Bioinformatics Group - Thesis projects

Last updated: 13 June 2024

In this document you can find a list of thesis projects that are available at the Bioinformatics group. Picking the right topic and supervisor for your Msc thesis is very important. Please spend some time thinking about your research interests and your strong and weak points, reading carefully through this document, and choosing among these topics.

To guide your decision, we mention the following requirements, that you can also use for a text search in this document:

- Supervisor – you can find more information on our group here: https://www.wur.nl/en/Research-Results/Chair-groups/Plant-Sciences/Bioinformatics/People.htm, in addition co-supervisors from other chair groups may be listed

- Type – the type of project
  - **Data analysis**: analysis of a specific dataset to address a biological question, often with collaborators; work usually involves the construction of one or more dedicated pipelines and/or data analysis scripts (in R or Python)
  - **Workflow development**: combining existing software tools to analyze any dataset in a specific setting
  - **Method/algorithm development**: writing new tools or routines, to solve a specific problem or to add functionality to an already existing tool
  - **Software engineering**: developing user-friendly software for biological data analysis

- Requirements – the courses required for that project
  - Advanced Bioinformatics (BIF30806)
  - Advanced Statistics (MAT20306) or similar
  - Algorithms in Bioinformatics (BIF31306)
  - Data Analysis & Visualization (BIF51306)
  - Deep Learning (GRS34806)
  - Machine Learning (FTE35306)
  - Molecular Systems Biology (SSB30306)
  - Plant Plasticity and Adaptation (PPH30806)
  - Programming in Python (INF22306)
  - Software Engineering (INF32306)

- Skills – the skills that you will need and train in the project
  - Algorithm development
  - Biological sequence analysis
  - Chemistry
  - Chemical biology
  - Comparative genomics
  - Databases
  - Deep learning

    - Genome assembly
    - Genomics
    - Machine learning
    - Mass spectrometry
    - Metabolism
    - Metabolomics
    - Metagenomics
    - Microbial ecology
    - Pangenomics
    - Phylogenetics
    - Population genetics
    - Programming
    - Proteomics
    - Python
    - Script programming
    - Statistics
    - Transcriptomics
    - Visualization

- Organisms – the type of organisms the project is about
    - General (from various sources)
    - Bacteria/archaea
    - Fungi
    - Viruses
    - Protists/algae
    - Animals
    - Plants
    - Humans

Project list

Developing strategies to enhance gene cluster family assignment algorithms

Exploring the origin and evolution of desiccation tolerance in plants using comparative transcriptomics.

Mapping the salt-induced gene regulatory network that guides root branching

Functional annotation of a non-redundant proteome

Supervisor	Lakhansing Pardeshi, Sandra Smit
Type	Method development
Requirements	Advanced Bioinformatics
Skills	Programming, Pangenomics, Algorithm development
Timestamp	June 2024

Description

Running InterProScan, a tool for functional annotation of proteins, takes a very long time when annotating all proteins in a bacterial pangenome, containing hundreds of genomes. To avoid recalculation of results, InterProScan provides a lookup match service, where it searches query sequences for an exact match in the pre-calculated results of the UniProtKB sequences. However, this set is not comprehensive to cover the prokaryotic proteome.

Pangenomics of prokaryotes usually involves 100s of genomes and this can easily take up to a month or two to process all the proteins from these genomes. These pangenomes are usually at the species level, meaning there will be significant redundancy at the protein sequence level. Recalculation of InterProScan results of these redundant proteins in the panproteome can be avoided to speed-up the processing.

In this project you will develop a pipeline that will run InterProScan only on the non-redundant panproteome, encoded in the pangenome. We cluster sequences at 100% identity and run InterProScan only once. It needs to be investigated what the most efficient way is to create the non-redudnant set (inside the pangenome/panproteome or before pangenome construction) and to re-assign all terms across the redundant set. Various bacterial pangenomes (e.g. Pectobacterium) are available for development.

This project will result in a pre-processing pipeline to efficiently perform functional annotation of bacterial genomes before/while building a pangenome or panproteome. This pipeline can be easily extended to include other functional annotation tools.

Optionally, if time allows and required skills are present, you can develop/extend a PanTools functionality to expand the functional annotation of one representative sequence to remaining identical sequences in the pan-proteome. This would involve Java programming in PanTools or possibly a prototype implementation in Python to manipulate the pangenome graph in Neo4j.

