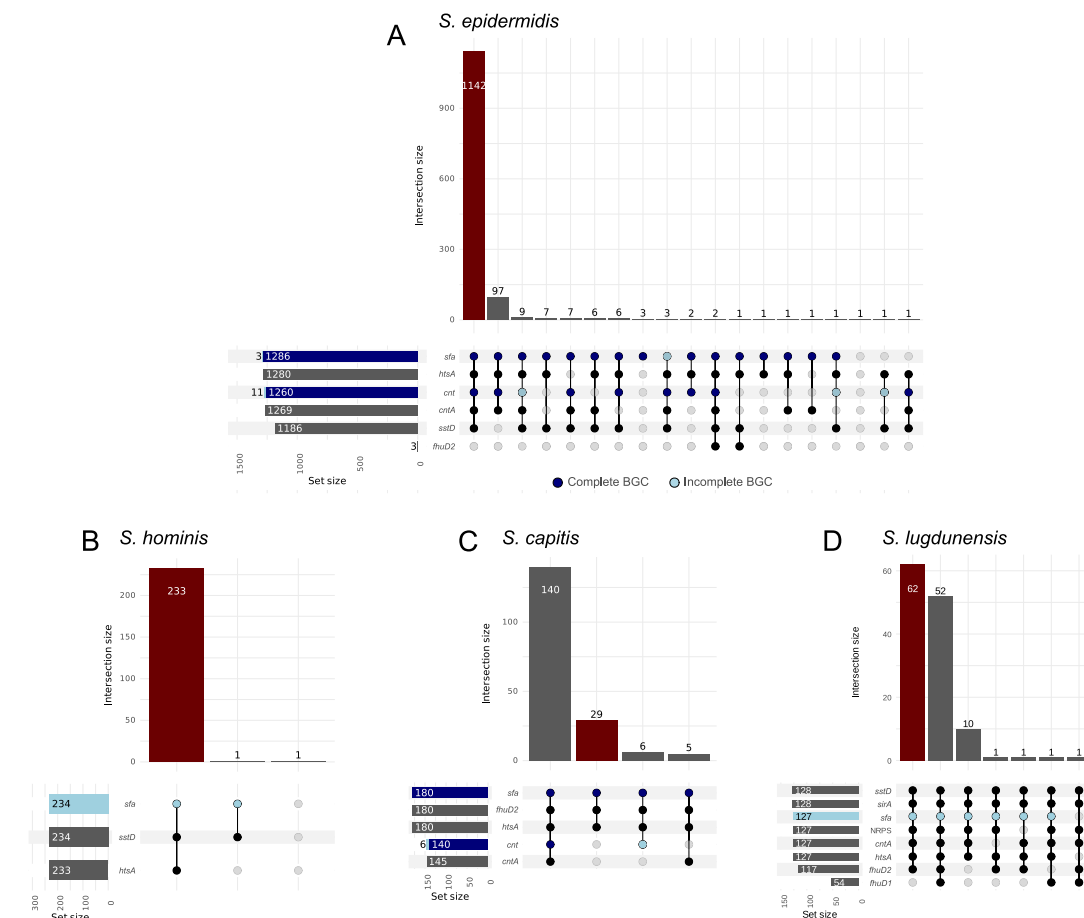


Figure IV.6: UpSet plots for the strain analysis on four selected species: (A) *S. epidermidis*, (B) *S. hominis*, (C) *S. capitis*, and (D) *S. lugdunensis*. The BGCs are classified as complete (dark blue) if they show 100% similarity to the reference BGCs in the MIBiG database. Else, they are classified as incomplete (light blue). A lipoprotein homolog is detected for a strain if the similarity to the used reference is higher than 50%.

the capacity to synthesize SF-A, SF-B, or STP itself, suggesting a strong dependence on the metabolic output of other community members.

While our primary focus was on SF-A, SF-B, and STP, the presence of another siderophore *aureochelin* has been suggested (Courcol *et al.* 1997). However, the molecule is not characterized, and biosynthetic gene clusters are not known. In our study, we also did not identify any biosynthetic gene cluster corresponding to this compound. This likely reflects current limitations in reference databases, such as the MIBiG database, which lack entries for these siderophores and thus prevent their detection by tools like *antiSMASH*. A similar issue applies to the NRPS system in *S. lugdunensis*, previously suggested to have iron-binding potential (Heilbronner *et al.* 2011). However, due to the absence of a known reference BGC, this system is annotated solely as an unclassified NRPS cluster, missing its putative function as a metallophore. Close inspection of the *antiSMASH* entry along with knowledge regarding the siderophore equibactin was needed to associate the NRPS with metallophore activity. Accordingly, our analysis might underestimate the occurrence of metallophore biosynthesis gene clusters. Similar problems associated with the bioinformatic mining of genomes for siderophore BGCs have also been reported by others (Butler *et al.* 2023; Gu, Shao, Rehm, *et al.* 2024; Reitz and Medema 2022).

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

Our study has analyzed 77 representatives of the *Staphylococci* genus, with a detailed analysis of four species, with a total analysis of over 1,800 strains for their potential of metallophores synthesis and uptake. For three of the in-depth strain analyses, we identify that the representatives reflected a similar profile of most of the analyzed strains. Nevertheless, we have also emphasized how for some strains a strain-dependency is observed, such as the cases for *S. urelyticus*, *S. capitis* and *S. lugdunensis*. Additional strain-level analyses of other specific species can provide additional insight into intra-species diversity. This phenomenon is most likely of immense ecological relevance. The emergence of cheaters put the biosynthesis and receptor genes under constant evolutionary pressure to create privatized siderophores (Kramer *et al.* 2020). This phenomenon is well studied, for example, in *Vibrionaceae* (Cordero *et al.* 2012) or *Pseudomonadaceae* (Gu, Shao, Qu, *et al.* 2025). In the latter genus, the advanced understanding of siderophore-modification and associated receptor variation has allowed accurate prediction of interactions on the strain level. Our study shows that cheaters specialized in the acquisition of staphyloferrin A, B and staphylopin produced by other cells of the same or of another species seem prominent within the genus *Staphylococcus*, suggesting conserved ecological principles. Accordingly, it is tempting to speculate that also metallophore variants and associated variants of HtsA, SirA and CntA might exist between staphylococcal species and strains. However, up to the best of our knowledge, this is not described today and further experimental evidence will be needed to assess this.

While we found divergence of the full-length receptor alleles, the amino acid residues known to be crucial for metallophore-binding were, with few exceptions, conserved between species. Additionally, the crucial amino acids of the HtsA protein showed perfect conservation between hundreds of *S. epidermidis* isolates. This suggests that diversification of staphyloferrin A and the associated receptors to create privatized variants might not be a prominent phenomenon. However, further biochemical characterization of staphyloferrin A derived from diverse species/strains is needed to support this hypothesis.

Moreover, our analyses investigated the uptake potential with respect to the similarity of the membrane-anchored receptors. However, the metallophore uptake mechanisms rely on further proteins, such as ABC transporters and ATPases (Cotton *et al.* 2009; Dale, Sebulsky, *et al.* 2004; Ghseini, Brutescio, *et al.* 2016). Hence, further analyses could expand on this study to identify the diversity of a complete uptake system.

In summary, our study provides a comprehensive description of metallophore-related systems in staphylococcal species, revealing both conserved features and diverse potentials. These findings lay the foundation for future experimental work to characterize novel BGC products and to further investigate the ecological and clinical implications of metallophore diversity, metabolic interdependencies, and cheating strategies within this genus.

Data availability

The scripts used for the data acquisition and processing, as well as for the generation of Figures 2 and 6, can be found via Zenodo at <https://doi.org/10.5281/zenodo.17201851>. The accession codes of all 1928 assemblies throughout this study can be found in the repository as well as a supplementary file.

Competing Interests

No competing interest is declared.

Author Contributions Statement

All authors conceived the idea. MWP conducted the computational analyses and created the figures for the results. MWP and SH wrote the first draft of the manuscript. All authors reviewed the final manuscript.

Acknowledgments and Funding

We acknowledge infrastructural funding from the Cluster of Excellence EXC 2124 ‘Controlling Microbes to Fight Infections’ [project ID 390838134] from the German Research Foundation (DFG). The authors acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 37/935-1 FUGG to KN, INST 37/1080-1 to SH, and HE 8381/4-1 to SH. We acknowledge support from the Open Access Publication Fund of the University of Tübingen.

References

- Beasley, F. C., Marolda, C. L., Cheung, J., Buac, S., and Heinrichs, D. E. (2011). “*Staphylococcus aureus* Transporters Hts, Sir, and Sst Capture Iron Liberated from Human Transferrin by Staphyloferrin A, Staphyloferrin B, and Catecholamine Stress Hormones, Respectively, and Contribute to Virulence”. In: *Infection and Immunity* vol. 79, no. 6, pp. 2345–2355. ISSN: 1098-5522. DOI: [10.1128/IAI.00117-11](https://doi.org/10.1128/IAI.00117-11).
- Beasley, F. C., Vinés, E. D., Grigg, J. C., Zheng, Q., Liu, S., Lajoie, G. A., Murphy, M. E. P., and Heinrichs, D. E. (2009). “Characterization of Staphyloferrin A Biosynthetic and Transport Mutants in *Staphylococcus aureus*”. In: *Molecular Microbiology* vol. 72, no. 4, pp. 947–963. ISSN: 1365-2958. DOI: [10.1111/j.1365-2958.2009.06698.x](https://doi.org/10.1111/j.1365-2958.2009.06698.x).
- Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J. N., van Wezel, G. P., Medema, M. H., and Weber, T. (2023). “antiSMASH 7.0: New and Improved Predictions for Detection, Regulation, Chemical Structures and Visualisation”. In: *Nucleic Acids Research* vol. 51, no. W1, W46–W50. ISSN: 0305-1048. DOI: [10.1093/nar/gkad344](https://doi.org/10.1093/nar/gkad344).
- Bomar, L., Brugger, S. D., and Lemon, K. P. (2018). “Bacterial Microbiota of the Nasal Passages across the Span of Human Life”. In: *Current Opinion in Microbiology* vol. 41, pp. 8–14. ISSN: 1879-0364. DOI: [10.1016/j.mib.2017.10.023](https://doi.org/10.1016/j.mib.2017.10.023).
- Brozyna, J. R., Sheldon, J. R., and Heinrichs, D. E. (2014). “Growth Promotion of the Opportunistic Human Pathogen, *Staphylococcus lugdunensis*, by Heme, Hemoglobin, and Coculture with *Staphylococcus aureus*”. In: *MicrobiologyOpen* vol. 3, no. 2, pp. 182–195. ISSN: 2045-8827. DOI: [10.1002/mbo3.162](https://doi.org/10.1002/mbo3.162).
- Butler, A., Jelowicki, A. M., Ogasawara, H. A., Reitz, Z. L., Stow, P. R., and Thomsen, E. (2023). “Mining Elements of Siderophore Chirality Encoded in Microbial Genomes”. In: *FEBS letters* vol. 597, no. 1, pp. 134–140. ISSN: 1873-3468. DOI: [10.1002/1873-3468.14539](https://doi.org/10.1002/1873-3468.14539).
- Cabrera, G., Xiong, A., Uebel, M., Singh, V. K., and Jayaswal, R. K. (2001). “Molecular Characterization of the Iron-Hydroxamate Uptake System in *Staphylococcus aureus*”. In: *Applied and Environmental Microbiology* vol. 67, no. 2, pp. 1001–1003. ISSN: 0099-2240. DOI: [10.1128/AEM.67.2.1001-1003.2001](https://doi.org/10.1128/AEM.67.2.1001-1003.2001).

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

- Cordero, O. X., Ventouras, L.-A., DeLong, E. F., and Polz, M. F. (2012). “Public Good Dynamics Drive Evolution of Iron Acquisition Strategies in Natural Bacterioplankton Populations”. In: *Proceedings of the National Academy of Sciences of the United States of America* vol. 109, no. 49, pp. 20059–20064. ISSN: 1091-6490. DOI: [10.1073/pnas.1213344109](https://doi.org/10.1073/pnas.1213344109).
- Cotton, J. L., Tao, J., and Balibar, C. J. (2009). “Identification and Characterization of the *Staphylococcus Aureus* Gene Cluster Coding for Staphyloferrin A”. In: *Biochemistry* vol. 48, no. 5, pp. 1025–1035. ISSN: 1520-4995. DOI: [10.1021/bi801844c](https://doi.org/10.1021/bi801844c).
- Courcol, R. J., Trivier, D., Bissinger, M. C., Martin, G. R., and Brown, M. R. (1997). “Siderophore Production by *Staphylococcus aureus* and Identification of Iron-Regulated Proteins”. In: *Infection and Immunity* vol. 65, no. 5, pp. 1944–1948. ISSN: 0019-9567, 1098-5522. DOI: [10.1128/iai.65.5.1944-1948.1997](https://doi.org/10.1128/iai.65.5.1944-1948.1997).
- Dale, S. E., Doherty-Kirby, A., Lajoie, G., and Heinrichs, D. E. (2004). “Role of Siderophore Biosynthesis in Virulence of *Staphylococcus aureus*: Identification and Characterization of Genes Involved in Production of a Siderophore”. In: *Infection and Immunity* vol. 72, no. 1, pp. 29–37. ISSN: 0019-9567. DOI: [10.1128/IAI.72.1.29-37.2004](https://doi.org/10.1128/IAI.72.1.29-37.2004).
- Dale, S. E., Sebulsky, M. T., and Heinrichs, D. E. (2004). “Involvement of SirABC in Iron-Siderophore Import in *Staphylococcus Aureus*”. In: *Journal of Bacteriology* vol. 186, no. 24, pp. 8356–8362. ISSN: 0021-9193. DOI: [10.1128/JB.186.24.8356-8362.2004](https://doi.org/10.1128/JB.186.24.8356-8362.2004).
- Flannagan, R. S., Brozyna, J. R., Kumar, B., Adolf, L. A., Power, J. J., Heilbronner, S., and Heinrichs, D. E. (2022). “In Vivo Growth of *Staphylococcus lugdunensis* Is Facilitated by the Concerted Function of Heme and Non-Heme Iron Acquisition Mechanisms”. In: *The Journal of Biological Chemistry* vol. 298, no. 5, p. 101823. ISSN: 1083-351X. DOI: [10.1016/j.jbc.2022.101823](https://doi.org/10.1016/j.jbc.2022.101823).
- Galpern, P., Manseau, M., Hettinga, P., Smith, K., and Wilson, P. (2012). “AlleleMatch: An R Package for Identifying Unique Multilocus Genotypes Where Genotyping Error and Missing Data May Be Present”. In: *Molecular Ecology Resources* vol. 12, no. 4, pp. 771–778. ISSN: 1755-0998. DOI: [10.1111/j.1755-0998.2012.03137.x](https://doi.org/10.1111/j.1755-0998.2012.03137.x).
- Ghssein, G., Brutesco, C., Ouerdane, L., Fojcik, C., Izaute, A., Wang, S., Hajjar, C., Lobinski, R., Lemaire, D., Richaud, P., Voulhoux, R., Espaillet, A., Cava, F., Pignol, D., Borezée-Durant, E., and Arnoux, P. (2016). “Biosynthesis of a Broad-Spectrum Nicotianamine-like Metallophore in *Staphylococcus aureus*”. In: *Science* vol. 352, no. 6289, pp. 1105–1109. ISSN: 1095-9203. DOI: [10.1126/science.aaf1018](https://doi.org/10.1126/science.aaf1018).
- Ghssein, G. and Ezzeddine, Z. (2022). “The Key Element Role of Metallophores in the Pathogenicity and Virulence of *Staphylococcus aureus*: A Review”. In: *Biology* vol. 11, no. 10, p. 1525. ISSN: 2079-7737. DOI: [10.3390/biology11101525](https://doi.org/10.3390/biology11101525).
- Gilchrist, C. L. M. and Chooi, Y.-H. (2021). “Clinker & Clustermap.Js: Automatic Generation of Gene Cluster Comparison Figures”. In: *Bioinformatics* vol. 37, no. 16. Ed. by Robinson, P., pp. 2473–2475. ISSN: 1367-4803, 1367-4811. DOI: [10.1093/bioinformatics/btab007](https://doi.org/10.1093/bioinformatics/btab007).
- Grigg, J. C., Cheung, J., Heinrichs, D. E., and Murphy, M. E. P. (2010). “Specificity of Staphyloferrin B Recognition by the SirA Receptor from *Staphylococcus aureus*”. In: *The Journal of Biological Chemistry* vol. 285, no. 45, pp. 34579–34588. ISSN: 1083-351X. DOI: [10.1074/jbc.M110.172924](https://doi.org/10.1074/jbc.M110.172924).
- Grigg, J. C., Cooper, J. D., Cheung, J., Heinrichs, D. E., and Murphy, M. E. P. (2010). “The *Staphylococcus aureus* Siderophore Receptor HtsA Undergoes Localized Conformational Changes to Enclose Staphyloferrin A in an Arginine-Rich Binding Pocket”. In: *The Journal of Biological Chemistry* vol. 285, no. 15, pp. 11162–11171. ISSN: 1083-351X. DOI: [10.1074/jbc.M109.097865](https://doi.org/10.1074/jbc.M109.097865).

- Gu, S., Shao, Y., Rehm, K., Bigler, L., Zhang, D., He, R., Xu, R., Shao, J., Jousset, A., Friman, V.-P., Bian, X., Wei, Z., Kümmerli, R., and Li, Z. (2024). “Feature Sequence-Based Genome Mining Uncovers the Hidden Diversity of Bacterial Siderophore Pathways”. In: *eLife* vol. 13, RP96719. ISSN: 2050-084X. DOI: [10.7554/eLife.96719](https://doi.org/10.7554/eLife.96719).
- Gu, S., Shao, Z., Qu, Z., Zhu, S., Shao, Y., Zhang, D., Allen, R., He, R., Shao, J., Xiong, G., Jousset, A., Friman, V.-P., Wei, Z., Kümmerli, R., and Li, Z. (2025). “Siderophore Synthetase-Receptor Gene Coevolution Reveals Habitat- and Pathogen-Specific Bacterial Iron Interaction Networks”. In: *Science Advances* vol. 11, no. 3, eadq5038. ISSN: 2375-2548. DOI: [10.1126/sciadv.adq5038](https://doi.org/10.1126/sciadv.adq5038).
- Heather, Z., Holden, M. T. G., Steward, K. F., Parkhill, J., Song, L., Challis, G. L., Robinson, C., Davis-Poynter, N., and Waller, A. S. (2008). “A Novel Streptococcal Integrative Conjugative Element Involved in Iron Acquisition”. In: *Molecular Microbiology* vol. 70, no. 5, pp. 1274–1292. ISSN: 1365-2958. DOI: [10.1111/j.1365-2958.2008.06481.x](https://doi.org/10.1111/j.1365-2958.2008.06481.x). PMID: [18990191](https://pubmed.ncbi.nlm.nih.gov/18990191/).
- Heilbronner, S., Holden, M. T. G., van Tonder, A., Geoghegan, J. A., Foster, T. J., Parkhill, J., and Bentley, S. D. (2011). “Genome Sequence of *Staphylococcus lugdunensis* N920143 Allows Identification of Putative Colonization and Virulence Factors”. In: *FEMS microbiology letters* vol. 322, no. 1, pp. 60–67. ISSN: 1574-6968. DOI: [10.1111/j.1574-6968.2011.02339.x](https://doi.org/10.1111/j.1574-6968.2011.02339.x).
- Koedooder, C., Zhang, F., Wang, S., Basu, S., Haley, S. T., Tolic, N., Nicora, C. D., Glavina Del Rio, T., Dyhrman, S. T., Gledhill, M., Boiteau, R. M., Rubin-Blum, M., and Shaked, Y. (2023). “Taxonomic Distribution of Metabolic Functions in Bacteria Associated with *Trichodesmium* Consortia”. In: *mSystems* vol. 8, no. 6. Ed. by Raina, J.-B., e00742–23. ISSN: 2379-5077. DOI: [10.1128/msystems.00742-23](https://doi.org/10.1128/msystems.00742-23).
- Kraemer, S. M., Duckworth, O. W., Harrington, J. M., and Schenkeveld, W. D. C. (2015). “Metallophores and Trace Metal Biogeochemistry”. In: *Aquatic Geochemistry* vol. 21, no. 2, pp. 159–195. ISSN: 1573-1421. DOI: [10.1007/s10498-014-9246-7](https://doi.org/10.1007/s10498-014-9246-7).
- Kramer, J., Özkaya, Ö., and Kümmerli, R. (2020). “Bacterial Siderophores in Community and Host Interactions”. In: *Nature Reviews. Microbiology* vol. 18, no. 3, pp. 152–163. ISSN: 1740-1534. DOI: [10.1038/s41579-019-0284-4](https://doi.org/10.1038/s41579-019-0284-4).
- Lex, A., Gehlenborg, N., Strobel, H., Vuilleumot, R., and Pfister, H. (2014). “UpSet: Visualization of Intersecting Sets”. In: *IEEE transactions on visualization and computer graphics* vol. 20, no. 12, pp. 1983–1992. ISSN: 1941-0506. DOI: [10.1109/TVCG.2014.2346248](https://doi.org/10.1109/TVCG.2014.2346248).
- Lindsay, J. A., Riley, T. V., and Mee, B. J. (1994). “Production of Siderophore by Coagulase-Negative *Staphylococci* and Its Relation to Virulence”. In: *European Journal of Clinical Microbiology & Infectious Diseases: Official Publication of the European Society of Clinical Microbiology* vol. 13, no. 12, pp. 1063–1066. ISSN: 0934-9723. DOI: [10.1007/BF02111829](https://doi.org/10.1007/BF02111829).
- Lonergan, Z. R. and Skaar, E. P. (2019). “Nutrient Zinc at the Host-Pathogen Interface”. In: *Trends in Biochemical Sciences* vol. 44, no. 12, pp. 1041–1056. ISSN: 0968-0004. DOI: [10.1016/j.tibs.2019.06.010](https://doi.org/10.1016/j.tibs.2019.06.010).
- Murdoch, C. C. and Skaar, E. P. (2022). “Nutritional Immunity: The Battle for Nutrient Metals at the Host-Pathogen Interface”. In: *Nature Reviews. Microbiology* vol. 20, no. 11, pp. 657–670. ISSN: 1740-1534. DOI: [10.1038/s41579-022-00745-6](https://doi.org/10.1038/s41579-022-00745-6).
- Navarro-Muñoz, J. C., Selem-Mojica, N., Mullaney, M. W., Kautsar, S. A., Tryon, J. H., Parkinson, E. I., De Los Santos, E. L., Yeong, M., Cruz-Morales, P., Abubucker, S., Roeters, A., Lokhorst, W., Fernandez-Guerra, A., Cappelini, L. T. D., Goering, A. W., Thomson, R. J., Metcalf, W. W., Kelleher, N. L., Barona-Gomez, F., and Medema,