References

[1] Jonkheer, E. M., van Workum, D.-J. M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J. R., Berke, L., J van der Lee, T. A., de Ridder, D., & Smit, S. (2022). PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics*, *38*(18), 4403–4405.
[2] Jonkheer, Eef M., et al. "The Pectobacterium pangenome, with a focus on Pectobacterium brasiliense, shows a robust core and extensive exchange of genes from a shared gene pool." BMC genomics 22 (2021): 1-18.

Diversity and evolution of Arabidopsis latent virus 1 using public sequencing data

Supervisor	Anne Kupczok, René van der Vlugt (Virology)
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	RNASeq analysis, Comparative genomics
Organism	Viruses, plants
Timestamp	May 2024

Description

Many plant biology studies involve Arabidopsis thaliana as a model system. However, it is rarely tested whether the seeds and plants used for experiments do contain viruses. Plant virus infections can often be latent, i.e., they do not cause obvious disease symptoms, and can sometimes even be beneficial for the host, in particular under stress conditions (Xu et al. 2008, González et al. 2020). In these cases, it is difficult to detect the viruses based on the plant phenotype, but instead high-throughput sequencing methods can reveal their presence. Recent advances in algorithms for virus discovery have greatly enriched the known diversity of RNA viruses (Edgar et al. 2022). In addition, the analysis of plant genomes revealed the widespread occurrence of endogenous viral elements (EVEs), i.e., viral genomes or fragments of viral genomes integrated into the genomes of their eukaryote hosts (Chiba et al. 2011).

We have recently discovered ArLV1, an latent virus which can also be found in many public A. thaliana RNA sequencing data (Verhoeven et al. 2023) which suggests a global distribution. Initial analyses suggest different viral clades; however, its precise diversity remains be investigated. In addition the possible presence as EVEs in A. thaliana genomes potentially linked to the virus asymptomatic nature has not yet been investigated. Here we will study the diversity of ArLV1. To this end, we will apply a pipeline to reconstruct novel genomes from publicly available RNASeq data of A. thaliana. In addition, public A. thaliana genome data can be used to find EVEs with similarity to ArLV1. The detected genomes will be analyzed to investigate the sequence diversity and evolution of these viruses.

References

Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, Suzuki N. 2011. Widespread Endogenization of Genome Sequences of Non-Retroviral RNA Viruses into Plant Genomes. PLOS Pathogens 7:e1002146.

Edgar RC, Taylor J, Lin V, Altman T, Barbera P, Meleshko D, Lohr D, Novakovsky G, Buchfink B, Al-Shayeb B, et al. 2022. Petabase-scale sequence alignment catalyses viral discovery. Nature 602: 142–147.

González R, Butković A, Elena SF. 2020. Chapter Three - From foes to friends: Viral infections expand the limits of host phenotypic plasticity. In: Kielian M, Mettenleiter TC, Roossinck MJ, editors. Advances in Virus Research. Vol. 106. Academic Press. p. 85–121. Available from: http://www.sciencedirect.com/science/article/pii/S0065352720300038

Verhoeven A, Kloth KJ, Kupczok A, Oymans GH, Damen J, Rijnsburger K, Jiang Z, Deelen C, Sasidharan R, van Zanten M, et al. 2023. Arabidopsis latent virus 1, a comovirus widely spread in Arabidopsis thaliana collections. New Phytologist 237:1146–1153.

Xu P, Chen F, Mannas JP, Feldman T. Sumner LW, Roossinck MJ. 2008 Virus infection improves drought tolerance. New Phytologist 180; 911-921.

Pathway analysis in a Solanaceae pangenome

Supervisor	Christina Papastolopoulou, Sandra Smit
Type	Data analysis, workflow development
Requirements	Advanced Bioinformatics
Skills	Programming, Comparative (pan)genomics, Visualization
Timestamp	June 2024

Description

The Solanaceae family is one of the most economically important plant families containing diverse species with nutritional values such as tomato, pepper, potato and eggplant, medicinal species such as tobacco and *Datura* spp., and ornamentals such as petunia and many others. This diversity is reflected with variations in genome sizes, genomic rearrangements, repeat content and ultimately pathway variation among the species.

Researchers want to identify the way that pathways associated with crucial traits are evolving, but they are limited by the variation in the genes and genomic loci [1]. By providing a solution to efficiently compare the genes/pathways across large evolutionary distances and facilitating their grouping and functional annotation, we provide a solid source of information to understand pathway evolution and diversity. To tackle this issue we employ a pangenome, a graph-based representation to capture the variation in multiple genome assemblies. We will use PanTools [2, 3], a software package for pangenomics.