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

- M. H. (2020). “A Computational Framework to Explore Large-Scale Biosynthetic Diversity”. In: *Nature chemical biology* vol. 16, no. 1, pp. 60–68. ISSN: 1552-4450. DOI: [10.1038/s41589-019-0400-9](https://doi.org/10.1038/s41589-019-0400-9).
- Neumann, W., Gulati, A., and Nolan, E. M. (2017). “Metal Homeostasis in Infectious Disease: Recent Advances in Bacterial Metallophores and the Human Metal-Withholding Response”. In: *Current Opinion in Chemical Biology* vol. 37, pp. 10–18. ISSN: 1879-0402. DOI: [10.1016/j.cbpa.2016.09.012](https://doi.org/10.1016/j.cbpa.2016.09.012).
- Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2022). “GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy”. In: *Nucleic Acids Research* vol. 50, no. D1, pp. D785–D794. ISSN: 0305-1048. DOI: [10.1093/nar/gkab776](https://doi.org/10.1093/nar/gkab776).
- Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). “NCBI Reference Sequences (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins”. In: *Nucleic Acids Research* vol. 35, no. Database, pp. D61–D65. ISSN: 0305-1048, 1362-4962. DOI: [10.1093/nar/gkl842](https://doi.org/10.1093/nar/gkl842).
- Reitz, Z. L. and Medema, M. H. (2022). “Genome Mining Strategies for Metallophore Discovery”. In: *Current Opinion in Biotechnology* vol. 77, p. 102757. ISSN: 1879-0429. DOI: [10.1016/j.copbio.2022.102757](https://doi.org/10.1016/j.copbio.2022.102757).
- Rempel, S., Stanek, W. K., and Slotboom, D. J. (2019). “ECF-Type ATP-Binding Cassette Transporters”. In: *Annual Review of Biochemistry* vol. 88, pp. 551–576. ISSN: 1545-4509. DOI: [10.1146/annurev-biochem-013118-111705](https://doi.org/10.1146/annurev-biochem-013118-111705).
- Reydam, H., Toledo-Silva, B., Mertens, K., Piepers, S., Vereecke, N., Souza, F. N., Haesebrouck, F., and De Vliegher, S. (2024). “Phenotypic and Genotypic Assessment of Iron Acquisition in Diverse Bovine-Associated Non-aureus *Staphylococcal* Strains”. In: *Veterinary Research* vol. 55, no. 1, p. 6. ISSN: 1297-9716. DOI: [10.1186/s13567-023-01260-z](https://doi.org/10.1186/s13567-023-01260-z).
- Schaible, U. E. and Kaufmann, S. H. E. (2004). “Iron and Microbial Infection”. In: *Nature Reviews. Microbiology* vol. 2, no. 12, pp. 946–953. ISSN: 1740-1526. DOI: [10.1038/nrmicro1046](https://doi.org/10.1038/nrmicro1046).
- Sebulsky, M. T., Hohnstein, D., Hunter, M. D., and Heinrichs, D. E. (2000). “Identification and Characterization of a Membrane Permease Involved in Iron-Hydroxamate Transport in *Staphylococcus aureus*”. In: *Journal of Bacteriology* vol. 182, no. 16, pp. 4394–4400. ISSN: 0021-9193. DOI: [10.1128/JB.182.16.4394-4400.2000](https://doi.org/10.1128/JB.182.16.4394-4400.2000).
- Sebulsky, M. T. and Heinrichs, D. E. (2001). “Identification and Characterization of *fhuD1* and *fhuD2*, Two Genes Involved in Iron-Hydroxamate Uptake in *Staphylococcus aureus*”. In: *Journal of Bacteriology* vol. 183, no. 17, pp. 4994–5000. ISSN: 0021-9193, 1098-5530. DOI: [10.1128/JB.183.17.4994-5000.2001](https://doi.org/10.1128/JB.183.17.4994-5000.2001).
- Sebulsky, M. T., Speziali, C. D., Shilton, B. H., Edgell, D. R., and Heinrichs, D. E. (2004). “FhuD1, a Ferric Hydroxamate-Binding Lipoprotein in *Staphylococcus aureus*: A Case of Gene Duplication and Lateral Transfer”. In: *The Journal of Biological Chemistry* vol. 279, no. 51, pp. 53152–53159. ISSN: 0021-9258. DOI: [10.1074/jbc.M409793200](https://doi.org/10.1074/jbc.M409793200).
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). “Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega”. In: *Molecular Systems Biology* vol. 7, no. 1, p. 539. ISSN: 1744-4292, 1744-4292. DOI: [10.1038/msb.2011.75](https://doi.org/10.1038/msb.2011.75).
- Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M., and Carriço, J. A. (2018). “chewBBACA: A Complete Suite for Gene-by-Gene Schema

- Creation and Strain Identification”. In: *Microbial Genomics* vol. 4, no. 3, e000166. ISSN: 2057-5858. DOI: [10.1099/mgen.0.000166](https://doi.org/10.1099/mgen.0.000166).
- Song, L., Zhang, Y., Chen, W., Gu, T., Zhang, S.-Y., and Ji, Q. (2018). “Mechanistic Insights into Staphylopin-Mediated Metal Acquisition”. In: *Proceedings of the National Academy of Sciences* vol. 115, no. 15, pp. 3942–3947. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1718382115](https://doi.org/10.1073/pnas.1718382115).
- Steinegger, M. and Söding, J. (2017). “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets”. In: *Nature Biotechnology* vol. 35, no. 11, pp. 1026–1028. ISSN: 1087-0156, 1546-1696. DOI: [10.1038/nbt.3988](https://doi.org/10.1038/nbt.3988).
- Tamura, K., Stecher, G., and Kumar, S. (2021). “MEGA11: Molecular Evolutionary Genetics Analysis Version 11”. In: *Molecular Biology and Evolution* vol. 38, no. 7. Ed. by Battistuzzi, F. U., pp. 3022–3027. ISSN: 1537-1719. DOI: [10.1093/molbev/msab120](https://doi.org/10.1093/molbev/msab120).
- Terlouw, B. R., Blin, K., Navarro-Muñoz, J. C., Avalon, N. E., Chevrette, M. G., Egbert, S., Lee, S., Meijer, D., Recchia, M. J. J., Reitz, Z. L., van Santen, J. A., Selem-Mojica, N., Tørring, T., Zaroubi, L., Alanjary, M., Aleti, G., Aguilar, C., Al-Salihi, S. A. A., Augustijn, H. E., Avelar-Rivas, J. A., Avitia-Domínguez, L. A., Barona-Gómez, F., Bernaldo-Agüero, J., Bielinski, V. A., Biermann, F., Booth, T. J., Carrion Bravo, V. J., Castelo-Branco, R., Chagas, F. O., Cruz-Morales, P., Du, C., Duncan, K. R., Gavriilidou, A., Gayraud, D., Gutiérrez-García, K., Haslinger, K., Helfrich, E. J. N., van der Hooft, J. J. J., Jati, A. P., Kalkreuter, E., Kalyvas, N., Kang, K. B., Kautsar, S., Kim, W., Kunjapur, A. M., Li, Y.-X., Lin, G.-M., Loureiro, C., Louwen, J. J. R., Louwen, N. L. L., Lund, G., Parra, J., Philmus, B., Pourmohsenin, B., Pronk, L. J. U., Rego, A., Rex, D. A. B., Robinson, S., Rosas-Becerra, L. R., Roxborough, E. T., Schorn, M. A., Scobie, D. J., Singh, K. S., Sokolova, N., Tang, X., Udvary, D., Vigneshwari, A., Vind, K., Vromans, S. P. J. M., Waschulin, V., Williams, S. E., Winter, J. M., Witte, T. E., Xie, H., Yang, D., Yu, J., Zdouc, M., Zhong, Z., Collemare, J., Linington, R. G., Weber, T., and Medema, M. H. (2023). “MIBiG 3.0: A Community-Driven Effort to Annotate Experimentally Validated Biosynthetic Gene Clusters”. In: *Nucleic Acids Research* vol. 51, no. D1, pp. D603–D610. ISSN: 0305-1048. DOI: [10.1093/nar/gkac1049](https://doi.org/10.1093/nar/gkac1049).
- Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). “Jalview Version 2—a Multiple Sequence Alignment Editor and Analysis Workbench”. In: *Bioinformatics* vol. 25, no. 9, pp. 1189–1191. ISSN: 1367-4811, 1367-4803. DOI: [10.1093/bioinformatics/btp033](https://doi.org/10.1093/bioinformatics/btp033).
- Weinberg, E. D. (2000). “Microbial Pathogens with Impaired Ability to Acquire Host Iron”. In: *Biomaterials: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine* vol. 13, no. 1, pp. 85–89. ISSN: 0966-0844. DOI: [10.1023/a:1009293500209](https://doi.org/10.1023/a:1009293500209).
- Wertheim, H. F. L., Melles, D. C., Vos, M. C., van Leeuwen, W., van Belkum, A., Verbrugh, H. A., and Nouwen, J. L. (2005). “The Role of Nasal Carriage in *Staphylococcus aureus* Infections”. In: *The Lancet. Infectious Diseases* vol. 5, no. 12, pp. 751–762. ISSN: 1473-3099. DOI: [10.1016/S1473-3099\(05\)70295-4](https://doi.org/10.1016/S1473-3099(05)70295-4).
- Wertheim, H. F. L., Vos, M. C., Ott, A., van Belkum, A., Voss, A., Kluytmans, J. A. J. W., van Keulen, P. H. J., Vandenbroucke-Grauls, C. M. J. E., Meester, M. H. M., and Verbrugh, H. A. (2004). “Risk and Outcome of Nosocomial *Staphylococcus aureus* Bacteraemia in Nasal Carriers versus Non-Carriers”. In: *Lancet (London, England)* vol. 364, no. 9435, pp. 703–705. ISSN: 1474-547X. DOI: [10.1016/S0140-6736\(04\)16897-9](https://doi.org/10.1016/S0140-6736(04)16897-9).
- Zackular, J. P., Chazin, W. J., and Skaar, E. P. (2015). “Nutritional Immunity: S100 Proteins at the Host-Pathogen Interface”. In: *The Journal of Biological Chemistry* vol. 290, no. 31, pp. 18991–18998. ISSN: 1083-351X. DOI: [10.1074/jbc.R115.645085](https://doi.org/10.1074/jbc.R115.645085).

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

- Zhao, Y., Bitzer, A., Power, J. J., Belikova, D., Torres Salazar, B. O., Adolf, L. A., Gerlach, D., Krismer, B., and Heilbronner, S. (2024). “Nasal Commensals Reduce *Staphylococcus aureus* Proliferation by Restricting Siderophore Availability”. In: *The ISME journal* vol. 18, no. 1, wræ123. ISSN: 1751-7370. DOI: [10.1093/ismejo/wrae123](https://doi.org/10.1093/ismejo/wrae123).
- Zipperer, A., Konnerth, M. C., Laux, C., Berscheid, A., Janek, D., Weidenmaier, C., Burian, M., Schilling, N. A., Slavetinsky, C., Marschal, M., Willmann, M., Kalbacher, H., Schitteck, B., Brötz-Oesterhelt, H., Grond, S., Peschel, A., and Krismer, B. (2016). “Human Commensals Producing a Novel Antibiotic Impair Pathogen Colonization”. In: *Nature* vol. 535, no. 7613, pp. 511–516. ISSN: 1476-4687. DOI: [10.1038/nature18634](https://doi.org/10.1038/nature18634).

Supplementary Material

Table IV.S1: Overview of species and assemblies used in the analysis. The classification from the NCBI and the GTDB is provided.

NCBI Accession Code	NCBI classification	GTDB classification
GCA_000433035.1	<i>Staphylococcus</i> sp. CAG:324	CAG-313 sp000433035
GCA_020544305.1	<i>Staphylococcus taiwanensis</i>	<i>Staphylococcus taiwanensis</i>
GCA_023665865.1	<i>Staphylococcus</i> sp.	<i>Duncaniella</i> sp023665865
GCA_945870555.1	<i>Staphylococcus epidermidis</i>	UBA970 sp945870555
GCF_000010125.1	<i>Staphylococcus saprophyticus</i> subsp. saprophyticus ATCC 15305 = NCTC 7292	<i>Staphylococcus saprophyticus</i>
GCF_000236925.1	<i>Staphylococcus argenteus</i>	<i>Staphylococcus argenteus</i>
GCF_000298075.1	<i>Staphylococcus massiliensis</i> CCUG 559 27	<i>Staphylococcus massiliensis</i>
GCF_000816085.1	<i>Staphylococcus hyicus</i>	<i>Staphylococcus hyicus</i>
GCF_000875895.1	<i>Staphylococcus gallinarum</i>	<i>Staphylococcus gallinarum</i>
GCF_000934465.1	<i>Staphylococcus microti</i>	<i>Staphylococcus microti</i>
GCF_001027105.1	<i>Staphylococcus aureus</i>	<i>Staphylococcus aureus</i>
GCF_001224225.1	<i>Staphylococcus borealis</i>	<i>Staphylococcus borealis</i>
GCF_001431205.1	<i>Staphylococcus</i> sp. NAM3COL9	<i>Staphylococcus</i> sp001431205
GCF_001500315.1	<i>Staphylococcus auricularis</i>	<i>Staphylococcus auricularis</i>
GCF_001618885.1	<i>Staphylococcus condimentii</i>	<i>Staphylococcus condimentii</i>
GCF_001792775.2	<i>Staphylococcus pseudintermedius</i>	<i>Staphylococcus pseudintermedius</i>
GCF_002101335.1	<i>Staphylococcus lutrae</i>	<i>Staphylococcus lutrae</i>
GCF_002614725.1	<i>Staphylococcus edaphicus</i>	<i>Staphylococcus edaphicus</i>
GCF_002732165.1	<i>Staphylococcus xylosus</i>	<i>Staphylococcus xylosus</i>
GCF_002901705.1	<i>Staphylococcus lugdunensis</i>	<i>Staphylococcus lugdunensis</i>
GCF_002901845.1	<i>Staphylococcus hominis</i> subsp. <i>hominis</i>	<i>Staphylococcus hominis</i>
GCF_002901865.1	<i>Staphylococcus agnetis</i>	<i>Staphylococcus agnetis</i>
GCF_002901995.1	<i>Staphylococcus coagulans</i>	<i>Staphylococcus coagulans</i>
GCF_002902085.1	<i>Staphylococcus simiae</i>	<i>Staphylococcus simiae</i>
GCF_002902145.1	<i>Staphylococcus rostri</i>	<i>Staphylococcus rostri</i>
GCF_002902235.1	<i>Staphylococcus ureilyticus</i>	<i>Staphylococcus ureilyticus</i>

GCF_002902285.1	<i>Staphylococcus simulans</i>	<i>Staphylococcus simulans</i>
GCF_002902305.1	<i>Staphylococcus argensis</i>	<i>Staphylococcus argensis</i>
GCF_002902325.1	<i>Staphylococcus capitis subsp. capitis</i>	<i>Staphylococcus capitis</i>
GCF_002902345.1	<i>Staphylococcus arlettae</i>	<i>Staphylococcus arlettae</i>
GCF_002902365.1	<i>Staphylococcus cohnii subsp. cohnii</i>	<i>Staphylococcus cohnii</i>
GCF_002902385.1	<i>Staphylococcus intermedius NC TC 11048</i>	<i>Staphylococcus intermedius</i>
GCF_002902405.1	<i>Staphylococcus schweitzeri</i>	<i>Staphylococcus schweitzeri</i>
GCF_002902565.1	<i>Staphylococcus petrasii</i>	<i>Staphylococcus petrasii</i>
GCF_002902575.1	<i>Staphylococcus croceilyticus</i>	<i>Staphylococcus croceilyticus</i>
GCF_002902625.1	<i>Staphylococcus devriesei</i>	<i>Staphylococcus devriesei</i>
GCF_002902685.1	<i>Staphylococcus pettenkoferi</i>	<i>Staphylococcus pettenkoferi</i>
GCF_002902725.1	<i>Staphylococcus caprae</i>	<i>Staphylococcus caprae</i>
GCF_003012915.1	<i>Staphylococcus felis</i>	<i>Staphylococcus felis</i>
GCF_003019255.1	<i>Staphylococcus kloosii</i>	<i>Staphylococcus kloosii</i>
GCF_003019275.1	<i>Staphylococcus muscae</i>	<i>Staphylococcus muscae</i>
GCF_003697915.1	<i>Staphylococcus pseudoxylus</i>	<i>Staphylococcus pseudoxylus</i>
GCF_003718735.1	<i>Staphylococcus debuckii</i>	<i>Staphylococcus debuckii</i>
GCF_003970495.1	<i>Staphylococcus pasteurii</i>	<i>Staphylococcus pasteurii</i>
GCF_004785665.1	<i>Staphylococcus pragensis</i>	<i>Staphylococcus pragensis</i>
GCF_006094395.1	<i>Staphylococcus haemolyticus</i>	<i>Staphylococcus haemolyticus</i>
GCF_006742205.1	<i>Staphylococcus epidermidis</i>	<i>Staphylococcus epidermidis</i>
GCF_010365305.1	<i>Staphylococcus sp. MI 10-1553</i>	<i>Staphylococcus sp010365305</i>
GCF_013391405.1	<i>Staphylococcus sp. GSSP0090</i>	<i>Staphylococcus sp013391405</i>
GCF_013463155.1	<i>Staphylococcus sp. 17KM0847</i>	<i>Staphylococcus sp013463155</i>
GCF_014635045.1	<i>Staphylococcus nepalensis</i>	<i>Staphylococcus nepalensis</i>
GCF_015594545.1	<i>Staphylococcus durrellii</i>	<i>Staphylococcus durrellii</i>
GCF_015775975.1	<i>Staphylococcus lloydii</i>	<i>Staphylococcus lloydii</i>
GCF_016238445.1	<i>Staphylococcus canis</i>	<i>Staphylococcus canis</i>
GCF_016238465.1	<i>Staphylococcus caledonicus</i>	<i>Staphylococcus caledonicus</i>
GCF_017583065.1	<i>Staphylococcus shinii</i>	<i>Staphylococcus shinii</i>
GCF_020883535.1	<i>Staphylococcus ratti</i>	<i>Staphylococcus ratti</i>
GCF_022346615.1	<i>Staphylococcus sp. ACRSN</i>	<i>Staphylococcus sp022346615</i>
GCF_022493055.1	<i>Staphylococcus roterodami</i>	<i>Staphylococcus singaporensis</i>
GCF_022815905.2	<i>Staphylococcus sp. NRL 16/872</i>	<i>Staphylococcus sp022815905</i>
GCF_024814435.1	<i>Staphylococcus americanisciur i</i>	<i>Staphylococcus americanisciuri</i>
GCF_025519785.1	<i>Staphylococcus marylandisciur i</i>	<i>Staphylococcus marylandisciuri</i>
GCF_025558425.1	<i>Staphylococcus sp. IVB6240</i>	<i>Staphylococcus sp025558425</i>

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*

GCF_025558585.1	<i>Staphylococcus sp. IVB6214</i>	<i>Staphylococcus sp025558585</i>
GCF_026659735.1	<i>Staphylococcus pettenkoferi</i>	<i>Staphylococcus sp943908435</i>
GCF_029024625.1	<i>Staphylococcus chromogenes</i>	<i>Staphylococcus chromogenes</i>
GCF_029024945.1	<i>Staphylococcus succinus</i>	<i>Staphylococcus succinus</i>
GCF_029024965.1	<i>Staphylococcus equorum</i>	<i>Staphylococcus equorum</i>
GCF_900097965.1	<i>Staphylococcus caeli</i>	<i>Staphylococcus caeli</i>
GCF_900183575.1	<i>Staphylococcus cornubiensis</i>	<i>Staphylococcus cornubiensis</i>
GCF_900186985.1	<i>Staphylococcus piscifermentans</i>	<i>Staphylococcus piscifermentans</i>
GCF_900458435.1	<i>Staphylococcus carnosus</i>	<i>Staphylococcus carnosus</i>
GCF_900458815.1	<i>Staphylococcus saccharolyticus</i>	<i>Staphylococcus saccharolyticus</i>
GCF_900458895.1	<i>Staphylococcus schleiferi</i>	<i>Staphylococcus schleiferi</i>
GCF_900636325.1	<i>Staphylococcus delphini</i>	<i>Staphylococcus delphini</i>
GCF_900636385.1	<i>Staphylococcus warneri</i>	<i>Staphylococcus warneri</i>
GCF_943737015.1	<i>Staphylococcus sp. Marseille- Q5304</i>	<i>Staphylococcus sp943737015</i>

Table IV.S2: Overview of additional strains used to assess the variability of the antiSMASH results within selected species.

Analysed for metabolite	Species	Accession Code	Similarity to annotated BGC in %			Number of uncharacterized BGC
			SF-A	SF-B	STP	
SF-A	<i>S. argensis</i>	GCF_007671915.1	75	0	0	1
		GCF_007682325.1	75	0	0	1
		GCF_007682285.1	75	0	0	1
		GCF_002902305.1	75	0	0	1
		GCF_026740095.1	75	0	0	1
	<i>S. equorum</i>	GCF_029531845.1	75	100	0	0
		GCF_030249045.1	75	100	0	0
		GCF_033447145.1	50	100	0	0
		GCF_001747785.1	75	100	0	0
		GCF_001747895.1	75	100	0	0
	<i>S. pettenkoferi</i>	GCF_002884615.1	75	0	0	1
		GCF_026660515.1	75	0	0	1
		GCF_007678805.1	75	0	0	1
		GCF_026659915.1	75	0	0	1
		GCF_030218345.1	75	0	0	1

Table IV.S3: Analysis of binding site conservation of *htsA*, *sirA* and *cntA*.

Position	Amino Acid	Identical	Total Sequences	Identical %
<i>htsA</i>				
86	R	46	67	68.66
104	R	67	67	100.00
126	R	67	67	100.00
203	K	46	67	68.66
209	H	67	67	100.00
239	Y	67	67	100.00
299	R	67	67	100.00
304	R	46	67	68.66
306	R	67	67	100.00
<i>sirA</i>				
81	W	19	19	100.00
125	R	19	19	100.00
144	T	18	19	94.74
201	R	19	19	100.00
206	R	18	19	94.74
208	Y	19	19	100.00
304	N	19	19	100.00
<i>cntA</i>				
52	Y	29	29	100.00
128	W	29	29	100.00
165	R	29	29	100.00
250	R	29	29	100.00
418	R	29	29	100.00
431	W	29	29	100.00
435	Y	29	29	100.00
448	N	13	29	44.83
522	Y	29	29	100.00

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*



Figure IV.S1: Multiple sequence alignment of identified homologs of *HtsA*. Visualized using Jalview (Waterhouse *et al.* 2009)

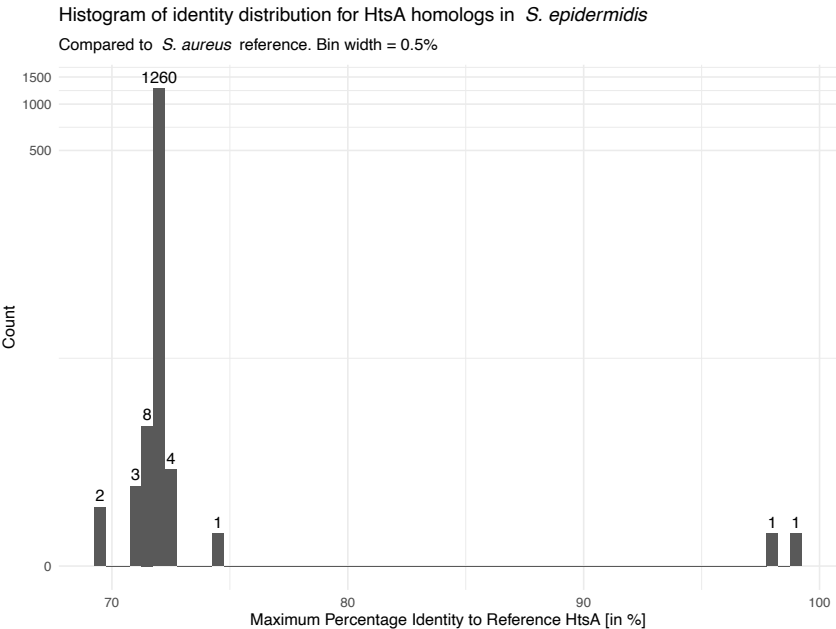
AA076_RS00330|GCF_001027105.1|Staphylococcus_aureus WRTRRYN
 SAMSHR1132_RS00465|GCF_000236925.1|Staphylococcus_argenteus
 ML436_RS00345|GCF_022493055.1|Staphylococcus_singaporensis
 CD116_RS00940|GCF_002902405.1|Staphylococcus_schweitzeri
 CD041_RS10505|GCF_002901705.1|Staphylococcus_lugdunensis ..N.....
 CCE82_RS05305|GCF_900183575.1|Staphylococcus_cornubiensis
 BJK46_RS05250|GCF_001792775.2|Staphylococcus_pseudintermedius
 EL101_RS11370|GCF_900636325.1|Staphylococcus_delphini
 CD138_RS02245|GCF_002902385.1|Staphylococcus_intermedius
 GZH82_RS11655|GCF_010365305.1|Staphylococcus_sp010365305
 CD118_RS08195|GCF_002901995.1|Staphylococcus_coagulans
 JM183_RS00365|GCF_900458895.1|Staphylococcus_schleiferi
 CD036_RS07075|GCF_002902345.1|Staphylococcus_arlettae
 B5P37_RS07585|GCF_002101335.1|Staphylococcus_lutrae
 PYW44_RS00580|GCF_029024965.1|Staphylococcus_equorum
 LN051_RS00225|GCF_020883535.1|Staphylococcus_ratti
 CD172_RS11315|GCF_002901865.1|Staphylococcus_agnetis
 SHYC_RS00645|GCF_000816085.1|Staphylococcus_hyicus
 PYW36_RS00295|GCF_029024625.1|Staphylococcus_chromogenesC..

Figure IV.S2: Multiple sequence alignment of identified homologs of *SirA*. Visualized using Jalview (Waterhouse *et al.* 2009)

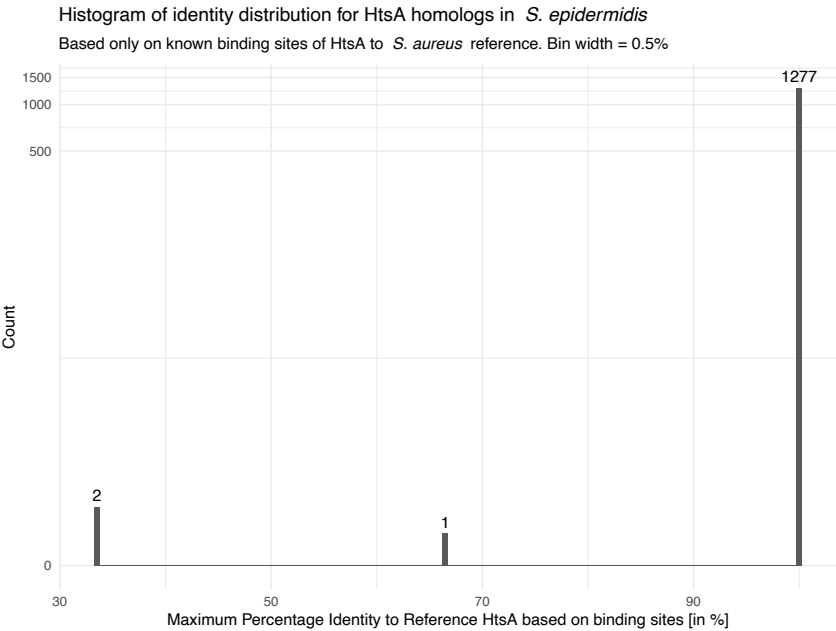
AA076_RS12655|GCF_001027105.1|Staphylococcus_aureus YWRRRWYN
 CD116_RS09400|GCF_002902405.1|Staphylococcus_schweitzeri
 ML436_RS12530|GCF_022493055.1|Staphylococcus_singaporensis
 SAMSHR1132_RS12165|GCF_000236925.1|Staphylococcus_argenteus
 ISP02_RS00280|GCF_015594545.1|Staphylococcus_durrellii
 J5E45_RS06025|GCF_017583065.1|Staphylococcus_shinii
 ISP08_RS00455|GCF_015775975.1|Staphylococcus_lloydii
 BTM19_RS04735|GCF_002732165.1|Staphylococcus_xylosus
 D9V42_RS05940|GCF_003697915.1|Staphylococcus_pseudoxylosus
 EL082_RS00425|GCF_900636385.1|Staphylococcus_warneriT.
 CD155_RS06575|GCF_002902725.1|Staphylococcus_capraeT.
 CD121_RS00345|GCF_003970495.1|Staphylococcus_pasteuriT.
 CD041_RS10300|GCF_002901705.1|Staphylococcus_lugdunensis
 FNL83_RS00850|GCF_006742205.1|Staphylococcus_epidermidisT.
 BJS11_RS04660|GCF_900097965.1|Staphylococcus_caeli
 HYE69_RS08110|GCF_013391405.1|Staphylococcus_sp013391405
 BTJ66_RS05135|GCF_002614725.1|Staphylococcus_edaphicus
 C7J90_RS03780|GCF_003012915.1|Staphylococcus_felisT.
 CD172_RS06415|GCF_002901865.1|Staphylococcus_agnetisT.
 CD118_RS05655|GCF_002901995.1|Staphylococcus_coagulansT.
 EL101_RS12590|GCF_900636325.1|Staphylococcus_delphiniT.
 SHYC_RS11365|GCF_000816085.1|Staphylococcus_hyicusT.
 GZH82_RS13255|GCF_010365305.1|Staphylococcus_sp010365305T.
 PYW36_RS10780|GCF_029024625.1|Staphylococcus_chromogenesT.
 CCE82_RS07340|GCF_900183575.1|Staphylococcus_cornubiensisT.
 LN051_RS10970|GCF_020883535.1|Staphylococcus_rattiT.
 BJK46_RS00445|GCF_001792775.2|Staphylococcus_pseudintermediusT.
 CD138_RS05270|GCF_002902385.1|Staphylococcus_intermediusT.
 B5P37_RS06440|GCF_002101335.1|Staphylococcus_lutraeT.

Figure IV.S3: Multiple sequence alignment of identified homologs of *CntA*. Visualized using Jalview (Waterhouse *et al.* 2009)

IV. A landscape of metallophore synthesis and uptake potential of the genus *Staphylococcus*



(a) Accounts full length sequence for the comparison.



(b) Accounts only for known binding sites.

Figure IV.S4: Distribution of sequence identify of HtsA homologs in *S. epidermidis* when compared to the reference of *S. aureus*.

Appendix V

Characterizing Siderophore Uptake and their Regulation Mechanisms in *Corynebacteria* based on Structural Homology and Transcriptome Profiling

Mathias Witte Paz, Alina Bitzer, Simon Heilbronner, Kay Nieselt

Manuscript in progress

Abstract

Iron acquisition is critical for bacterial proliferation in nutrient-limited environments such as the human nasal environment. To overcome this, *Staphylococcus aureus* produces siderophores like staphyloferrin A and B for iron uptake. Other bacterial species typically present in the nasal microbiome are, for example, *Corynebacteria*. Some species of this genera have been shown to exploit these siderophores without synthesizing them, an observed mechanism of siderophore cheating. To investigate the molecular mechanisms behind this phenotype, the pipeline **PRESERVE** has been commonly developed and applied to multiple *Corynebacteria* and *S. aureus* samples. By integrating a lipoprotein prediction method, sequence- and structure-based approaches, and motif enrichment analysis, we were able to characterize a set of 52 strains from 7 *Corynebacterium* species, as well as 8 *S. aureus* strains. We were able to identify a set of 444 structural homologs of *S. aureus* siderophore-binding lipoproteins, revealing extensive conservation across *Corynebacteria* despite low sequence similarity. The integration of transcriptomic experiments under iron-deprived conditions showed a significant up-regulation of the homologs, as well as enriched siderophore-related gene ontology terms. The identified homologs were analyzed further by an orthology analysis across *Corynebacteria*, confirming the conservation across the species, but highlighted limitations in linking structural similarity to specific siderophore phenotypes. Lastly, promoter motif analyses showed a difference in that *Corynebacteria* are regulated by the Iron-dependent Regulator (IdeR) and not the *S. aureus* associated ferric uptake regulator (Fur). Overall, our findings provide a multi-layered framework for understanding iron uptake strategies in *Corynebacteria*, with implications for microbial competition and niche colonization in host-associated microbiomes.

Keywords: Structure-based Homology, Iron-uptake Mechanisms, Transcriptomic Regulation, Transcriptome Profiling

Introduction

Bacteria require a diverse palette of nutrients for their proliferation, including iron (Wooldridge and Williams 1993). However, iron molecules are hardly soluble in many environments, such as the nasal mucosa (Adolf and Heilbronner 2022). To overcome such challenging environments, bacteria require mechanisms with high iron affinities to proliferate. One of these mechanisms are siderophores (SIDs), secondary metabolites that are excreted from the cell to form an iron-SID complex, before being transported back to the cell (Saha *et al.* 2013). SIDs are synthesized by multiple genes located within biosynthetic gene clusters (BGC). In general, BGCs are sets of genes in a proximal location mainly responsible for the synthesis of metabolites. In the case of siderophores, four different types are defined depending on the structure of the metabolite: hydroxamates, catecholates, carboxylates, and mixed-type siderophores (Saha *et al.* 2013).

Besides synthesis genes, BGCs can also include other genes related to the general function of siderophores, such as the uptake of the SID-iron complex. Regardless of the type, the general mode of action for the uptake of the SID-iron complex within gram-positive bacteria is similar. Outer membrane lipoproteins interact with the SID-iron complex to signal ATP-binding cassette (ABC) transporter units to channel the substrate into the cell (Saha *et al.* 2013; Schalk 2025). However, lipoproteins typically exhibit specificity towards particular classes of structurally related siderophores, or even individual, specific siderophores (Caza and Kronstad 2013). For example, the gram-positive pathogen *Staphylococcus aureus*, the most prominent pathogen in adults for the nasal environment, is able to synthesize two different carboxylate-type siderophores. While staphyloferrin A and B (SF-A and SF-B) belong to the same siderophore type, they differ in their synthesis, but more importantly in their uptake mechanisms. SF-A uptake is induced by the *htsABC* operon, with HtsA as the respective lipoprotein in this operon, but the uptake of SF-B requires the operon *sirABC* and SirA being the lipoprotein (Beasley *et al.* 2011). Moreover, *S. aureus* is able to uptake a larger repertoire of siderophores, although it is not capable of synthesizing them, hence termed as xenosiderophore usage (Conroy *et al.* 2019). This has been observed with the transporter system *sstABCD* for the usage of catechol-type siderophores (Morrissey *et al.* 2000), and the operon *fhuBGC2-D2/D1* to uptake hydroxamate siderophores (Sebulsky and Heinrichs 2001).

Other organisms are also present in the nasal environment, with members of the *Corynebacterium* genera being the most abundant species (Liu *et al.* 2015). Of the 170 species that are known for *Corynebacteria*, only a few of them have been identified as pathogenic, such as *C. diphtheriae* (Cunningham *et al.* 2025). More interestingly, it has been shown that a high abundance of members of this species correlates with the absence of opportunistic pathogens, such as *S. aureus* (Bomar *et al.* 2016; Kelly *et al.* 2022). Due to the previously described usage of *xenosiderophore* in *S. aureus*, siderophores could be expected to play an important role in this reciprocal competition and support between organisms in the nasal environment, as has been shown for other microbiomes (Cordero *et al.* 2012; Kramer *et al.* 2020).

Recently, Zhao *et al.* (2024) studied the interaction via siderophores between *S. aureus* and other species, mainly from the genus *Corynebacterium*. They identified an important role of siderophores for the establishment of specific communities within the nasal environment, even though only one of the studied corynebacterial species (*C. propinquum*) was able to synthesize siderophores by itself. Due to the special phenotype of *Corynebacteria* regarding siderophore acquisition, the hypothesis was formulated that their genomes show different states of the siderophore receptor genes SF-A and SF-B. Zhao *et al.* (2024)

were able to characterize two lipoproteins for the species *C. hesseae*. However, further high-throughput computational analyses are required to identify the molecular mechanisms that explain the iron-uptake mechanisms observed for multiple *Corynebacteria* samples. Hence, the goal of this work is to investigate the iron-regulon of *Corynebacteria* to identify iron-regulated lipoproteins that might serve as staphyloferrin receptors. For this, a homology-based computational pipeline was established to identify SID-related genes. This was complemented by a comparison of the transcriptome of four different *Corynebacteria* species grown in two media (one iron-rich and one iron-depleted).

Multiple tools have been developed to identify genes related to siderophore production and uptake. One commonly used tool is **antiSMASH** (Blin *et al.* 2023), a hidden Markov model (HMM) based approach to identify and annotate BGCs. Although **antiSMASH** is also able to annotate proteins related to metabolite transport, it requires the identification of core synthesis genes for annotation. Since these genes are missing in *cheating* organisms such as *Corynebacteria*, **antiSMASH** cannot be used to identify the uptake mechanisms. Other studies have exploited the sequence similarity of lipoproteins using machine learning algorithms, such as HMMs, to decipher the key-lock principles within SID-producing and exploiting species (Gu *et al.* 2025). The focus of their study was the diversity of SID exploitation within *Pseudomonas*. While sequence similarity can provide insight into similar functions of proteins, it can be limiting, especially when comparing species from distinct genera, as in the case of *S. aureus* and *Corynebacteria* species.

In addition to identifying the necessary mechanisms for xenosiderophore usage, we aim to explore how they are regulated in *Corynebacteria* and compared that to the regulation mechanisms of *S. aureus*. In the latter organism, genes related to iron homeostasis are regulated by ferric uptake regulator (*Fur*) (Xiong *et al.* 2000). Identifying whether this regulating mechanism is also conserved in siderophore-cheating *Corynebacteria* might help gain more insight on these mechanisms and also identify undiscovered iron-related genes in the species.

To address these goals, we proceed as follows. For a set of multiple *Corynebacteria* species, we identified putative homologs to known lipoproteins of *S. aureus*, thus predicting putative staphyloferrin receptors. Since diverse protein sequences can be folded into the same protein structure, the sequence-based homology analysis is complemented by a structure-based conservation analysis (Koehl 2001; Shatsky *et al.* 2008), providing more reliable bases for comparison between genera. To take advantage of advances in computational methods for the prediction of protein structure, we developed pipeline **PRESERVE**¹ (**P**rotein conservation survey) for this analysis. **PRESERVE** is a computational approach focused on the structural prediction process of sequence proteomes originating from multiple species. In this study, the pipeline has been applied to go beyond protein structure prediction and compare the structures to a reference set of structures. This enhancement makes **PRESERVE** a generalizable approach that combines homology search methods based on the primary and tertiary structure of the protein. Moreover, the transcriptomic profile of four selected *Corynebacteria* species was assessed in an iron-deprived condition via RNA-seq, to understand how *Corynebacteria* react in such conditions and to contextualize the findings of **PRESERVE**. Lastly, **PRESERVE** was further developed to gain insight into the regulation mechanisms of the identified genes by including a downstream analysis on the promoter regions. By this thorough characterization of the SID uptake systems in a cohort of nasal *Corynebacteria*, we aim to understand the mechanisms for the use of such xenosiderophores.

¹Commonly developed with Simon Hackl, <https://github.com/Integrative-Transcriptomics/preserve-nextflow>

V. Characterizing Siderophore Uptake and their Regulation Mechanisms in *Corynebacteria* based on Structural Homology and Transcriptome Profiling

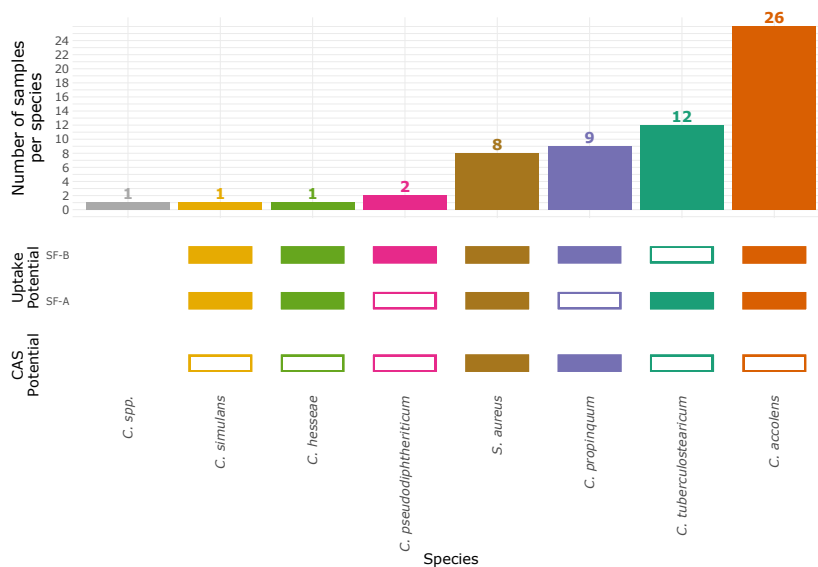


Figure V.1: Overview of the dataset used for the computational characterization of the mechanistic siderophore uptake potential in *Corynebacteria* and *Staphylococcus aureus* samples. It shows the distribution of samples per species (upper bar chart). The other heatmaps show the results of the experimental characterization presented by Zhao *et al.* (2024), where the uptake potential of specific siderophores was analyzed via co-cultures with *S. aureus* mutants (middle binary heatmap) as well as their potential of synthesizing any siderophore (lower binary heatmap). Full rectangles of the two lower heatmap represent the potential of the species to uptake the corresponding siderophore (uptake potential) or to produce any siderophore (CAS potential), while rectangles without a color fill represent that the species is not able to uptake them or produce them. Empty spaces represent species that were not characterized in the study.