The overall aim of this project is to devise a strategy to perform efficient pathway analysis in a family-level pangenome and apply it to a Solanaceae pangenome. This involves the construction of a family-level pangenome with representative genomes from the Solanaceae family. Within this pangenome you will perform in-depth analysis of some relevant gene families (e.g. fruit development). Devising a strategy for this will include optimization of homology grouping, and testing novel functionalities to analyze Regions of Interest (ROIs) and to perform gene-set analysis in a pangenome. This will deliver important feedback to the PanTools developers. In addition, we will gain insights into the evolution of the target genes/pathways.

Optionally, if time allows, a comparison could be made between application in a family-level pangenome and in a genus-level pangenome (e.g. capsicum or tomato only), resulting in pros/cons and important considerations for this methodology.

References:

[1] Keithellakpam S (2024) Genetics and evolutionary insights from Solanaceae genome sequences. Plant Systematics and Evolution, 1-16, 310(1).

[2] Jonkheer, E. M., van Workum, D.-J. M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J. R., Berke, L., J van der Lee, T. A., de Ridder, D., & Smit, S. (2022). PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics*, *38*(18), 4403–4405.

[3] Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S (2016) PanTools: representation, storage and exploration of pan-genomic data. Bioinformatics 32(17):i487-i49.

Responses of environmental microbiota to extrinsic synthetic microbial communities

Supervisor	Jun Liu, Anne Kupczok
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Metagenomics, Metatranscriptomics
Organism	Bacteria/archaea
Timestamp	March 2024

Description

In recent years, synthetic microbial communities have become a powerful tool to achieve some challenging tasks in various areas, such as plant protection, waste recycling, and disease prevention of humans (Shahab, et al. 2020; Cheng, et al. 2022). Previous studies focused on designing and building workable microbial communities applied in complex environments, for example in guts (Clark, et al. 2021). However, it is still unclear how the native microbiota responds to the extrinsic synthetic microbial community and how that influences the performance of the synthetic microbial community. These knowledge gaps limit our understanding of applying synthetic microbial communities in complex ecosystems.

In this study, metagenomics and metatranscriptomics data will be analyzed to explore the response of manure microbiota to an established synthetic community of lactic acid bacteria species (LAB SynCom). We have already shown that the LAB SynCom strategy mitigated ammonia emissions (a dominant precursor of PM2.5 and global air pollution) produced by pig manure microbiota with a high mitigation efficiency (Liu et al. 2023). However, how the manure microbiota responds to the LAB SynCom is still unknown. This project will include the following five aims:

1) Building a pipeline to analyze metagenomic and metatranscriptomic data;

2) Classifying manure local microbial communities according to metabolic function and ecological niches;

3) Reconstructing the C and N metabolic pathways from metagenomic data;

4) Characterizing the gene expression differences in C and N metabolic pathways;

5) Establishing the metabolic interaction pattern between LAB SynCom and native communities with the metatranscriptomic data.

References

Cheng, A, G. et al. (2022). Design, construction, and in vivo augmentation of a complex gut microbiome. Cell, 185 (19), 3617-3636. Doi: 10.1016/j.cell.2022.08.003.

Clark, R, L. et al. (2021). Design of synthetic human gut microbiome assembly and butyrate production. Nature Communications, 12(1), doi: 10.1038/s41467-021-22938-y.

Liu, J. et al. (2023). Highly efficient reduction of ammonia emissions from livestock waste by the synergy of novel manure acidification and inhibition of ureolytic bacteria, Environment International. 172. Doi: 10.1016/j.envint.2023.107768.

Shahab, R, L. et al. (2020). A heterogeneous microbial consortium producing short-chain fatty acids from lignocellulose. Science, 369(6507). doi: 10.1126/science.abb1214.