Material & Methods

Materials

To characterize iron uptake mechanisms in *Corynebacteria* a set of 52 strains covering different species of nasal *Corynebacteria* has been collected. Moreover, eight samples for the nasal pathogen *Staphylococcus aureus* have been included in the dataset for pipeline validation purposes.

Most samples originate from two collections sampled by the group led by Simon Heilbronner: the LaCa (Camus *et al.* 2025) and the Münster collection (Zhao *et al.* 2024) (see Table V.S1 for a detailed overview). For all strains, a sequenced genome and an annotation computed with BAKTA have been provided. In addition, a further commercial strain of *Corynebacterium accolens* was added to our dataset (NCBI accession ID: GCF_000478135.1). The strains have been experimentally tested for their potential to uptake staphyloferrin A and B (co-culture experiments), as well as the potential to synthesize any siderophore (chrome azurol S, short CAS, assay), as described by Zhao *et al.* (2024). A short overview of the covered species and their expected SID uptake and synthesis potential can be found in Fig. V.1.

To investigate the potential of siderophore uptake in *Corynebacteria*, five lipoproteins responsible for uptake in *Staphylococcus aureus* were identified from the literature: *HtsA* (Grigg, Cooper, *et al.* 2010), *SirA* (Grigg, Cheung, *et al.* 2010), *FhuD1*, *FhuD2* (Sebulsky and Heinrichs 2001) and *SstD* (Beasley *et al.* 2011). Their protein structures were acquired from the PDB (Burley *et al.* 2017) or from the AlphaFold

Table V.1: Origin of reference protein structures used for structural comparison. Reference structures of five *Staphylococcus aureus* siderophore-binding lipoproteins (HtsA, SirA, FhuD1, FhuD2, SstD) were retrieved from the PDB or the AlphaFold database, depending on availability.

Reference Protein	Origin	ID
HtsA	PDB	3EIW
SirA	PDB	3MWG
FhuD1	AlphaFold DB	A0A0H2WWK9
FhuD2	AlphaFold DB	A0A0H2WZDT
SstD	AlphaFold DB	A0A0H2WYC5

database (Varadi *et al.* 2022), depending on the availability of the data (Table V.1).

For the transcriptomics analysis, representative strains of four *Corynebacterium* species (*C. hesseae*: CH, *C. simulans*: CS, *C. propinquum*: CP and *C. accolens*: CA) were selected for an RNA-seq experiment based on their diverse phenotypic profile of utilizing *S. aureus*' siderophores, as well as some interactions observed in the study by Zhao *et al.* (2024), such as their ability to inhibit or support *S. aureus*. It should be noted that CP is the only corynebacterial species in the data set with the potential to produce siderophores, as shown by the CAS assay (Fig. V.1). Three strains were selected from the previously described collections (Münster: CH and CS, LaCa: CP), while the sample for CA was a commercial strain. The RNA-seq experiment consisted of measuring gene expression under two conditions: an iron-deprived environment with the EDDHA chelator vs. a control environment. Three biological replicates per condition were collected.

Prior analysis of validation samples

The eight samples of *S. aureus* were analyzed independently of the PRESERVE pipeline to confirm the presence of all five reference lipoproteins in their annotated coding sequences. Due to assembly and/or annotation problems, it is possible that the corresponding coding sequences have been annotated incorrectly, and hence introduce a bias in our validation set. For this, the amino acid sequence of all five reference lipoproteins was searched against all annotated coding sequences of the samples using MMseqs2 (Hauser *et al.* 2016). If the coding sequence returned a hit with sequence similarity more than 0.9 and covers the reference lipoprotein by more than 0.99, the corresponding lipoprotein is considered present in the sample.

The PRESERVE pipeline

PRESERVE combines different computational methods to efficiently identify homologous lipoproteins between multiple proteomes and a set of reference structures (for an overview, see Fig. V.2). The first step of the pipeline is to determine the proteins of interest; in this case, lipoproteins. Here, SignalP6 (Teufel *et al.* 2022) is applied to each proteome to identify lipoproteins transported by the Sec (labeled as *LIPO*) or by the Tat translocon (*TATLIPO*) and cleaved by the Signal Peptidase II. To address a possible redundancy across all samples, the sequences of the identified lipoproteins are clustered using MMseqs2 (Hauser *et al.* 2016), prior to structure prediction. MMseqs2 is used with a sequence similarity threshold of at least 0.45 and a coverage of at least 0.8. Moreover, MMseqs2 identifies a sequence as the representative of each cluster.

The structures of the resulting MMseqs2 cluster representatives are then predicted using ColabFold (Mirdita *et al.* 2022) and compared to the reference set of lipoprotein structures

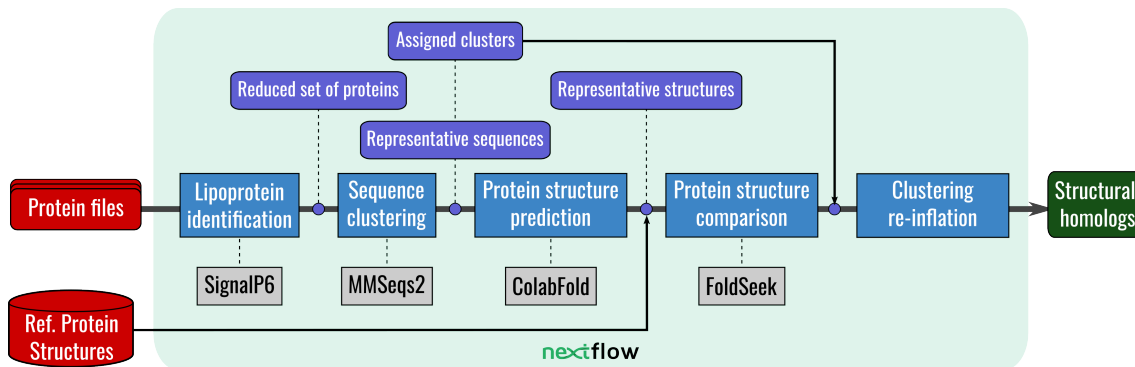


Figure V.2: Overview of the pipeline PRESERVE. The protein input files (red) are processed (blue) with tools (gray) to produce interim (purple) results, such as the identification of lipoproteins, the clustering of them and their structure prediction. The predicted structures are then compared to the input reference structures to produce a set of putative structural homologs.

using FoldSeek (van Kempen *et al.* 2024). To identify putative structural homologs, a minimum local distance difference test (LDDT) score of 0.5 and a coverage of 0.7 have been set as threshold, close to the threshold defined by van Kempen *et al.* (2024) for true positives in a structural comparison study. Since the results of structural comparison are based on the representative sequence, the respective clusters are reinflated, i.e. the information of the representative is extrapolated to all other proteins of the cluster to obtain a per-sample resolution.

RNA-seq analysis and GO-enrichment analysis

The four strains selected for transcriptomic analysis were exposed to two conditions: an iron-deprived environment via the chelator EDDH and a control environment. With three replicates per condition, a total of 24 samples were subject to single-end strand-specific RNA-sequencing using Illumina technology. The resulting 24 FASTQ sequences were processed using an in-house developed Nextflow-based pipeline (Harbig 2023) for read pre-processing, mapping, and gene expression quantification. For quality assessment, FastQC (Andrews 2010) was used to compute common quality metrics, such as the GC content per sequence, the quality scores of the sequences, as well as the analysis of overrepresented sequences. The reads were then mapped to the reference genomes using HISAT2 (Kim *et al.* 2015). Spliced alignments were disabled and the reads were mapped according to the strand-specific protocol (`-no-spliced-alignment` and `-rna-strandness "R"`). The resulting mapping files in SAM format were converted to BAM format using SAMtools (Li *et al.* 2009). For the assessment of the mapping results, QualiMap2 (García-Alcalde *et al.* 2012) was used to provide a general overview. Furthermore, all results of the quality assessment were summarized using MultiQC (Ewels *et al.* 2016) for a better overview of the data. The calculation of the gene count table was performed using featureCounts (Liao *et al.* 2014), allowing for multi-mapping (`-M`) of the reads as well as overlapping (`-0`, `-fracOverlap 0.7`, `-minOverlap 10`). The gene expression quantification was based on the entries labeled as *transcripts*, and with overlapping and multi-mapping reads counted as fractions (`-fraction`).

Differentially expressed genes between the two conditions (control vs. EDDH treated) were identified using DESeq2 (Love *et al.* 2014). Since multi-mapping and overlapping reads were counted as fractions, the count tables produced by featureCounts were rounded to the next integer for their use in DESeq2. Genes with an adjusted *p*-value below 0.01 and

an absolute fold change larger than 1 were considered as differentially expressed.

In addition, a gene ontology (GO) enrichment analysis was performed to identify terms that might be shared by differentially expressed genes between the tested species. For this, the GO background of each species was predicted using the module **FaCoP** from the web-server **FUNAGE-Pro** (de Jong *et al.* 2022). Afterwards, the GO background, the list of differentially expressed genes as well as their fold change value were uploaded to the tool **GO-Compass** (Harbig *et al.* 2023). **GO-Compass** not only conducts a GO-enrichment analysis but also clusters the resulting GO terms with respect to their dispensability. This means that terms that are less relevant receive a higher dispensability value, providing a better overview of the enriched GO terms.

Genes annotated with GO terms related to siderophore uptake were compared to the final set of structural homologous lipoproteins to evaluate the consistency between both analyses.

Downstream analysis of homologous genes with transcriptomic data

To contextualize the results of the **PRESERVE** pipeline, we integrated the set of structural homologous lipoproteins with the transcriptomic analysis results for each species. This allowed us to explore the expression of the genes across both analyzed environments and evaluate the correctness of our identified homologs based on the expression, as we expect true homologous genes to be up-regulated in the treated condition. Moreover, we integrated all identified lipoproteins with **SignalP6** to identify any putative missing lipoprotein involved in iron-acquisition, as they should also be up-regulated in the transcriptomic analysis.

In addition, we compared the identified homologs and their expression across the four species by computing orthologous groups with **OrthoFinder**. Although protein sequences have previously been clustered with **MMseqs2** within the **PRESERVE** framework, the focus was to reduce the redundancy of the sequence space of the proteins. As **OrthoFinder** is specialized in the identification of orthologous genes, it was more suited for this downstream analysis. From all identified groups, we focused on those with at least one structural homolog identified via the **PRESERVE** pipeline.

Analyzing promoter regions of structural homologs

Following the identification of homologous lipoproteins with **PRESERVE**, we analyzed all structural homologs to identify the primary transcription factors involved in their regulation via the identification of putative binding sites. For this, the upstream region of the homologs identified by **PRESERVE** was extracted and analyzed for enriched motifs using a downstream analysis pipeline (Fig. V.3).

As mentioned above, genes related to the uptake of SID-iron complexes, such as lipoproteins and ABC transporters, are found in operons. Due to the co-regulation of operons, we expect the transcription factor binding site to be found upstream of the first gene of the operon (Osbourn and Field 2009). Hence, we identified all putative operons within the input bacterial genomes by grouping genes found within a distance of 50 nt. This threshold was selected to capture the wide range of intergenic distances observed between co-transcribed coding sequences in bacteria (Yada *et al.* 1999). After operon detection, only those groups (either operons or single genes) with at least one identified structural homolog are used further. Based on these sets of genes, the first gene of each group is selected and its upstream region is extracted (400 nt). Such a long upstream sequence is required as annotation files focus on the coding sequence rather than the 5' – UTR, which might hinder the identification of the promoter region and any transcription factor binding

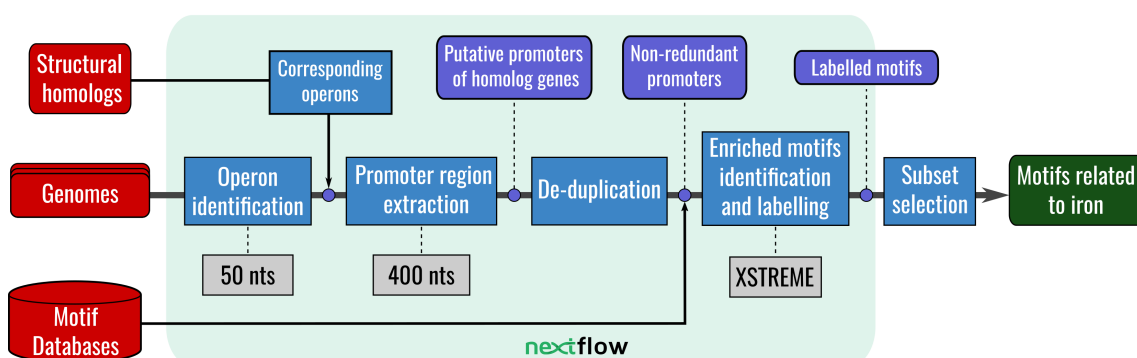


Figure V.3: Overview of the downstream analysis used to identify possible transcription factor binding sites related to the regulation of the identified structural homologs. The input files (red) are processed (blue) with tools or parameters (gray) to produce interim (purple) or final results (green). First, genes are grouped into operons and the upstream region of the first gene is extracted, as long as the operon contains a gene identified as structural homolog. For structural homologs not found in any operon, the direct upstream region of the gene is extracted. To avoid sample bias, identical sequences are de-duplicated and passed to XSTREME (Grant and Bailey 2021) for *de novo* motif identification and labeling via database comparison. The results of XSTREME were investigated manually to get an overview of those motifs related to iron homeostasis.

site. Due to the high sequence similarity between the genomes of closely related strains, it can occur that identical upstream regions are extracted, producing a bias towards specific corynebacterial strains. Hence, the set of extracted upstream sequences was de-duplicated to create a non-redundant set for motif analysis with XSTREME (Grant and Bailey 2021). XSTREME identifies statistically enriched sequence motifs and compares them to a database of known transcription factor binding sites (such as PRODORIC (Dudek and Jahn 2022) or RegTransBase (Cipriano *et al.* 2013)) to identify the corresponding regulator. Based on the XSTREME output, motifs known in the literature to be related to iron regulation were manually selected. For each motif, we computed their occurrence distribution across all analyzed species.

Results

Analysis of validation samples

To verify that the eight *S. aureus* samples used for validation correctly annotate the five lipoproteins used as a reference, a search using MMseqs2 was run against all coding sequences of the samples, expecting high sequence similarity (> 0.9) and high coverage of the reference lipoprotein (> 0.99). For each coding sequence, its best hit with respect to the mentioned parameters was identified (Table V.2 and Suppl. Table V.S2). This search identified that four of the eight samples contain exactly one copy of each reference lipoprotein. The other samples differ from the expected number of lipoproteins. Sample 199 contained two identical copies of the HtsA protein. However, this result should be taken with caution, as the respective genome sequence for this strain had a large number of contigs ($n = 181$), suggesting a bias from a low-quality assembly. Samples 19 and 36 miss a locus for FhuD1 transcription, and sample 36 misses a transcript for SstD.

For the validation purposes of this study, these results reflect that instead of expecting a total of 40 structural homologs for the *S. aureus* samples (eight samples $\times$ five reference proteins), the structural homolog identified can at most identify 38 homologs.

Table V.2: Overview of the number of reference lipoproteins present in *S. aureus* samples used for pipeline validation purposes. These results are based on a MMseqs2 search, considering the best hit for every target protein of the *S. aureus* samples (Supl. Table V.S2).

sample	FhuD1	FhuD2	HtsA	SirA	SstD
sample_1	1	1	1	1	1
sample_2	1	1	1	1	1
sample_3	1	1	1	1	1
sample_19	0	1	1	1	1
sample_36	1	1	1	1	0
sample_145	0	1	1	1	1
sample_186	1	1	1	1	1
sample_199	1	1	2	1	1

Identification and clustering of lipoproteins

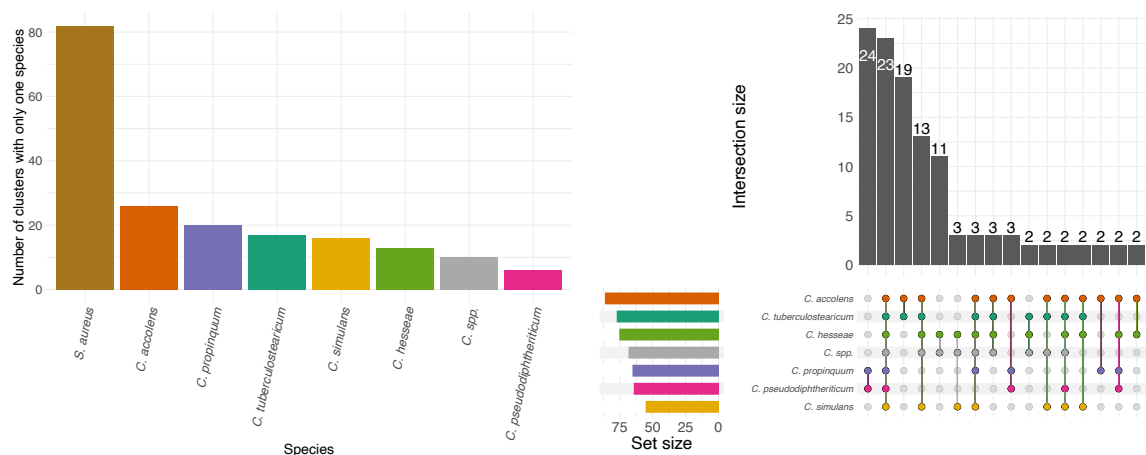
For the 60 analyzed annotated protein sequences (52 from the *Corynebacterium* genus, 8 *S. aureus* strains), SignalP6 predicted between 69 and 107 lipoproteins per strain (Supl. Fig. V.S1), resulting in a total of 4,520 predicted lipoproteins across all strains. Using MMseqs2 for sequence clustering, PRESERVE reduced this amount to a total of 322 clusters. The clusters contain either sequences from only one species, or from two or more species. 190 of 322 (59%) of the clusters contained sequences from just one species (Fig. V.4a), with *S. aureus* having here the most clusters of this class. The remaining clusters contained sequences from two or more of the corynebacterial species (Fig. V.4b and Supl. Fig. V.S2). 23 clusters cover all *Corynebacteria* species, showing a high lipoprotein conservation across the species. Specific species pairs, such as *C. propinquum* and *C. pseudodiphtheriticum*, or *C. accolens* and *C. tuberculoastericum*, also harbor unique lipoproteins not shared with other species and could indicate species-specific mechanisms. Here, is important to note that no cluster showed conserved lipoproteins between *S. aureus* and any of the corynebacterial species (Fig. V.S2). This is due to the low sequence similarity between *S. aureus* and the *Corynebacteria* species.

Structural homolog identification via PRESERVE

After sequence clustering with MMseqs2, a structural homology analysis was performed using ColabFold for structure prediction and Foldseek for the comparison to known *S. aureus* proteins. This returned 28 cluster representatives (21 with sequences of corynebacterial species and 7 for sequences retrieved from *S. aureus* strains) surpassing the defined coverage and LDDT thresholds. After clustering re-inflation, a total of 444 lipoproteins across all 60 samples (63 for *S. aureus* and 381 for corynebacterial samples) showing homologous structures to the known *S. aureus* lipoproteins involved in SID-uptake (Fig. V.5).

Interestingly, for most species, the number of lipoproteins structurally homologous to either of the five *S. aureus* lipoproteins is constant across strains, with *C. tuberculoastericum* having the least ($n = 3$) and *C. propinquum* having the most ($n = 13$) structural homologs. Only *C. accolens* showed a higher lipoprotein diversity across strains (between 4 and 10 homologs), suggesting a possible greater intra-species variability in the interaction with iron-binding molecules.

V. Characterizing Siderophore Uptake and their Regulation Mechanisms in Corynebacteria based on Structural Homology and Transcriptome Profiling



(a) Number of clusters with protein sequences originating from samples labeled as the same species. (b) Upset plot showing the distribution of clusters across multiple species. For simplification, combinations occurring only more than one time are shown. The complete plot can be found in Suppl. Fig. [V.S2](#).

Figure V.4: Overview of the composition for the protein sequence clusters returned by MMseqs2 (Hauser *et al.* [2016](#)). For a species to be accounted for a cluster, at least one protein sequence originating from the species should be found in the cluster. It has to be noted that more sequences from the same sample, or from a different sample but labeled with the same species can be part of the cluster. Two separate views are provided: (a) for clusters with sequences originating from one species and (b) for clusters with protein sequences from multiple species.

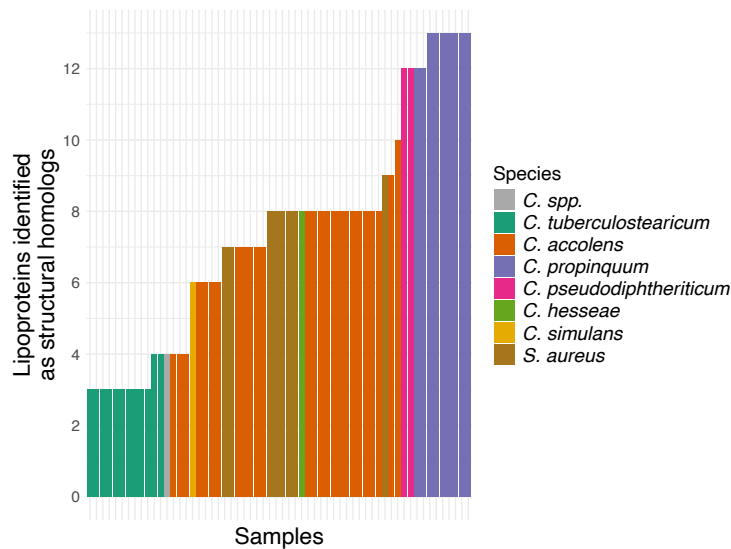


Figure V.5: Number of structural homologs identified via PRESERVE, using Colabfold and Foldseek, after clustering reflation. Each bar is colored with respect to the analyzed species.

Validation of structural homolog identification with *S. aureus* samples

Although for *S. aureus* samples 38 homologous lipoproteins were expected to be identified, a total of 63 structural homologs were found (between 7 and 9, depending on the strain). The identified homologs were further characterized in an independent BLAST search against the UniProt database (Zaru *et al.* [2023](#)), and AureoWiki was used to further characterize the results (Fuchs *et al.* [2018](#)) (Suppl. Table [V.S3](#)). For each identified structural homolog, the best hit was chosen with respect to sequence identity and summarized in a sample-wise

Table V.3: Overview of the identified structural homologs by gene and sample for *S. aureus* strains, based on the *UniProt* search results. The left side of the table shows the expected lipoproteins used as a reference, while the right side shows false positive homologs.

Sample ID	FhuD1	FhuD2	HtsA	SirA	SstD	BtuF	IsdE	SAOUHSC_00976
sample_1	1	1	1	1	1	1	1	1
sample_2	1	1	1	1	1	1	1	1
sample_3	1	1	1	1	1	1	1	1
sample_19	0	1	1	1	1	1	1	1
sample_36	1	1	1	1	0	1	1	2
sample_145	0	1	1	1	1	1	1	1
sample_186	1	1	1	1	1	1	1	1
sample_199	1	1	2	1	1	1	1	1

Table V.4: Overview of the cluster representatives of *S. aureus* identified as structural homologs. The second column shows the gene names identified with the *UniProt* search to be found within the cluster of the representative. Lastly, the last columns show the reference structure that aligned best with the representative and the respective *LDDT* score.

Locus tag of representative	Gene names in cluster	Struc. aligned to	LDDT
OEHLIH_00205	<i>fhuD1</i>	FhuD1	98.67
OKMLLP_05960	<i>sstD</i>	SstD	97.56
OKMLLP_11655	<i>fhuD2</i>	FhuD2	95.90
JNHCPM_11275	<i>sirA</i>	SirA	94.61
KJDHGP_03365	SAOUHSC_00976 & <i>htsA</i>	HtsA	81.45
CGIHNC_05605	<i>btuF</i>	HtsA	59.56
OKMLLP_04230	<i>isdE</i>	FhuD2	58.26

view (Table V.3).

The **UniProt** search confirmed that **PRESERVE** was able to correctly identify the expected reference proteins in *S. aureus* samples, as it mirrors the results of the sequence search with **MMseqs2** (see previous Table V.2). In addition, three different proteins were also found by **PRESERVE**: the vitamin B12 binding protein *BtuF*, the *heme* receptor *IsdE*, and an uncharacterized protein with iron binding potential (locus tag of the target hit: *SAOUHSC_00976*). To further analyze the results of false positives, the clusters and their respective representatives used for structural structural were evaluated with respect to the annotated gene names from the **UniProt** search, as well as with the *LDDT* values of the structural alignment (Table V.4). The analysis showed that almost all clusters consisted of the same gene, with the exception of one cluster which contained *htsA* sequences as well as the uncharacterized protein.

Moreover, it can be seen that the *LDDT* value of the true homologs is higher than the rest. An overview of the complete distribution of *LDDT* values across all structural alignments that exceed the defined thresholds (Fig. V.6) shows that the lowest *LDDT* of the true homologs are generally higher than the *LDDT* values of all corynebacterial hits, with some corynebacterial hits found between the true homologs and false positives of *S. aureus*.

Analysis of RNA-seq data

The RNA-seq experiment showed large changes in gene expression for the four chosen *Corynebacterium* species when exposed to iron-deprived conditions. Each of the four

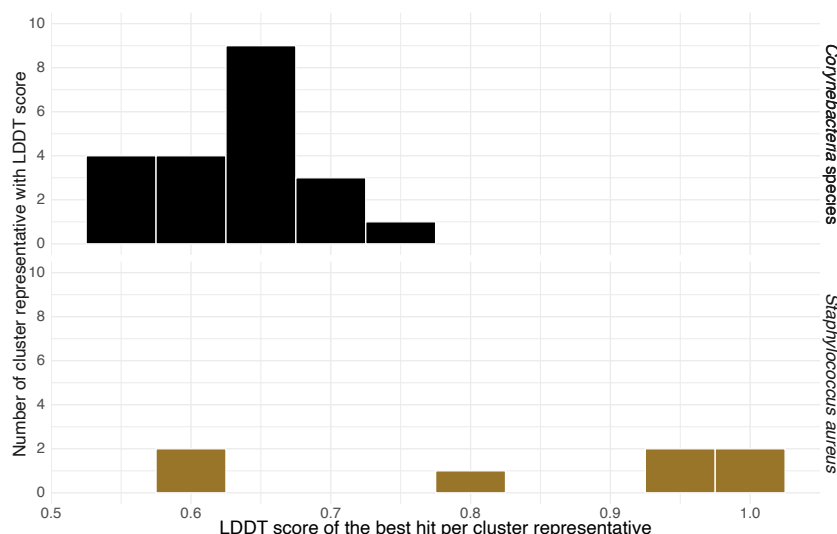


Figure V.6: Histogram on the distribution of *LDDT* scores for all structural homologs, divided by representative with only *S. aureus* sequences or only corynebacterial samples.

species exhibited over 310 differentially expressed genes, using a corrected *p*-value threshold of 0.01 and an absolute $\log_2$ fold change of 1 (Table V.5 and Fig. V.7). Interestingly, the species seem to differ with respect to the ratio of up- and down-regulated genes ($|\log_2(FC)| > 1$). While in *C. hesseae* and *C. propinquum* more differentially expressed genes are up-regulated (56.17% and 55.03%, respectively), the contrary can be seen for *C. accolens* and *C. simulans*, as only 46.95% and 41.04% are up-regulated. The gene ontology enrichment analysis using **GO-Compass** returned a total of 130 enriched GO terms, across all three main ontologies: 80 for *biological processes*, 29 for *cellular component* and 21 for *molecular function*. Here, 23 of the terms were found to be enriched in all four species, such as the terms associated with siderophore-mediated iron transport and iron homeostasis (Suppl. Fig. V.S4), as well as the term *ferric-enterobactin import into the cell*. Although enterobactin is a siderophore originally identified in *Salmonella Typhimurium* (Raymond *et al.* 2003), this GO term underscores the functional necessity for siderophore import systems in the four analyzed *Corynebacterium* species. Species-specific tree maps (Suppl. Figs. V.S5a–V.S5d) further show that iron-related terms were mainly associated with up-regulated genes, while down-regulated genes were not linked to specific processes in the four species.

To evaluate the structural lipoprotein homologs predicted by **PRESERVE** based on the enriched GO terms, we compared the 37 structural homologs identified across the four species with proteins that have been linked to the enriched GO terms (Suppl. Fig. V.S3). As **PRESERVE** only accounts for lipoproteins, proteins linked to GO terms were filtered to only include genes that have been labeled by **SignalP6** as lipoproteins during the **PRESERVE** analysis. Interestingly, only 9 structural homologs overlapped with genes associated with GO annotations. **PRESERVE** identified 28 unique lipoproteins, suggesting that sequenced-based GO annotations alone may miss key components of the siderophore uptake machinery. However, seven genes were not identified by **PRESERVE**. The relevance of these genes in the siderophore-uptake process was further analyzed in a further step.

To further evaluate the structural homologs, we assessed their expression level in the context of the transcriptomic setting. For those lipoproteins identified as structural homologs, most of them are strongly up-regulated under EDDHA treatment (Fig. V.7, red

Table V.5: Summary of differentially expressed genes, distributed by the magnitude of their $\log_2$ fold change ($\log_2(FC)$). Genes with an adj. p -value lower than 0.01 were called differentially expressed.

	CA	CH	CP	CS
Diff. expr. genes (adj. $p < 0.01$)	1,296	1,671	983	1,435
Diff. expr. genes (adj. $p < 0.01$) and filtered for fold change ($ \log_2(FC) > 1$)	311	616	338	541
Upregulated				
Total	659	798	492	683
$\log_2(FC) > 1$	146	346	186	222
$\log_2(FC) > 2$	47	119	105	73
Downregulated				
Total	637	873	491	752
$\log_2(FC) < -1$	165	270	152	319
$\log_2(FC) < -2$	16	96	55	105

triangles), regardless of the overall expression ratio in the species. Only five of the structural homologs (13.5%, Suppl. Table [V.S4](#)), two in *C. accolens* and three in *C. propinquum*, showed low fold-changes and non-differential expression across conditions. However, their functional annotations suggest a substrate-binding role with respect to ABC transporters. The annotations imply a goal similar to that of the known iron-binding lipoproteins in *S. aureus* but with a different role, as they are expressed differently in these *Corynebacteria*.

To explore whether **PRESERVE** may have missed any relevant lipoproteins, we analyzed the 23 significantly up-regulated genes (p -adj. < 0.01 and $\log_2(FC) > 1$) that were not called structural homologs after **PRESERVE**, but identified as lipoproteins by **SignalP6** (Suppl. Table [V.S5](#)). First, we compared these genes to the seven lipoproteins uniquely identified via GO annotation. Only one lipoprotein of this set (marked with a + in Suppl. Table [V.S5](#)) intersected with the up-regulated genes; nevertheless, it does not show a high LDDT score nor a high coverage to the provided reference, suggesting a different kind of lipoprotein involved in iron-uptake. As of the remaining six lipoproteins intersecting with the GO annotation, they do not show a significant up-regulation. This indicates that they might not play a role in siderophore-uptake and hence are not relevant for this study.

The other 22 significantly up-regulated lipoproteins were analyzed with respect to the results of the **Foldseek** comparison between them and the reference lipoproteins (Suppl. Table [V.S5](#)). This analysis showed that most lipoproteins do not resemble any of the reference genes, since structural alignment returns low coverages (< 0.5) or very low LDDT values (< 0.4). This could indicate that these differentially expressed genes represent other types of lipoproteins also involved in the uptake of siderophores, such as ABC transporters (Schalk [2025](#)), as seen in the genes' descriptions (Suppl. Table [V.S5](#)). The only exceptions are four up-regulated proteins (marked with an * in Suppl. Table [V.S5](#)) that returned a high alignment coverage (> 0.9), but LDDT values close to the selected threshold of 0.5 (≈ 0.41). Structural alignments showed that while the protein core shows a high level of conservation, the differences between the *Corynebacterium* gene and the *S. aureus* reference genes are found mainly in unstructured regions, suggesting that these genes are false negatives (an example can be seen in Fig. [V.8](#)).

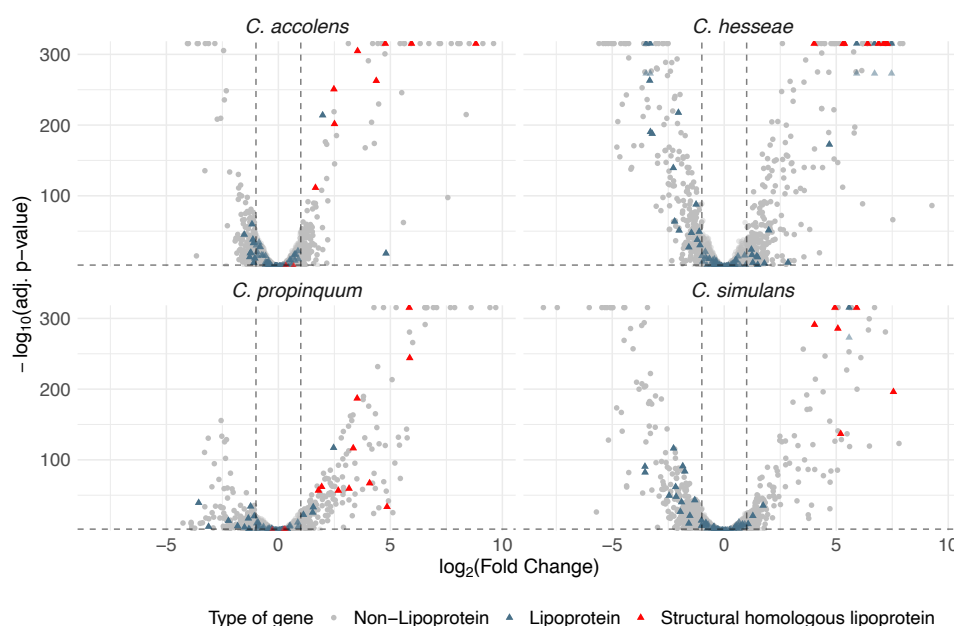


Figure V.7: Volcano plots summarizing the complete transcriptomic response for a control vs. treatment with the chelator EDDHA (Fold change: treated – control). The vertical dashed lines highlight where the absolute $\log_2$ fold-change ($|\log_2(FC)|$) is equal to one, while the horizontal dashed lines show where $-\log_{10}(p - \text{adj.})$ equals 0.01. Additionally, the volcano plot differentiates between lipoproteins identified only by **SignalP6** (blue triangles), lipoproteins additionally identified as structural homologs by **Foldseek** (red triangles) and other genes not identified as lipoproteins (gray circles).

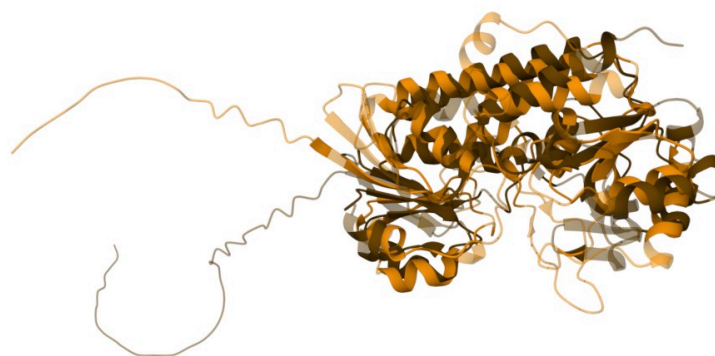


Figure V.8: Structural alignment of gene *RS06480* from *C. accolens* (orange) and the reference lipoprotein *HtsA* from *S. aureus* (brown) with an LDDT of 0.40. Similar structural regions are shown with full opacity, while diverging regions are more transparent. The structural alignment and visualization was computed using the alignment tool from the RCSB PDB (Bittrich *et al.* [2024](#)).

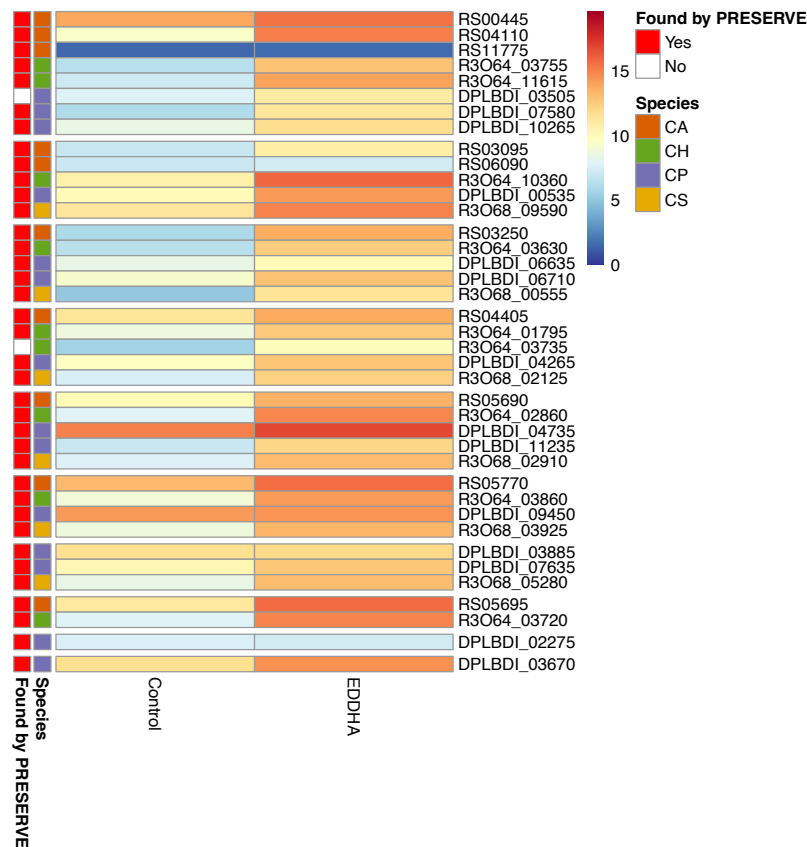


Figure V.9: Heatmap showing median regularized log-transformed expression values per condition for a subset of genes grouped by orthogroups. Only orthogroups containing at least one gene identified by **PRESERVE** as a structural homolog to reference proteins are shown. Each row corresponds to a gene, annotated by its species of origin (color-coded) and whether it was identified by **PRESERVE** (red = yes, white = no). Columns of the heatmap represent the different conditions from the RNA-seq experiment (Control vs. EDDHA).

Cross-species comparison of identified homologs

While the previous section focused on comparing the response of the individual *Corynebacteria* to the iron depletion as well as contextualizing the results of the **PRESERVE** pipeline, the following analyses will focus on the cross-species comparison of the homologs across the four *Corynebacteria*. To investigate the consistency of the results across species, we used orthogroups computed by **OrthoFinder** to compare the structural homologs across the four *Corynebacterium* species for which RNA-seq data are available. We focused only on orthogroups that contained at least one predicted structural homolog. Almost for all orthologous groups all genes were additionally identified as structural homologs after processing with **PRESERVE** (Fig. [V.9](#)). Only two exceptions (genes *DPLBDI_03505* from *C. propinquum* and *R3O64_03735* from *C. accolens*) were identified, as they are part of the filtered orthologous group, but were not called structural homologs.

Detailed analyses of the intermediate steps of **PRESERVE** detected possible reasons for this discrepancy. First, the gene *DPLBDI_03505* was not identified as a lipoprotein by **SignalP6** and thus was directly excluded from any subsequent analysis. However, its annotation as *Fe/B12 periplasmic-binding domain-containing protein* suggests a lipoprotein function, indicating a putative false negative in the lipoprotein prediction step of **PRESERVE**. For the second case, gene *R3O64_03735*, was found to be shorter (128 aa) than the other

V. Characterizing Siderophore Uptake and their Regulation Mechanisms in *Corynebacteria* based on Structural Homology and Transcriptome Profiling

members of its orthogroup (minimal length of 313 aa), as well as all five reference lipoproteins (ranging from 302 to 342 aa). This discrepancy suggests that either **OrthoFinder** allocate this gene into the incorrect orthogroup, or that the structure of the gene did not manage to pass the coverage threshold required for structural homology, leading to its exclusion by **PRESERVE**.

This comparison also allows to compare the expression of the genes across conditions. For this, the normalized expression values based on the regularized log (*RLOG*) provided by **DESeq2** per gene were visualized for both conditions. The corresponding expression heatmap reflects a general increase in the expression of the genes shown when exposed to an iron-deprived condition. In addition, most of the genes in the same group not only share the tendency of up-regulation, but also a similar expression value. This could indicate that the species react also similarly to the exposed conditions. However, it also shows differences in the absolute expression values across conditions. Some genes, for example *DPLBDI_06635* and *DPLBDI_04735*, have differences to the rest of the members of the orthogroup. This could suggest differences in their regulation, despite **Orthofinder** grouping them in the same orthogroup via sequence similarity.

Among the genes with unchanged expression across both conditions, two distinct patterns emerge. The genes *RS11775* and *RS06090* from *C. accolens* and *DPLBDI_02275* from *C. propinquum* show low expression under both conditions, suggesting limited relevance in this environmental context. In contrast, genes *DPLBDI_09450* and *DPLBDI_03885* (also from *C. propinquum*) are constantly highly expressed, suggesting a role that is independent of iron availability. Notably, *C. propinquum* is the only species in the dataset known to synthesize siderophores itself (Fig. [V.1](#)), which could explain the persistent expression of certain uptake-related lipoproteins also in the control environment.

Finally, the identified orthogroups also offer the possibility of comparing the structural homolog predictions with the experimental phenotypes (see Fig. [V.1](#) for experimental metadata). As Zhao *et al.* ([2024](#)) were able to show that the gene *R3O64_11615* is responsible for the uptake of both SF-A and SF-B siderophores in *C. hessae*, it could be expected that the other members of its orthogroup are involved in a similar mechanism. This assumption could be extrapolated to the genes of *C. accolens*, since the experimentally observed phenotype of *C. accolens* and *C. hessae* in the co-cultures by Zhao *et al.* ([2024](#)) is the same. Nevertheless, since *C. propinquum* genes are also found within the orthogroup, it does not explain why this species is not able to uptake SF-A or why any gene of *C. simulans* is missing to fully mirror the experimentally observed SF-B phenotype. Additionally, no other orthogroup reflects the composition of the SF-A phenotype observed experimentally, where only *C. propinquum* lacks the ability to import this siderophore, as the structural homologs of this species are found in almost all orthogroups. These findings suggest that the current combination of structural homology and sequence-based orthology comparisons is insufficient to link specific lipoproteins to individual siderophore usage with high confidence.

Identification of transcription factor binding sites for structural homologs

Besides investigating the expression of structural homologs for selected species, we analyzed the putative promoter regions of the operons containing a structural homologs identified by **PRESERVE**, for all 60 strains of this study. Using **XSTREME**, the promoter regions were analyzed to identify enriched motifs of length between 6 and 15 nts and compared to databases for further characterization. Via manual selection, we reduced those motifs to only those known to be involved in iron regulation in the literature and computed an average number of binding sites found per strain for each species.