Microbial pathways of nitrogen emissions from the global livestock sector

Supervisor	Jun Liu, Anne Kupczok
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Metagenomics
Organism	Bacteria/archaea
Timestamp	March 2024

Description

Nitrogen emissions mainly including NO, N2O, and NH3 are dominant precursors of the global climate change, air pollution, biodiversity loss, and threat to human health (Mueller, et al. 2020). It is highly urgent to mitigate nitrogen emissions to achieve a sustainable development of the planet (Schulte, et al. 2022). The global sources and hotspots of nitrogen emissions are well characterized. Previous studies have revealed that Asia, Europe, and America contribute to more than 80% of nitrogen emissions and that the livestock sector is one of the major hotspots of nitrogen emissions (Mueller, et al. 2020). Specifically, the farming systems of dairy cattle, beef cattle, pig, chicken, and buffalo milk accounted for around 80% of nitrogen emissions. In the livestock farming system, the manure management chain is one of the largest sources of nitrogen emissions, which mainly originate from the microbial nitrogen metabolism (Uwizeye, et al. 2020; Sigurdarson, et al. 2018). Pathways of nitrogen emissions are well studied in single microbes. However, microbes in feces and manure reside in communities and it is unclear how nitrogen emissions are generated on the community level. Moreover, the global geographic distribution of nitrogen emission producers is still unknown. Answering these challenging but pivotal questions is crucial for the sustainable development of the global livestock sector.

In this project, we will analyze public metagenomics data to identify functional genes involved in nitrogen metabolisms. We aim to reconstruct livestock fecal microbial nitrogen metabolic pathways that generate NO, N2O, and NH3. A pipeline will be developed for identifying functional genes from genomes and metagenome-assembled genomes (MAGs), where the data will be collected from world-wide distributed metagenomic samples of a typical livestock (beef for example). Based on the identified genes, the contributing taxa will be inferred and relative abundances of contributing taxa will be computed on species or strain level (e.g., with MetaPhlAn 4) (Blanco, et al. 2023). Based on the identified functional genes and contributing species, the global distribution of nitrogen metabolic genes and contributing microbes of livestock fecal microbiota will be characterized.

References

Blanco, A. et al. (2023). Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. Nature Biotechnology, 41(11):1633-1644. doi: 10.1038/s41587-023-01688-w.

Mueller, N. D. et al. (2020). Nitrogen challenges in global livestock systems. Nature Food, 7(1), 400-401. doi: 10.1038/s43016-020-0117-7.

Schulte-Uebbing, L. F. et al. (2022). From planetary to regional boundaries for agricultural nitrogen pollution. Nature, 610(7932), 507-512. doi: 10.1038/s41586-022-05158-2.

Sigurdarson, J. J. et al. (2018). The molecular processes of urea hydrolysis in relation to ammonia emissions from agriculture. Reviews in Environmental Science and Bio/Technology, 2(17), 241-258. doi: 10.1007/s11157-018-9466-1.

Uwizeye, A. et al. (2020). Nitrogen emissions along global livestock supply chains. Nature Food, 7 (1), 437–446. doi: 10.1038/s43016-020-0113-y.

Implicit Neural Representations of Molecular Surfaces

Supervisor	Daniel Probst
Type	Method/algorithm development, Software engineering
Requirements	Advanced Statistics, Programming in Python, Machine Learning
Skills	Chemistry, Chemical biology, Machine learning, Programming
Organism	General
Timestamp	June 2024

Description

Coordinate-based neural networks, sometimes called Implicit Neural Representations (INRs) [1], used for instance in NeRFs [2], learn to represent highly complex functions by treating pairs of coordinates and associated function evaluations as a dataset. These representations turn out to be extremely useful because they are resolution-independent and allow one to easily manipulate the underlying function via manipulations of its latent representation. Most of the applications so far have occurred in the realm of images, videos, or computer graphics. There are, however, situations - in chemistry, biology, and physics - where one would like to represent a (family of) continuous closed surfaces efficiently. In chemistry, for instance, such a function could represent the van der Waals surface of a molecule and be used to model chemical interactions. In this project, we will train INR-type neural architecture to efficiently represent families of surfaces by mapping them first to spherical functions [3]. This common representation will allow us to build useful inductive biases, such as rotation invariance, and provide a unique domain to compare surfaces for classification or regression tasks. We will apply this system to learn representations of small molecules and solve benchmark tasks in chemistry.