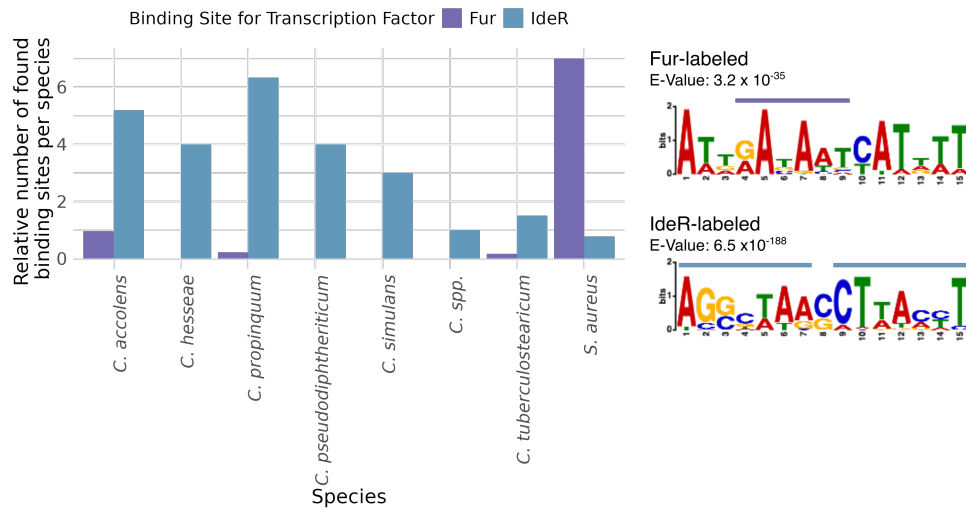


Figure V.10: Overview of the results for motif enrichment across the analyzed species. (left) Relative abundance of transcription factor binding sites associated with iron homeostasis, identified per species. The number of identified motifs was normalized by the number of samples per species, providing an estimate of each transcription factor's overall relevance within the respective species. (right) Motifs identified as binding sites for Fur or IdeR, known iron-regulators. The corresponding consensus binding sequences have been highlighted for each motif (bar on top of the subsequence).

The analysis of the upstream regulatory regions of the operons showed that *Staphylococcus aureus* and the *Corynebacteria* strains have different regulatory mechanisms for the identified genes (Fig. V.10). One statistically significant motif ($E\text{-value} = 3.2 \times 10^{-35}$) was labeled by XSTREME as the binding site for the ferric uptake regulator (Fur) known to be relevant in *S. aureus*, and also showed the expected consensus binding sequence GATAAT (Sheldon and Heinrichs 2012). As expected, *S. aureus* strains have a high number of binding sites for Fur upstream of the identified homologs, correlating also with the amount of identified structural homologs by PRESERVE (7-9 for *S. aureus* samples).

In contrast, analyses of *Corynebacteria* genomes present a different picture (Fig. V.10). Binding sites for Fur were only found sparsely across these species, with *C. accolens* having the most binding sites for this transcription factor, with an average of one binding site per strain. However, the binding site for the Iron-dependent Regulator IdeR was found as the most significant motif for this analysis ($E\text{-value} = 6.5 \times 10^{-188}$). This binding site is found in the upstream regions across all analyzed corynebacterial species, ranging from one up to six times per sample, dependent on the species. However, for *S. aureus* it was only found sparsely, with an average of less than one time per sample. IdeR is a member of the diphtheria toxin repressor (DtxR) family and has been previously characterized to regulate iron in other corynebacterial species, such as *Corynebacterium diphtheria* (Marcos-Torres et al. 2021). Moreover, the identified motif also resembles the described consensus sequence of the literature, as it covers partially the sequence TTAGGTTAG, and is followed by its reverse complement CTAACCTAA (Lee and Holmes 2000; Marcos-Torres et al. 2021). The distribution analysis suggests that IdeR, not Fur, appears to be the primary transcription factor regulating the expression of the iron-related structural homologs in the studied *Corynebacterium* species.

Discussion

Understanding siderophore uptake mechanisms in microbial communities is essential for gaining insight on how bacteria compete for limited resources such as iron (Zhao *et al.* 2024). However, the molecular mechanisms of siderophore uptake are highly diverse, since transport strategies and regulatory architectures can differ substantially between bacterial species, complicating the overall characterization. In this study, we combined structural homology detection, transcriptional analysis, and regulatory motif analysis to characterize the siderophore uptake potential in *Corynebacterium* species. Although certain representatives have been experimentally shown to exploit siderophores produced by *Staphylococcus aureus*, the molecular mechanisms behind this exploitation remain poorly characterized (Zhao *et al.* 2024).

To achieve this goal, we developed the **PRESERVE** pipeline to go beyond structural prediction and integrate sequence- and structure-based methods to detect homologous lipoproteins in distantly related genera and applied it to the strains analyzed in this study, originating from multiple *Corynebacteria* species and *S. aureus*. These predictions were complemented by RNA-seq analyses to evaluate the transcriptional landscape of candidate uptake systems under iron-limited environments.

By prioritizing structural over primary sequence conservation, **PRESERVE** facilitated a more sensitive identification of potential functional homologs implicated in siderophore uptake. From the total 4,520 lipoproteins first identified using primary sequence conservation, the approach narrowed to a candidate set of 444 lipoproteins with a structural similarity to known *S. aureus* siderophore-binding proteins, 63 of which occur in *S. aureus* and 381 were found in *Corynebacteria* strains.

However, validation using *S. aureus* samples provided more insight into the results. First, the strains used for the validation already showed a diversity for the present lipoproteins, as only 38 sequence homologs were identified by sequence homology analysis. Although this might suggest assembly or annotation problems, it could also indicate diversity in the use of lipoproteins in *S. aureus* strains, as it has been studied for other staphylococcal samples (Witte Paz *et al.* 2025).

Moreover, the validation of the structural homologs shows that the approach is overly sensitive. Across these samples, the pipeline detected more homologs than the number of reference proteins used (63 homologs versus the expected 38). First, it was shown that the *htsA* gene was clustered together with an uncharacterized iron-binding protein, hence increasing the number of false positive hits. This suggests that the thresholds defined for the *MMseqs2* clustering should increase, which in contrary would also require a higher computational time for the protein structure prediction.

Moreover, other proteins that were not expected to be called homologs were identified via a **BLAST** search. IsdE and BtuF are known to participate in iron uptake (Grigg, Vermeiren, *et al.* 2007; Krewulak and Vogel 2008), but are not associated with the siderophore binding mechanism. In general, this suggests that proteins involved in iron homeostasis may have structures too similar and require an increase of the *LDDT* thresholds defined in this study. Nevertheless, the *LDDT* values of the corynebacterial hits are also generally lower and therefore indicate that too strict threshold might increase the chances of missing too many proteins for the cross-genera comparison. A detailed analysis on the optimal thresholds for such a comparison remains pending.

In addition, the characterization of corynebacterial strains revealed four proteins that can represent false negatives. For example, the gene *RS06480* from *C. accolens* was not identified as a structural homolog, though it only showed a high divergence in non-globular,

unstructured regions, while its core structure remained conserved, suggesting a putative role in iron homeostasis. Such unstructured regions are crucial for lipoprotein binding to the outer membrane (El Rayes *et al.* 2021), yet they are poorly predicted by structure-prediction tools like ColabFold (Bertoline *et al.* 2023). To reduce this limitation, PRESERVE can be extended with structure processing tools, such as AlphaCutter (Tam and Iwasaki 2023), which remove unreliable predicted regions and could therefore improve the accuracy of the structural comparison.

Despite the identified putative false positives and negatives, the integration of RNA-seq data and the orthology analysis supports the relevance of the identified homologs. RNA-seq results showed that most of the identified homologous lipoproteins are strongly up-regulated under iron-deprived conditions, highlighting their potential role in iron acquisition. Moreover, the orthogroup analysis revealed high conservation of these structural homologs within the corynebacterial species. However, our results suggest that siderophore uptake specificity in *Corynebacteria* might not be determined by structural homology alone, as computational predictions did not align with experimentally observed uptake phenotypes. Some further approaches not described in this study tried to provide further analyses to resolve this linkage by attempting to replicate the results presented by Zhao *et al.* (2024) on a computational level. One approach included the training of Hidden Markov Models based on staphylococcal uptake systems to identify corynebacterial lipoprotein classes, following a similar idea presented by Gu *et al.* (2025). Also, molecular docking experiments were conducted to assess the binding affinity of a few selected structural homologs to the metabolites staphyloferrin A and B. Yet, none of these approaches reproduced the experimentally observed phenotypes and were thus not applied to the full protein set. The inability to link specific orthogroups to a single siderophore might rely on differences between the uptake methods of *Corynebacteria* and *S. aureus*. For example, in *S. aureus*, multiple siderophores of a specific class are imported via single lipoproteins, as seen with catecholates and hydroxamates siderophores and the uptake lipoproteins *SstD* and *FhuD1/2* (Beasley *et al.* 2011; Sebelsky and Heinrichs 2001). A comparable mechanism could be expected in *C. hesseae*, for example, with the genes *R3O64_03755* and *R3O64_11615* being both identified as structural homologs via PRESERVE and members of the same orthogroup via OrthoFinder. However, experimental data from Zhao *et al.* (2024) showed a functional difference. While *R3O64_11615* is responsible for the uptake of SF-A and SF-B, *R3O64_03755* does not. These results highlight that even small differences can have significant phenotypic consequences, and that structural or sequence homology alone may be insufficient to predict substrate specificity.

Despite the challenges in providing a detailed mapping of siderophore uptake to individual genes, our analysis of promoter regions and regulatory motifs provides insight into the regulation of the uptake mechanisms in *Staphylococcus aureus* and *Corynebacteria* species. In *S. aureus*, the identified upstream motifs correspond to binding sites of the ferric uptake regulator (Fur), consistent with its well-described role in controlling iron homeostasis and siderophore-related operons in the literature (Friedman *et al.* 2006). Hence, this finding validates our methodological approach by identifying the expected regulation of the reference lipoproteins used in this study. For corynebacterial species, our approach suggests a different regulatory landscape. Here, binding sites of the iron-dependent regulator *IdeR* were enriched upstream of the identified homologous genes, with the motif identified as the most significant of the dataset (lowest *E*-value). *IdeR* has previously been described as a major regulator of iron metabolism in *Actinobacteria*, including *Corynebacterium diphtheriae* (Marcos-Torres *et al.* 2021), which strengthens our interpretation that *Fur* might play a smaller role for the regulation of these genes

in these species. This difference in the regulation systems highlights the idea that even if homologous functions occur, they may be embedded in different functional contexts. Moreover, the detection of *IdeR* motifs across multiple corynebacterial species suggests that this signal can be used to uncover additional iron-related genes not captured by homology searches (Reitz and Medema [2022]). An approach to refine species-specific motifs would be to remove any identified false hits, such as proteins that are not up-regulated in the transcriptomic analysis, before computing the motifs. Nevertheless, as these proteins still play a role in iron-acquisition, the effect might not be drastic. With such refined motifs, motif-searching tools such as **FIMO** (Bailey *et al.* [2009]) could then be applied to identify genome-wide instances of *IdeR* binding sites and the downstream genes they might regulate, indicating further iron-related genes that have been so far overlooked.

Nevertheless, it is also important to mention the limitations of our promoter analysis approach. The identification of enriched regulatory motifs depends strongly on the chosen parameters, such as the maximum intergenic distance for operon definition and the length of upstream regions used to approximate untranslated and promoter regions. Although the 50nt chosen for operon computation and 400nt for upstream sequence extraction are based on parameters used in previous studies (Yada *et al.* [1999]), these thresholds may not always accurately capture species-specific variation in operon structure or untranslated region lengths. Hence, some identified motifs may not correspond to true binding sites, while others might have been missed. Nevertheless, the identification of the *Fur* motif in *S. aureus* and the evidence of *IdeR* being found in *C. diphtheria* show that, despite parameter sensitivity, the approach is robust enough to recover biologically meaningful signals.

Overall, this study provides a high-throughput characterization of siderophore uptake mechanisms in multiple corynebacterial samples through the integration of structure-based homology, transcriptional data, and regulatory motif analysis. By providing the **PRESERVE** pipeline, we have established a computational basis for future cross-genera comparisons. While a detailed functional description of individual siderophore uptake mechanisms remains to be solved, the results identified several putative candidates for further experimental validation. Moreover, we were able to provide an alternative for the regulation of siderophore-uptake mechanisms in corynebacterial strains, which could be further used to identify further iron-related mechanisms in these species. Achieving the understanding of these uptake mechanisms could then provide insight into their relevance in the microbial interactions within a microbiome environment, as well as further mechanisms that allow the control via engineering of microbial communities.

References

- Adolf, L. A. and Heilbronner, S. (2022). “Nutritional Interactions between Bacterial Species Colonising the Human Nasal Cavity: Current Knowledge and Future Prospects”. In: *Metabolites* vol. 12, no. 6, p. 489. ISSN: 2218-1989. DOI: [10.3390/metabo12060489](https://doi.org/10.3390/metabo12060489).
- Andrews, S. (2010). *A Quality Control Tool for High Throughput Sequence Data*. URL: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). “MEME Suite: tools for motif discovery and searching”. In: *Nucleic Acids Research* vol. 37, no. suppl_2, W202–W208. ISSN: 0305-1048. DOI: [10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335).
- Beasley, F. C., Marolda, C. L., Cheung, J., Buac, S., and Heinrichs, D. E. (2011). “Staphylococcus aureus Transporters Hts, Sir, and Sst Capture Iron Liberated from Human Transferrin by staphyloferrin A, staphyloferrin B, and Catecholamine Stress

- Hormones, Respectively, and Contribute to Virulence”. In: *Infection and Immunity* vol. 79, no. 6, pp. 2345–2355. ISSN: 1098-5522. DOI: [10.1128/IAI.00117-11](https://doi.org/10.1128/IAI.00117-11).
- Bertoline, L. M. F., Lima, A. N., Krieger, J. E., and Teixeira, S. K. (2023). “Before and after AlphaFold2: An Overview of Protein Structure Prediction”. In: *Frontiers in Bioinformatics* vol. 3. ISSN: 2673-7647. DOI: [10.3389/fbinf.2023.1120370](https://doi.org/10.3389/fbinf.2023.1120370).
- Bittrich, S., Segura, J., Duarte, J. M., Burley, S. K., and Rose, Y. (2024). “RCSB Protein Data Bank: Exploring Protein 3D Similarities via Comprehensive Structural Alignments”. In: *Bioinformatics* vol. 40, no. 6, btae370. ISSN: 1367-4811. DOI: [10.1093/bioinformatics/btae370](https://doi.org/10.1093/bioinformatics/btae370).
- Blin, K., Shaw, S., Augustijn, H. E., Reitz, Z. L., Biermann, F., Alanjary, M., Fetter, A., Terlouw, B. R., Metcalf, W. W., Helfrich, E. J. N., van Wezel, G. P., Medema, M. H., and Weber, T. (2023). “antiSMASH 7.0: New and Improved Predictions for Detection, Regulation, Chemical Structures and Visualisation”. In: *Nucleic Acids Research* vol. 51, no. W1, W46–W50. ISSN: 0305-1048. DOI: [10.1093/nar/gkad344](https://doi.org/10.1093/nar/gkad344).
- Bomar, L., Brugger, S. D., Yost, B. H., Davies, S. S., and Lemon, K. P. (2016). “Corynebacterium accolens Releases Antipneumococcal Free Fatty Acids from Human Nostril and Skin Surface Triacylglycerols”. In: *mBio* vol. 7, no. 1, 10.1128/mbio.01725–15. DOI: [10.1128/mbio.01725-15](https://doi.org/10.1128/mbio.01725-15).
- Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). “Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive”. In: *Protein Crystallography*, pp. 627–641.
- Camus, L., Franz, J., Gerlach, D., Lange, A., Power, J. J., Navarro-Díaz, M., Rapp, J., Otto, M., Klages, L. J., Starostin, V., Mößner, J., Bilici, K., Ham, S., Kalinowski, J., Angenent, L. T., Link, H., and Heilbronner, S. (2025). “Tyrosine Availability Shapes Staphylococcus aureus Nasal Colonization and Interactions with Commensal Communities”. In: *bioRxiv*. DOI: [10.1101/2025.05.06.651429](https://doi.org/10.1101/2025.05.06.651429).
- Caza, M. and Kronstad, J. (2013). “Shared and Distinct Mechanisms of Iron Acquisition by Bacterial and Fungal Pathogens of Humans”. In: *Frontiers in Cellular and Infection Microbiology* vol. 3. ISSN: 2235-2988. DOI: [10.3389/fcimb.2013.00080](https://doi.org/10.3389/fcimb.2013.00080).
- Cipriano, M. J., Novichkov, P. N., Kazakov, A. E., Rodionov, D. A., Arkin, A. P., Gelfand, M. S., and Dubchak, I. (2013). “RegTransBase—a Database of Regulatory Sequences and Interactions Based on Literature: A Resource for Investigating Transcriptional Regulation in Prokaryotes”. In: *BMC Genomics* vol. 14, pp. 1–8.
- Conroy, B. S., Grigg, J. C., Kolesnikov, M., Morales, L. D., and Murphy, M. E. P. (2019). “Staphylococcus aureus Heme and Siderophore-Iron Acquisition Pathways”. In: *Biometals* vol. 32, no. 3, pp. 409–424. ISSN: 1572-8773. DOI: [10.1007/s10534-019-00188-2](https://doi.org/10.1007/s10534-019-00188-2).
- Cordero, O. X., Ventouras, L.-A., DeLong, E. F., and Polz, M. F. (2012). “Public Good Dynamics Drive Evolution of Iron Acquisition Strategies in Natural Bacterioplankton Populations”. In: *Proceedings of the National Academy of Sciences* vol. 109, no. 49, pp. 20059–20064. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.1213344109](https://doi.org/10.1073/pnas.1213344109).
- Cunningham, A. L., Zhbannikov, I. Y., Myers, R., Tran, T. H., Gao, W., Lemon, K. P., Aquino, J. N., Hurst, J. H., Yoon, J. W., Seed, P. C., and Kelly, M. S. (2025). “Genome Mining Identifies a Diversity of Natural Product Biosynthetic Capacity in Human Respiratory Corynebacterium Strains”. In: *mSphere* vol. 10, no. 6, e00258–25. DOI: [10.1128/msphere.00258-25](https://doi.org/10.1128/msphere.00258-25).
- de Jong, A., Kuipers, O. P., and Kok, J. (2022). “FUNAGE-pro: Comprehensive Web Server for Gene Set Enrichment Analysis of Prokaryotes”. In: *Nucleic Acids Research* vol. 50, no. W1, W330–W336. ISSN: 0305-1048. DOI: [10.1093/nar/gkac441](https://doi.org/10.1093/nar/gkac441).

- Dudek, C.-A. and Jahn, D. (2022). “PRODORIC: State-of-the-Art Database of Prokaryotic Gene Regulation”. In: *Nucleic Acids Research* vol. 50, no. D1, pp. D295–D302.
- El Rayes, J., Szewczyk, J., Deghelt, M., Csoma, N., Matagne, A., Iorga, B. I., Cho, S.-H., and Collet, J.-F. (2021). “Disorder Is a Critical Component of Lipoprotein Sorting in Gram-Negative Bacteria”. In: *Nature Chemical Biology* vol. 17, no. 10, pp. 1093–1100. ISSN: 1552-4469. DOI: [10.1038/s41589-021-00845-z](https://doi.org/10.1038/s41589-021-00845-z).
- Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). “MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report”. In: *Bioinformatics* vol. 32, no. 19, pp. 3047–3048. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btw354](https://doi.org/10.1093/bioinformatics/btw354).
- Friedman, D. B., Stauff, D. L., Pishchany, G., Whitwell, C. W., Torres, V. J., and Skaar, E. P. (2006). “Staphylococcus aureus Redirects Central Metabolism to Increase Iron Availability”. In: *PLOS Pathogens* vol. 2, no. 8, e87. ISSN: 1553-7374. DOI: [10.1371/journal.ppat.0020087](https://doi.org/10.1371/journal.ppat.0020087).
- Fuchs, S., Mehlan, H., Bernhardt, J., Hennig, A., Michalik, S., Surmann, K., Pané-Farré, J., Giese, A., Weiss, S., Backert, L., Herbig, A., Nieselt, K., Hecker, M., Völker, U., and Mäder, U. (2018). “AureoWiki - the Repository of the Staphylococcus aureus Research and Annotation Community”. In: *International Journal of Medical Microbiology* vol. 308, no. 6, pp. 558–568. ISSN: 1438-4221. DOI: [10.1016/j.ijmm.2017.11.011](https://doi.org/10.1016/j.ijmm.2017.11.011).
- García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T. F., and Conesa, A. (2012). “Qualimap: Evaluating next-Generation Sequencing Alignment Data”. In: *Bioinformatics* vol. 28, no. 20, pp. 2678–2679. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/bts503](https://doi.org/10.1093/bioinformatics/bts503).
- Grant, C. E. and Bailey, T. L. (2021). “XSTREME: Comprehensive Motif Analysis of Biological Sequence Datasets”. In: *bioRxiv*, p. 2021.09.02.458722. DOI: [10.1101/2021.09.02.458722](https://doi.org/10.1101/2021.09.02.458722).
- Grigg, J. C., Cheung, J., Heinrichs, D. E., and Murphy, M. E. (2010). “Specificity of Staphyloferrin B Recognition by the SirA Receptor from Staphylococcus aureus”. In: *Journal of Biological Chemistry* vol. 285, no. 45, pp. 34579–34588. ISSN: 0021-9258. DOI: [10.1074/jbc.m110.172924](https://doi.org/10.1074/jbc.m110.172924).
- Grigg, J. C., Cooper, J. D., Cheung, J., Heinrichs, D. E., and Murphy, M. E. (2010). “The Staphylococcus aureus Siderophore Receptor HtsA Undergoes Localized Conformational Changes to Enclose staphyloferrin A in an Arginine-Rich Binding Pocket”. In: *Journal of Biological Chemistry* vol. 285, no. 15, pp. 11162–11171. ISSN: 0021-9258. DOI: [10.1074/jbc.m109.097865](https://doi.org/10.1074/jbc.m109.097865).
- Grigg, J. C., Vermeiren, C. L., Heinrichs, D. E., and Murphy, M. E. P. (2007). “Heme Coordination by Staphylococcus aureus IsdE”. In: *Journal of Biological Chemistry* vol. 282, no. 39, pp. 28815–28822. ISSN: 0021-9258, 1083-351X. DOI: [10.1074/jbc.M704602200](https://doi.org/10.1074/jbc.M704602200).
- Gu, S., Shao, Z., Qu, Z., Zhu, S., Shao, Y., Zhang, D., Allen, R., He, R., Shao, J., Xiong, G., Jousset, A., Friman, V.-P., Wei, Z., Kümmerli, R., and Li, Z. (2025). “Siderophore Synthetase-Receptor Gene Coevolution Reveals Habitat- and Pathogen-Specific Bacterial Iron Interaction Networks”. In: *Science Advances* vol. 11, no. 3. ISSN: 2375-2548. DOI: [10.1126/sciadv.adq5038](https://doi.org/10.1126/sciadv.adq5038).
- Harbig, T. A. (2023). *Integrative-Transcriptomics/rnaSeq-pipeline: V1.0.0*. Zenodo. DOI: [10.5281/ZENODO.7500898](https://doi.org/10.5281/ZENODO.7500898).
- Harbig, T. A., Witte Paz, M., and Nieselt, K. (2023). “GO-Compass: Visual Navigation of Multiple Lists of GO Terms”. In: *Computer Graphics Forum* vol. 42, no. 3, pp. 271–281. DOI: [10.1111/cgf.14829](https://doi.org/10.1111/cgf.14829).

- Hauser, M., Steinegger, M., and Söding, J. (2016). “MMseqs Software Suite for Fast and Deep Clustering and Searching of Large Protein Sequence Sets”. In: *Bioinformatics* vol. 32, no. 9, pp. 1323–1330. ISSN: 1367-4811, 1367-4803. DOI: [10.1093/bioinformatics/btw006](https://doi.org/10.1093/bioinformatics/btw006).
- Kelly, M. S., Plunkett, C., Yu, Y., Aquino, J. N., Patel, S. M., Hurst, J. H., Young, R. R., Smieja, M., Steenhoff, A. P., Arscott-Mills, T., Feemster, K. A., Boiditswe, S., Leburu, T., Mazhani, T., Patel, M. Z., Rawls, J. F., Jawahar, J., Shah, S. S., Polage, C. R., Cunningham, C. K., and Seed, P. C. (2022). “Non-Diphtheriae *Corynebacterium* Species Are Associated with Decreased Risk of Pneumococcal Colonization during Infancy”. In: *The ISME Journal* vol. 16, no. 3, pp. 655–665. ISSN: 1751-7370. DOI: [10.1038/s41396-021-01108-4](https://doi.org/10.1038/s41396-021-01108-4).
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). “HISAT: A Fast Spliced Aligner with Low Memory Requirements”. In: *Nature Methods* vol. 12, no. 4, pp. 357–360. ISSN: 1548-7105. DOI: [10.1038/nmeth.3317](https://doi.org/10.1038/nmeth.3317).
- Koehl, P. (2001). “Protein Structure Similarities”. In: *Current Opinion in Structural Biology* vol. 11, no. 3, pp. 348–353. ISSN: 0959-440X. DOI: [10.1016/S0959-440X\(00\)00214-1](https://doi.org/10.1016/S0959-440X(00)00214-1).
- Kramer, J., Özkaya, Ö., and Kümmerli, R. (2020). “Bacterial Siderophores in Community and Host Interactions”. In: *Nature Reviews Microbiology* vol. 18, no. 3, pp. 152–163. ISSN: 1740-1526. DOI: [10.1038/s41579-019-0284-4](https://doi.org/10.1038/s41579-019-0284-4).
- Krewulak, K. D. and Vogel, H. J. (2008). “Structural Biology of Bacterial Iron Uptake”. In: *Biochimica et Biophysica Acta (BBA) - Biomembranes. Structural Proteomics of the Cell Envelope of Gram-negative Bacteria* vol. 1778, no. 9, pp. 1781–1804. ISSN: 0005-2736. DOI: [10.1016/j.bbamem.2007.07.026](https://doi.org/10.1016/j.bbamem.2007.07.026).
- Lee, J. H. and Holmes, R. K. (2000). “Characterization of Specific Nucleotide Substitutions in DtxR-specific Operators of *Corynebacterium diphtheriae* That Dramatically Affect DtxR Binding, Operator Function, and Promoter Strength”. In: *Journal of Bacteriology* vol. 182, no. 2, pp. 432–438. ISSN: 0021-9193. DOI: [10.1128/jb.182.2.432-438.2000](https://doi.org/10.1128/jb.182.2.432-438.2000).
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). “The Sequence Alignment/Map Format and Samtools”. In: *Bioinformatics* vol. 25, no. 16, pp. 2078–2079. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352).
- Liao, Y., Smyth, G. K., and Shi, W. (2014). “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features”. In: *Bioinformatics* vol. 30, no. 7, pp. 923–930. ISSN: 1367-4803. DOI: [10.1093/bioinformatics/btt656](https://doi.org/10.1093/bioinformatics/btt656).
- Liu, C. M., Price, L. B., Hungate, B. A., Abraham, A. G., Larsen, L. A., Christensen, K., Stegger, M., Skov, R., and Andersen, P. S. (2015). “*Staphylococcus aureus* and the Ecology of the Nasal Microbiome”. In: *Science Advances* vol. 1, no. 5, e1400216. DOI: [10.1126/sciadv.1400216](https://doi.org/10.1126/sciadv.1400216).
- Love, M. I., Huber, W., and Anders, S. (2014). “Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2”. In: *Genome Biology* vol. 15, no. 12, p. 550. ISSN: 1474-760X. DOI: [10.1186/s13059-014-0550-8](https://doi.org/10.1186/s13059-014-0550-8).
- Marcos-Torres, F. J., Maurer, D., Juniar, L., and Griesse, J. J. (2021). “The Bacterial Iron Sensor IdeR Recognizes Its DNA Targets by Indirect Readout”. In: *Nucleic Acids Research* vol. 49, no. 17, pp. 10120–10135. ISSN: 0305-1048. DOI: [10.1093/nar/gkab711](https://doi.org/10.1093/nar/gkab711).
- Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). “ColabFold: Making Protein Folding Accessible to All”. In: *Nature Methods* vol. 19, no. 6, pp. 679–682. ISSN: 1548-7091, 1548-7105. DOI: [10.1038/s41592-022-01488-1](https://doi.org/10.1038/s41592-022-01488-1).
- Morrissey, J. A., Cockayne, A., Hill, P. J., and Williams, P. (2000). “Molecular Cloning and Analysis of a Putative Siderophore ABC Transporter from *Staphylococcus aureus*”.

- In: *Infection and Immunity* vol. 68, no. 11, pp. 6281–6288. DOI: [10.1128/iai.68.11.6281-6288.2000](https://doi.org/10.1128/iai.68.11.6281-6288.2000).
- Osbourne, A. E. and Field, B. (2009). “Operons”. In: *Cellular and Molecular Life Sciences* vol. 66, no. 23, pp. 3755–3775. ISSN: 1420-9071. DOI: [10.1007/s00018-009-0114-3](https://doi.org/10.1007/s00018-009-0114-3).
- Raymond, K. N., Dertz, E. A., and Kim, S. S. (2003). “Enterobactin: An Archetype for Microbial Iron Transport”. In: *Proceedings of the National Academy of Sciences* vol. 100, no. 7, pp. 3584–3588. ISSN: 0027-8424, 1091-6490. DOI: [10.1073/pnas.0630018100](https://doi.org/10.1073/pnas.0630018100).
- Reitz, Z. L. and Medema, M. H. (2022). “Genome Mining Strategies for Metallophore Discovery”. In: *Current Opinion in Biotechnology* vol. 77, p. 102757. ISSN: 0958-1669. DOI: [10.1016/j.copbio.2022.102757](https://doi.org/10.1016/j.copbio.2022.102757).
- Saha, R., Saha, N., Donofrio, R. S., and Bestervelt, L. L. (2013). “Microbial Siderophores: A Mini Review”. In: *Journal of Basic Microbiology* vol. 53, no. 4, pp. 303–317. ISSN: 0233-111X, 1521-4028. DOI: [10.1002/jobm.201100552](https://doi.org/10.1002/jobm.201100552).
- Schalk, I. J. (2025). “Bacterial Siderophores: Diversity, Uptake Pathways and Applications”. In: *Nature Reviews Microbiology* vol. 23, no. 1, pp. 24–40. ISSN: 1740-1534. DOI: [10.1038/s41579-024-01090-6](https://doi.org/10.1038/s41579-024-01090-6).
- Sebulsky, M. T. and Heinrichs, D. E. (2001). “Identification and Characterization of *fluD1* and *fluD2*, Two Genes Involved in Iron-Hydroxamate Uptake in *Staphylococcus aureus*”. In: *Journal of Bacteriology* vol. 183, no. 17, pp. 4994–5000. ISSN: 0021-9193. DOI: [10.1128/JB.183.17.4994-5000.2001](https://doi.org/10.1128/JB.183.17.4994-5000.2001).
- Shatsky, M., Nussinov, R., and Wolfson, H. J. (2008). “Algorithms for Multiple Protein Structure Alignment and Structure-Derived Multiple Sequence Alignment”. In: *Methods in Molecular Biology* vol. 413, pp. 125–146. ISSN: 1064-3745. DOI: [10.1007/978-1-59745-574-9_5](https://doi.org/10.1007/978-1-59745-574-9_5).
- Sheldon, J. and Heinrichs, D. (2012). “The Iron-Regulated Staphylococcal Lipoproteins”. In: *Frontiers in Cellular and Infection Microbiology* vol. 2. ISSN: 2235-2988. DOI: [10.3389/fcimb.2012.00041](https://doi.org/10.3389/fcimb.2012.00041).
- Tam, C. and Iwasaki, W. (2023). “AlphaCutter: Efficient Removal of Non-Globular Regions from Predicted Protein Structures”. In: *Proteomics* vol. 23, no. 16, p. 2300176. ISSN: 1615-9861. DOI: [10.1002/pmic.202300176](https://doi.org/10.1002/pmic.202300176).
- Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gíslason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. (2022). “SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models”. In: *Nature Biotechnology* vol. 40, no. 7, pp. 1023–1025. ISSN: 1087-0156. DOI: [10.1038/s41587-021-01156-3](https://doi.org/10.1038/s41587-021-01156-3).
- van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. (2024). “Fast and Accurate Protein Structure Search with Foldseek”. In: *Nature Biotechnology* vol. 42, no. 2, pp. 243–246. ISSN: 1546-1696. DOI: [10.1038/s41587-023-01773-0](https://doi.org/10.1038/s41587-023-01773-0).
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., *et al.* (2022). “AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models”. In: *Nucleic Acids Research* vol. 50, no. D1, pp. D439–D444. DOI: [10.1093/nar/gkab1061](https://doi.org/10.1093/nar/gkab1061).
- Witte Paz, M.**, Bitzer, A., Nieselt, K., and Heilbronner, S. (2025). “A Landscape of Metallophore Synthesis and Uptake Potential of the Genus *Staphylococcus*”. In: *NAR Genomics and Bioinformatics* vol. 7, no. 4, lqaf183. DOI: [10.1093/nargab/lqaf183](https://doi.org/10.1093/nargab/lqaf183).

- Wooldridge, K. G. and Williams, P. H. (1993). “Iron Uptake Mechanisms of Pathogenic Bacteria”. In: *FEMS Microbiology Reviews* vol. 12, no. 4, pp. 325–348. ISSN: 0168-6445, 1574-6976. DOI: [10.1111/j.1574-6976.1993.tb00026.x](https://doi.org/10.1111/j.1574-6976.1993.tb00026.x).
- Xiong, A., Singh, V. K., Cabrera, G., and Jayaswal, R. K. (2000). “Molecular Characterization of the Ferric-Uptake Regulator, Fur, from *Staphylococcus aureus*”. In: *Microbiology* vol. 146, no. 3, pp. 659–668. ISSN: 1350-0872, 1465-2080. DOI: [10.1099/00221287-146-3-659](https://doi.org/10.1099/00221287-146-3-659).
- Yada, T., Nakao, M., Totoki, Y., and Nakai, K. (1999). “Modeling and Predicting Transcriptional Units of *Escherichia Coligenes* Using Hidden Markov Models”. In: *Bioinformatics* vol. 15, no. 12, pp. 987–993. ISSN: 1367-4811, 1367-4803. DOI: [10.1093/bioinformatics/15.12.987](https://doi.org/10.1093/bioinformatics/15.12.987).
- Zaru, R., Orchard, S., and Consortium, T. U. (2023). “UniProt Tools: BLAST, Align, Peptide Search, and ID Mapping”. In: *Current Protocols* vol. 3, no. 3, e697. ISSN: 2691-1299. DOI: [10.1002/cpz1.697](https://doi.org/10.1002/cpz1.697).
- Zhao, Y., Bitzer, A., Power, J. J., Belikova, D., Torres Salazar, B. O., Adolf, L. A., Gerlach, D., Krismer, B., and Heilbronner, S. (2024). “Nasal Commensals Reduce *Staphylococcus aureus* Proliferation by Restricting Siderophore Availability”. In: *The ISME Journal* vol. 18, no. 1, wrae123. DOI: [10.1093/ismejo/wrae123](https://doi.org/10.1093/ismejo/wrae123).

Supplementary Material

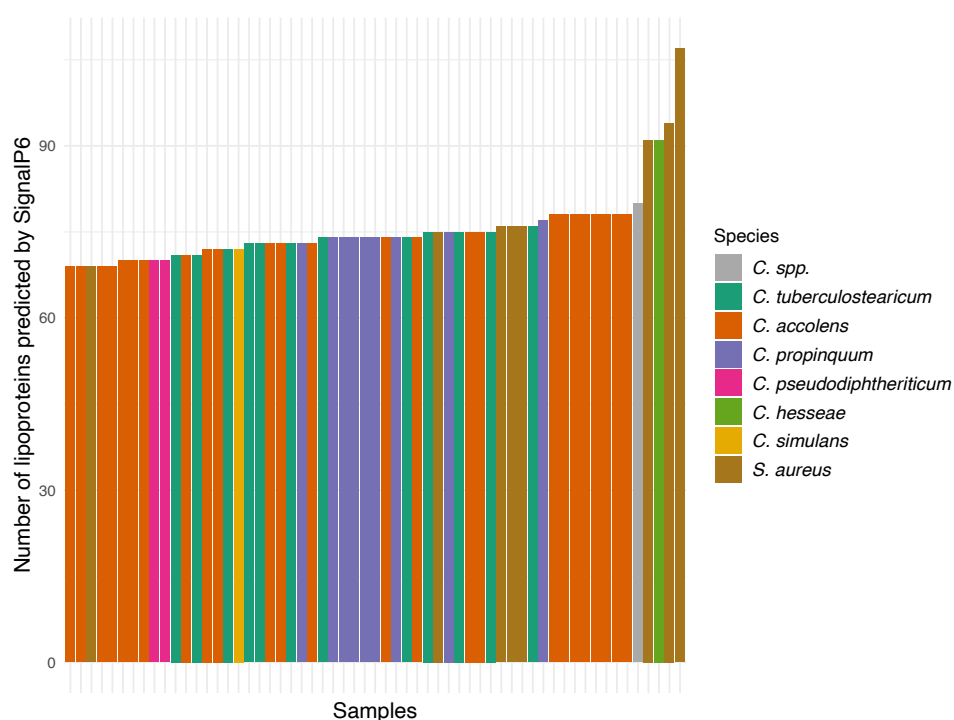


Figure V.S1: Number of lipoproteins (TAT/TATLIPO) per sample as predicted by **SignalP6** (Teufel *et al.* [2022](https://doi.org/10.1093/ismejo/wrae123)). Each color represents the species the sample has been labeled with. Only one sample (gray color) was not classified on the species level.

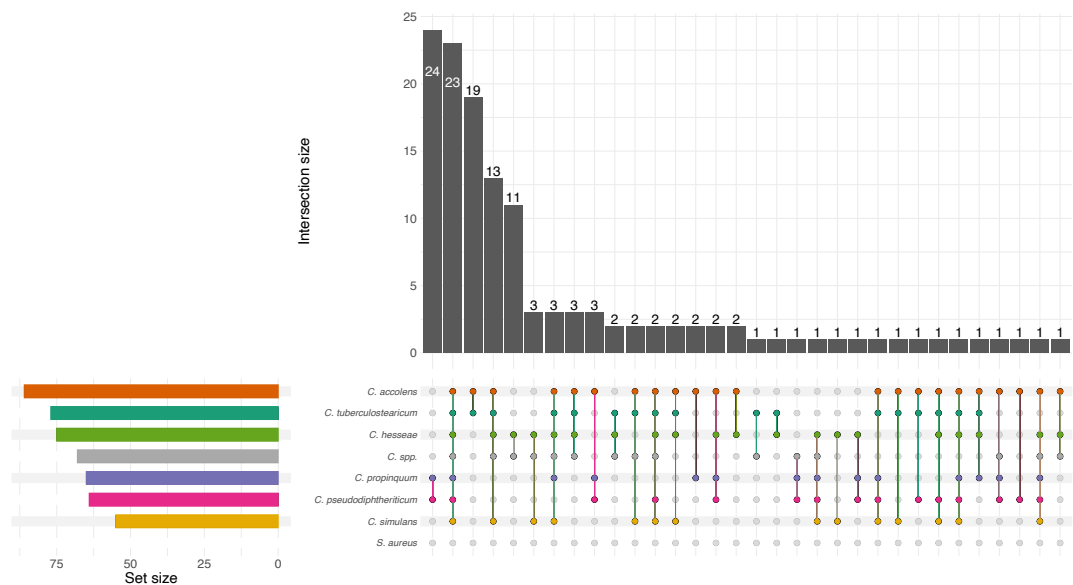


Figure V.S2: Upset plot for shared lipoproteins as identified with the PRESERVE pipeline.

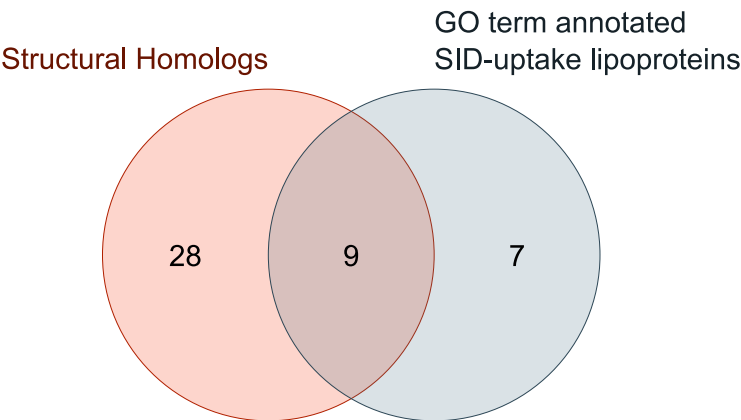


Figure V.S3: Venn diagram comparing structural homologs identified by PRESERVE and lipoproteins associated with siderophore uptake based on GO terms.

Table V.S1: Overview of corynebacterial and staphylococcal genome assemblies used for the identification of structural homologs as well as validation of the computational approaches. The table provides an overview of the sample ID of the strain, its classified species, the origin from the sample as well as the status of accessibility of the genomic sequence. Unpublished genomes were provided via internal communication.