References

[1] https://www.vincentsitzmann.com/siren

[2] http://www.matthewtancik.com/nerf

[3] https://arxiv.org/abs/2301.04695

Wavefunction collapse algorithm for de novo molecular design

Supervisor	Daniel Probst
Type	Method/algorithm development, Software engineering
Requirements	Advanced Statistics, Programming in Python, Machine Learning
Skills	Chemistry, Machine learning, Programming, Statistics
Organism	General
Timestamp	June 2024

Description

The wavefunction collapse algorithm is a conceptually simple generative algorithm inspired by wavefunction superposition [1]. The algorithm has been used for procedural level and model design [2,3] and explored in the context of discriminative learning [4]. Furthermore, it has been extended to graph generation [5]. De novo molecular design is developing as a sub-field of chem- and bioinformatics concerned with sampling molecules from a conceptual chemical space [6]. Recently, generative machine learning has led to an explosion in available methods capable of exploring this chemical space while imposing constraints such as drug-likeness or synthetic accessibility [7,8]. However, these models rely on large amounts of training data, potentially introducing unknown biases and largely limiting the scope of generated samples to already explored areas of the chemical space. A wavefunction collapse algorithm-based approach has the potential to solve this problem while being both interpretable and computationally efficient.

As a molecule is generally represented by its molecular graph, which may also include the spatial coordinates of the atoms, de novo molecular design can be posed as a graph generation problem. The main goal of this project is to explore the wavefunction collapse algorithm in the context of de novo molecular design, focusing on its extension to graphs and discriminative learning. Specifically the project aims to achieve the following:

- Implementing the Python scripts necessary to facilitate the application of the wavefunction collapse algorithm to molecular graphs.
- Training and testing a discriminative learning-based generator capable of creating molecular graphs with drug-like properties.
- (Bonus) Extending the approach to the tertiary structure of biological macromolecules, such as proteins.

References

[1] https://github.com/mxgmn/WaveFunctionCollapse

[2] https://andymakesgames.tumblr.com/post/182363131350/global-game-jam-2019-maureens-chaotic-dungeon

[3] https://github.com/marian42/wavefunctioncollapse

[4] https://arxiv.org/abs/1809.04432

[5] https://github.com/lamelizard/GraphWaveFunctionCollapse/tree/master

[6] https://www.sciencedirect.com/science/article/pii/S1359644621002531?via%3Dihub

[7] https://link.springer.com/chapter/10.1007/978-3-030-01418-6_41

[8] https://arxiv.org/abs/1805.11973

Explaining gene expression variation in lettuce by promoter sequence variation

Supervisor	Dirk-Jan van Workum, Sandra Smit
Type	Workflow development, data analysis
Requirements	Advanced Bioinformatics
Skills	Programming, Comparative (pan)genomics, Algorithm development
Timestamp	June 2024

Description

Understanding the relationship between genomic variation and gene expression is essential for deciphering many plant traits. Within the LettuceKnow project (https://lettuceknow.nl/), many RNA-seq experiments are performed with the goal of better understanding pathogen resistance and architecture traits in lettuce. Of special interest for architecture traits are understanding bolting (the start of flowering) time, the response to light and temperature stress and aging. Bolting makes lettuce taste bitter (and thus unmarketable); light and temperature stress are major factors in helping lettuce adapt to both climate change and vertical farming systems; and aging of leaves is undesirable as they have to be removed prior to selling. All of these traits involve dedicated molecular pathways for the integration of signals. The experiments are set up using hundreds of different lettuce accessions such that the variation in response can aid us in understanding the underlying genetics of gene regulation for these traits.

Binding of transcription factors (TFs) is one of the major forces behind regulating gene expression and integrating signals. TFs can both positively and negatively impact gene expression, which is often a process in which multiple TFs play a role. TFs physically interact with the DNA via so-called TF binding sites (TFBSs). TFBSs are short DNA sequences recognised by a DNA binding domain of the TF. In previous projects, a full inventory of TFs in lettuce has been compiled as well as a workflow defined for identifying TFBSs in a given promoter sequence (across hundreds of accessions based on homology). We hypothesize that variation in the presence/absence of TFBSs in a promoter of a gene can explain (part of) the expression of the gene.

In this project, you will look for correlations between TFBS variation and gene expression variation in lettuce. Starting with promising candidates, you will develop a (reproducible) workflow that enables large-scale analysis of these associations. This workflow has the potential to discover pathways and genes in lettuce that are of major importance to lettuce breeding. When a candidate is found relevant to the LettuceKnow project, there is a possibility for a detailed bioinformatic analysis of the gene which will be very helpful to the molecular biologist(s) working on it in the lab!