Sample Name	Species	Origin of sample	BioSample ID
sample_129	<i>C. accolens</i>	LaCa Collection	unpublished
sample_142	<i>C. accolens</i>	LaCa Collection	unpublished
sample_143	<i>C. accolens</i>	LaCa Collection	unpublished
sample_144	<i>C. accolens</i>	LaCa Collection	unpublished
sample_156	<i>C. accolens</i>	LaCa Collection	SAMN47812989
sample_173	<i>C. accolens</i>	LaCa Collection	unpublished
sample_174	<i>C. accolens</i>	LaCa Collection	unpublished
sample_175	<i>C. accolens</i>	LaCa Collection	unpublished
sample_191	<i>C. accolens</i>	LaCa Collection	unpublished
sample_193	<i>C. accolens</i>	LaCa Collection	SAMN47813004
sample_220	<i>C. accolens</i>	LaCa Collection	SAMN47813025
sample_221	<i>C. accolens</i>	LaCa Collection	SAMN47813026
sample_24	<i>C. accolens</i>	LaCa Collection	SAMN47812908
sample_248	<i>C. accolens</i>	LaCa Collection	SAMN47813049
sample_249	<i>C. accolens</i>	LaCa Collection	SAMN47813050
sample_25	<i>C. accolens</i>	LaCa Collection	SAMN47812909
sample_267	<i>C. accolens</i>	LaCa Collection	SAMN47813065
sample_268	<i>C. accolens</i>	LaCa Collection	SAMN47813066
sample_282	<i>C. accolens</i>	LaCa Collection	SAMN47813076
sample_42	<i>C. accolens</i>	LaCa Collection	SAMN47812921
sample_43	<i>C. accolens</i>	LaCa Collection	SAMN47812922
sample_77	<i>C. accolens</i>	LaCa Collection	unpublished
sample_78	<i>C. accolens</i>	LaCa Collection	SAMN45893581
sample_97	<i>C. accolens</i>	LaCa Collection	SAMN45893582
sample_99	<i>C. accolens</i>	LaCa Collection	unpublished
sample_CA	<i>C. accolens</i>	Commercial strain	GCF_000478135.1
sample_CH	<i>C. hesseae</i>	Münster collection	SAMN37848894
sample_15	<i>C. propinquum</i>	LaCa Collection	SAMN47812899
sample_152	<i>C. propinquum</i>	LaCa Collection	SAMN47812985
sample_17	<i>C. propinquum</i>	LaCa Collection	SAMN47812901
sample_170	<i>C. propinquum</i>	LaCa Collection	unpublished
sample_171	<i>C. propinquum</i>	LaCa Collection	unpublished
sample_172	<i>C. propinquum</i>	LaCa Collection	unpublished
sample_264	<i>C. propinquum</i>	LaCa Collection	SAMN47813062
sample_265	<i>C. propinquum</i>	LaCa Collection	SAMN47813063
sample_69	<i>C. propinquum</i>	LaCa Collection	SAMN47812943
sample_243	<i>C. pseudodiphtheriticum</i>	LaCa Collection	SAMN47813045
sample_244	<i>C. pseudodiphtheriticum</i>	LaCa Collection	SAMN47813046
sample_CS	<i>C. simulans</i>	Münster collection	SAMN37848892
sample_101	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN45893583
sample_102	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN47812967
sample_117	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN45893585
sample_204	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN47813011
sample_205	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN47813012
sample_222	<i>C. tuberculostearicum</i>	LaCa Collection	unpublished
sample_224	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN47813028
sample_54	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN45893580
sample_62	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN47812938
sample_7	<i>C. tuberculostearicum</i>	LaCa Collection	SAMN45893579
sample_8	<i>C. tuberculostearicum</i>	LaCa Collection	unpublished
sample_9	<i>C. tuberculostearicum</i>	LaCa Collection	unpublished
sample_116	<i>Corynebacterium sp.</i>	LaCa Collection	SAMN45893584
sample_1	<i>S. aureus</i>	LaCa Collection	unpublished
sample_145	<i>S. aureus</i>	LaCa Collection	unpublished
sample_186	<i>S. aureus</i>	LaCa Collection	SAMN47812998
sample_19	<i>S. aureus</i>	LaCa Collection	SAMN47812903
sample_199	<i>S. aureus</i>	LaCa Collection	unpublished
sample_2	<i>S. aureus</i>	LaCa Collection	unpublished
sample_3	<i>S. aureus</i>	LaCa Collection	unpublished
sample_36	<i>S. aureus</i>	LaCa Collection	unpublished

V. Characterizing Siderophore Uptake and their Regulation Mechanisms in *Corynebacteria* based on Structural Homology and Transcriptome Profiling

Table V.S2: Result of MMseqs2 search of reference lipoproteins within the *S. aureus* samples used for computational validation purposes. Only hits with a reference coverage of over 0.99 and a sequence similarity of 0.9 were allowed. For each protein sequence, its best hit (maximum coverage and similarity) was chosen.

<i>S. aureus</i> sample ID	Target protein	Ref. lipoprotein locus	Seq. identity	Ref. coverage
sample_1	OEHLIH_00205	<i>fhuD1</i>	1.000	1.000
sample_1	OEHLIH_12305	<i>fhuD2</i>	0.996	1.000
sample_1	OEHLIH_12815	<i>htsA</i>	1.000	1.000
sample_1	OEHLIH_09825	<i>sirA</i>	1.000	1.000
sample_1	OEHLIH_06605	<i>sstD</i>	0.997	1.000
sample_2	GOALGM_13575	<i>fhuD1</i>	1.000	1.000
sample_2	GOALGM_01135	<i>fhuD2</i>	0.996	1.000
sample_2	GOALGM_00625	<i>htsA</i>	1.000	1.000
sample_2	GOALGM_03610	<i>sirA</i>	1.000	1.000
sample_2	GOALGM_06825	<i>sstD</i>	0.997	1.000
sample_3	OKMLLP_12975	<i>fhuD1</i>	1.000	1.000
sample_3	OKMLLP_11655	<i>fhuD2</i>	0.996	1.000
sample_3	OKMLLP_12165	<i>htsA</i>	1.000	1.000
sample_3	OKMLLP_09180	<i>sirA</i>	1.000	1.000
sample_3	OKMLLP_05960	<i>sstD</i>	0.997	1.000
sample_19	CGIHNC_10565	<i>fhuD2</i>	1.000	1.000
sample_19	CGIHNC_11090	<i>htsA</i>	0.984	1.000
sample_19	CGIHNC_08085	<i>sirA</i>	0.993	1.000
sample_19	CGIHNC_04960	<i>sstD</i>	0.973	1.000
sample_36	PJGDJP_14580	<i>fhuD1</i>	0.970	0.997
sample_36	PJGDJP_17745	<i>fhuD2</i>	0.996	1.000
sample_36	PJGDJP_17060	<i>htsA</i>	1.000	1.000
sample_36	PJGDJP_02850	<i>sirA</i>	1.000	1.000
sample_145	BGIKGC_12995	<i>fhuD2</i>	1.000	1.000
sample_145	BGIKGC_09835	<i>htsA</i>	1.000	1.000
sample_145	BGIKGC_07025	<i>sirA</i>	0.996	1.000
sample_145	BGIKGC_05325	<i>sstD</i>	1.000	1.000
sample_186	KJDHGP_11450	<i>fhuD1</i>	1.000	1.000
sample_186	KJDHGP_10115	<i>fhuD2</i>	1.000	1.000
sample_186	KJDHGP_10635	<i>htsA</i>	1.000	1.000
sample_186	KJDHGP_07645	<i>sirA</i>	0.996	1.000
sample_186	KJDHGP_04450	<i>sstD</i>	0.997	1.000
sample_199	JNHCPM_00920	<i>fhuD1</i>	1.000	1.000
sample_199	JNHCPM_09720	<i>fhuD2</i>	0.996	1.000
sample_199	JNHCPM_10230	<i>htsA</i>	1.000	1.000
sample_199	JNHCPM_10260	<i>htsA</i>	1.000	1.000
sample_199	JNHCPM_11275	<i>sirA</i>	1.000	1.000
sample_199	JNHCPM_06095	<i>sstD</i>	0.997	1.000

Table V.S3: Results of **BLAST** search against the **UniProt** database for all identified structural homologs of the *S. aureus* samples. The individual query hits have been grouped by the representative sequence used for structural alignment within **Foldseek**.

Sample	Locus tag of Query	Target ID in <i>UniProt</i>	Gene Name	Similarity	Structure represented by
sample_145	BGIKGC_11210	A0A0D1I7H8	<i>btuF</i>	99.6	CGIHNC_05605
sample_19	CGIHNC_05605	A0A0D1I7H8	<i>btuF</i>	99.6	CGIHNC_05605
sample_2	GOALGM_06180	A0A0D1I7H8	<i>btuF</i>	99.6	CGIHNC_05605
sample_199	JNHCPM_14150	A0A0D1I7H8	<i>btuF</i>	99.6	CGIHNC_05605
sample_1	OEHLIH_07250	A0A0D1I7H8	<i>btuF</i>	99.6	CGIHNC_05605
sample_3	OKMLLP_06605	A0A0D1I7H8	<i>btuF</i>	99.6	CGIHNC_05605
sample_36	PJGDJP_12180	A0A0D1I7H8	<i>btuF</i>	99.2	CGIHNC_05605
sample_186	KJDHGP_05090	A0A0D1I7H8	<i>btuF</i>	98.9	CGIHNC_05605
sample_2	GOALGM_03610	A0A0H2XHK0	<i>sirA</i>	100	JNHCPM_11275
sample_199	JNHCPM_11275	A0A0H2XHK0	<i>sirA</i>	100	JNHCPM_11275
sample_1	OEHLIH_09825	A0A0H2XHK0	<i>sirA</i>	100	JNHCPM_11275
sample_3	OKMLLP_09180	A0A0H2XHK0	<i>sirA</i>	100	JNHCPM_11275
sample_36	PJGDJP_02850	A0A0H2XHK0	<i>sirA</i>	100	JNHCPM_11275
sample_19	CGIHNC_08085	A0A0H2XHK0	<i>sirA</i>	99.4	JNHCPM_11275
sample_145	BGIKGC_07025	A0A0H2XHK0	<i>sirA</i>	99.7	JNHCPM_11275
sample_186	KJDHGP_07645	A0A0H2XHK0	<i>sirA</i>	99.7	JNHCPM_11275
sample_145	BGIKGC_09835	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_2	GOALGM_00625	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_199	JNHCPM_10230	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_199	JNHCPM_10260	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_186	KJDHGP_10635	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_1	OEHLIH_12815	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_3	OKMLLP_12165	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_36	PJGDJP_17060	A0A0H2XEB0	<i>htsA</i>	100	KJDHGP_03365
sample_19	CGIHNC_11090	A0A7U7EVX6	<i>htsA</i>	99.7	KJDHGP_03365
sample_36	PJGDJP_11915	Q2FZM4	SAOUHSC_00976	100	KJDHGP_03365
sample_36	PJGDJP_12435	Q2FZM4	SAOUHSC_00976	100	KJDHGP_03365
sample_145	BGIKGC_09525	Q2FZM4	SAOUHSC_00976	99.7	KJDHGP_03365
sample_2	GOALGM_07915	Q2FZM4	SAOUHSC_00976	99.3	KJDHGP_03365
sample_199	JNHCPM_15125	Q2FZM4	SAOUHSC_00976	99.3	KJDHGP_03365
sample_186	KJDHGP_03365	Q2FZM4	SAOUHSC_00976	99.3	KJDHGP_03365
sample_1	OEHLIH_05515	Q2FZM4	SAOUHSC_00976	99.3	KJDHGP_03365
sample_3	OKMLLP_04865	Q2FZM4	SAOUHSC_00976	99.3	KJDHGP_03365
sample_19	CGIHNC_03875	Q2FZM4	SAOUHSC_00976	98.6	KJDHGP_03365
sample_2	GOALGM_13575	Q93PN3	<i>fhuD1</i>	100	OEHLIH_00205
sample_199	JNHCPM_00920	Q93PN3	<i>fhuD1</i>	100	OEHLIH_00205
sample_186	KJDHGP_11450	Q93PN3	<i>fhuD1</i>	100	OEHLIH_00205
sample_1	OEHLIH_00205	Q93PN3	<i>fhuD1</i>	100	OEHLIH_00205
sample_3	OKMLLP_12975	Q93PN3	<i>fhuD1</i>	100	OEHLIH_00205
sample_36	PJGDJP_14580	Q93PN3	<i>fhuD1</i>	97.2	OEHLIH_00205
sample_145	BGIKGC_01985	A0AAN2D5R8	<i>isdE</i>	100	OKMLLP_04230
sample_19	CGIHNC_03390	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_2	GOALGM_08555	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_199	JNHCPM_04095	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_186	KJDHGP_02890	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_1	OEHLIH_04875	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_3	OKMLLP_04230	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_36	PJGDJP_05395	A0AAN2D5R8	<i>isdE</i>	99.6	OKMLLP_04230
sample_145	BGIKGC_05325	O87828	<i>sstD</i>	99.1	OKMLLP_05960
sample_2	GOALGM_06825	O87828	<i>sstD</i>	98.8	OKMLLP_05960
sample_199	JNHCPM_06095	O87828	<i>sstD</i>	98.8	OKMLLP_05960
sample_186	KJDHGP_04450	O87828	<i>sstD</i>	98.8	OKMLLP_05960
sample_1	OEHLIH_06605	O87828	<i>sstD</i>	98.8	OKMLLP_05960
sample_3	OKMLLP_05960	O87828	<i>sstD</i>	98.8	OKMLLP_05960
sample_19	CGIHNC_04960	O87828	<i>sstD</i>	97.2	OKMLLP_05960
sample_145	BGIKGC_12995	A0A6H4EJ24	<i>fhuD2</i>	100	OKMLLP_11655
sample_19	CGIHNC_10565	A0A6H4EJ24	<i>fhuD2</i>	100	OKMLLP_11655
sample_186	KJDHGP_10115	A0A6H4EJ24	<i>fhuD2</i>	100	OKMLLP_11655
sample_2	GOALGM_01135	A0A6H4EJ24	<i>fhuD2</i>	99.6	OKMLLP_11655
sample_199	JNHCPM_09720	A0A6H4EJ24	<i>fhuD2</i>	99.6	OKMLLP_11655
sample_1	OEHLIH_12305	A0A6H4EJ24	<i>fhuD2</i>	99.6	OKMLLP_11655
sample_3	OKMLLP_11655	A0A6H4EJ24	<i>fhuD2</i>	99.6	OKMLLP_11655
sample_36	PJGDJP_17745	A0A6H4EJ24	<i>fhuD2</i>	99.6	OKMLLP_11655

V. Characterizing Siderophore Uptake and their Regulation Mechanisms in Corynebacteria based on Structural Homology and Transcriptome Profiling

Table V.S4: Genes identified by **PRESERVE** as structural homologs of the given references, but showing no significant expression changes under the conditions of the RNA-seq experiment.

Gene ID	Species	Description	$\log_2(\text{FC})$	$p\text{-adj}$
RS06090	CA	ABC transporter subst.-bind. prot.	0.33	0.063
RS11775	CA	iron-siderophore ABC transporter subst.-bind. prot.	0.68	0.571
DPLBDI_02275	CP	ABC transporter subst.-bind. prot.	-0.27	0.283
DPLBDI_03885	CP	Fe/B12 periplasmic-bind. domain-containing prot.	0.25	0.025
DPLBDI_09450	CP	iron ABC transporter subst.-bind. prot.	0.33	0.175

Table V.S5: Genes identified as differentially up-regulated genes ($p\text{-adj.} < 0.01$ and $\log_2(\text{FC}) > 1$) and labeled by **SignalP6** as lipoproteins, but not identified by **PRESERVE** as homologs of the reference proteins. Genes marked with an asterisk (*) represent those with a *LDDT* value slightly below the selected threshold, when compared with the reference proteins. Genes marked with a plus (+) represent those identified to be related to *SID*-uptake mechanisms via GO term annotation. The table has been divided into three sections: genes with a low coverage (< 0.5), genes with a low *LDDT* (< 0.4) and genes with a potential to be homologs (coverage > 0.9 and *LDDT* > 0.4)

Gene ID	Species	Description	$\log_2(\text{FC})$	$p\text{-adj}$	<i>LDDT</i>	Coverage
R3O64_00810	CH	sortase	1.21	9.2×10^{-25}	0.87	0.095
R3O68_02720	CS	hypothetical protein	1.28	1.6×10^{-21}	0.674	0.099
R3O64_08445	CH	hypothetical protein	1.98	4.6×10^{-52}	0.708	0.106
R3O64_01595	CH	hypothetical protein	1.79	3.8×10^{-05}	0.813	0.107
R3O64_07885	CH	AMIN domain-containing protein	1.49	3.2×10^{-03}	0.323	0.115
R3O68_04005	CS	glutaminy-peptide cyclotransferase	1.07	4.2×10^{-10}	0.223	0.158
RS13105	CA	hypothetical protein	4.8	2.7×10^{-19}	0.489	0.158
DPLBDI_01130	CP	Copper chaperone PCu(A)C	1.12	4.4×10^{-23}	0.463	0.175
R3O64_03745	CH	DUF4232 domain-containing protein	1.45	2.8×10^{-15}	0.47	0.184
R3O64_03770	CH	hypothetical protein	2.85	2.2×10^{-06}	0.245	0.209
R3O64_03735	CH	hypothetical protein	4.69	2.6×10^{-173}	0.372	0.211
+R3O64_01325	CH	iron uptake transporter deferriochelataze/eroxidase subunit	5.91	0	0.424	0.218
R3O64_08495	CH	choice-of-anchor I family protein	1.28	3.7×10^{-06}	0.18	0.242
DPLBDI_09965	CP	hypothetical protein	2.47	8.2×10^{-118}	0.358	0.482
R3O64_05235	CH	hypothetical protein	1.24	3.8×10^{-17}	0.222	0.49
R3O64_04910	CH	phosphate/phosphite/phosphonate ABC transporter subst.-bind. protein	1.49	5.3×10^{-14}	0.347	0.547
DPLBDI_00310	CP	Esterase PHB depolymerase	1.53	2.2×10^{-26}	0.226	0.623
R3O64_01330	CH	iron uptake system protein EfeO	6.71	0	0.266	0.826
R3O68_05250	CS	SGNH/GDSL hydrolase family protein	1.74	3.6×10^{-36}	0.343	0.985
*RS06480	CA	metal ABC transporter subst.-bind. prot.	1.98	1×10^{-214}	0.402	0.942
*R3O68_02005	CS	metal ABC transporter subst.-bind. prot.	5.58	0	0.402	0.942
*DPLBDI_04120	CP	zinc ABC transporter subst.-bind. prot.	1.56	7.4×10^{-34}	0.408	0.987
*R3O64_01665	CH	zinc ABC transporter subst.-bind. prot.	7.47	0	0.408	0.987



Figure V.S4: Overview of enriched GO terms of *Biological Process* ontology as computed using G0-Compass. The right heatmap shows the negative log of the significance value of the enrichment analysis and those below the significant threshold ($p - value < 0.05$) are marked with a dot (·). The shown terms are clustered into three distinct groups, rooted by the three least dispensable GO terms.

V. Characterizing Siderophore Uptake and their Regulation Mechanisms in Corynebacteria based on Structural Homology and Transcriptome Profiling

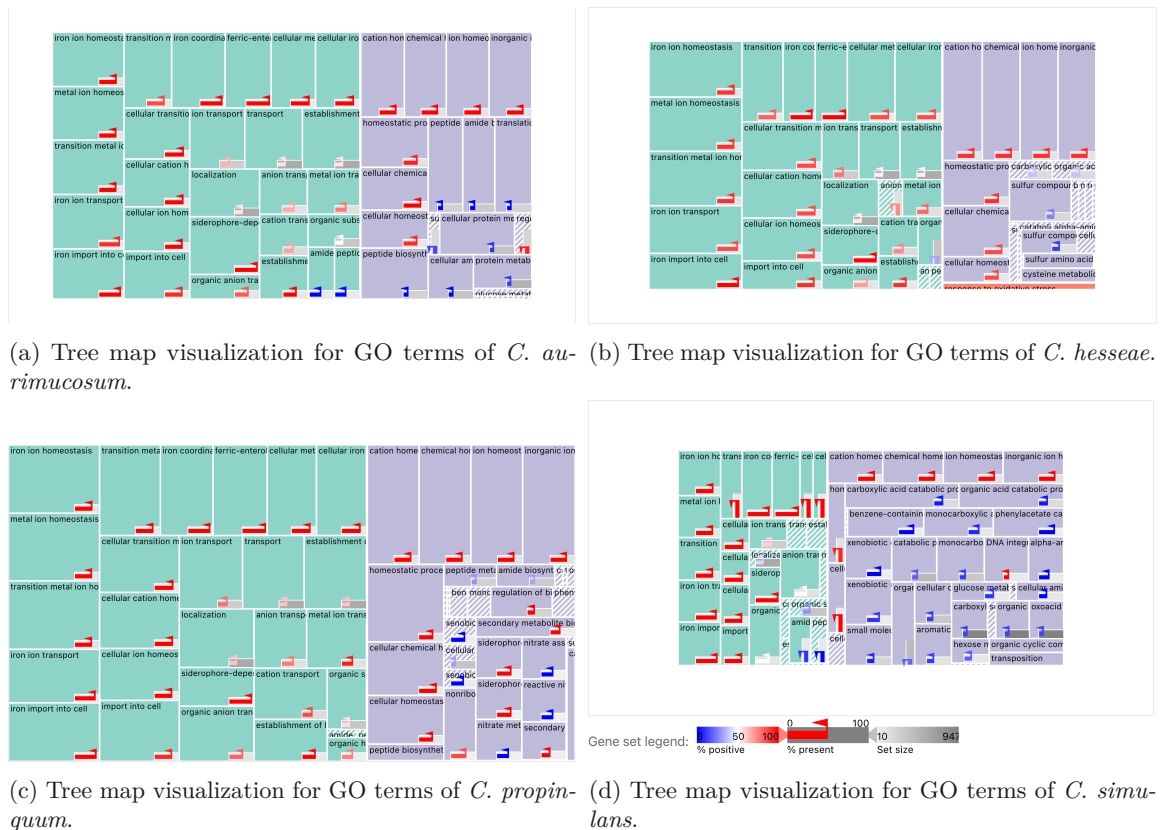


Figure V.S5: Tree maps summarizing the individual results of the gene ontology (GO) enrichment analysis with **GO-Compass** per species. For simplicity, the figure shows only results from the *Biological Process* ontology. The size of each tree map represents the $-\log(p\text{-value})$ of the term for the respective species. Rectangles with a pattern show those terms that are not significant for the species. The small glyph summarizes information of the genes related to the term. Most importantly, the color of the bar encodes the median log fold change of the set of genes related to the term. Additionally, the brightness represents how many genes are annotated with the GO term in the provided background, and the length of the bars represents the relative number of genes in the differentially expressed set.