Relevant literature

- Wei, T., *et al*. (2021). Whole-genome resequencing of 445 Lactuca accessions reveals the domestication history of cultivated lettuce. Nature Genetics, 53(5), 752-760.
- Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. Nature reviews Molecular cell biology, 19(10), 621-637.
- Swinnen, G., Goossens, A., & Pauwels, L. (2016). Lessons from domestication: targeting cis-regulatory elements for crop improvement. Trends in Plant Science, 21(6), 506-515.
- Takagi, H., Hempton, A. K., & Imaizumi, T. (2023). Photoperiodic flowering in Arabidopsis: Multilayered regulatory mechanisms of CONSTANS and the florigen FLOWERING LOCUS T. Plant Communications.

Machine learning based inference of horizontal gene transfer in bacterial pangenomics

Supervisor	Anne Kupczok
Type	Method development, Data analysis
Requirements	Advanced Bioinformatics, Machine Learning
Skills	Population genetics, Machine learning, Deep learning
Organism	Bacteria/archaea
Timestamp	February 2024

Description

Population genetics is the study of the genetic makeup of populations. In a population, genotype and phenotype frequency distributions vary in response to the processes of mutation, gene flow, genetic drift, and natural selection. Many powerful and established model-based estimators and inference tools for population genetics data exist and are widely applied. Nevertheless, some difficult problems like the detection of sites under selection benefit from the rise of machine learning tools for population genetics. However, modern machine learning tools require massive training datasets of independent data points which are usually not available for inference tasks in population genetics. This is a simple consequence of the fact that all individuals of a population are related to each other, such that no independent subpopulations exist. Thus, the best approach to train neural networks in population genetics is based on simulations.

In bacterial populations, individuals can vary in gene content to a large extent, where only a small fraction of genes can be found in all genomes of a species. The pangenome of a population, i.e., the set of all genes present in the population, arises due to horizontal gene transfer and differential gene loss. Horizontal gene transfer is one of the most puzzling features to reconstruct and assess in bacterial population genetics. In this project, we aim to combine modern efficient simulation tools for bacterial pangenomes with simulation-based machine learning techniques to infer the horizontal gene transfer rate of many bacterial species. The project will be carried out in close collaboration with Franz Baumdicker, Tübingen University, Germany.

References

Baumdicker F, Kupczok A. 2023. Tackling the Pangenome Dilemma Requires the Concerted Analysis of Multiple Population Genetic Processes. Genome Biology and Evolution 15:evad067.

Brockhurst MA, Harrison E, Hall JPJ, Richards T, McNally A, MacLean C. 2019. The Ecology and Evolution of Pangenomes. Current Biology 29:R1094–R1103.

Korfmann K, Gaggiotti OE, Fumagalli M. 2023. Deep learning in population genetics. Genome Biology and Evolution:evad008.

Schrider DR, Kern AD. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. Trends in Genetics 34:301–312.

Deciphering interactions among bacterial defense systems

Supervisor	Anne Kupczok
Type	Data analysis, Method/algorithm development
Requirements	Advanced Bioinformatics, Programming in Python
Skills	Comparative genomics, phylogenetics
Organism	Viruses, bacteria/archaea
Timestamp	February 2024

Description

Bacteriophages (short: phages) are viruses that infect bacteria. In the arms race of bacteria and phages, bacteria have evolved a huge diversity of defense strategies against invading phages (Georjon and Bernheim 2023). Such defense systems typically consist of multiple genes. One prominent member is the adaptive immune system CRISPR/Cas named after its repeat structure containing sequences specific for immunity (CRISPR) and the associated genes (Cas). Interestingly, these defense systems often occur closely together in the genome and they are often mobile, i.e., they only occur in some strains of a bacterial species and tend to move between strains by horizontal gene transfer. It has been observed that certain defense systems co-occur in bacterial genomes, suggesting that they have synergistic effects (Tesson and Bernheim 2023, Wu et al. 2024).

In this project, we will extend a novel phylogenetic method to study the co-occurrence or avoidance between genes (Gavriilidou et al. 2024) by including genomic location and genomic proximity in the output. The extended tool can then be applied to microbial pangenomes, with a particular focus on defense systems that have been predicted computationally (Tesson et al. 2022). This analysis will reveal which genes tend to co-occur with bacterial defense systems and where they can be found in the genome. The project will be carried out in close collaboration with Franz Baumdicker, Tübingen University, Germany

References

Gavriilidou A, Paulitz E, Resl C, Ziemert N, Kupczok A, Baumdicker F. 2024. Goldfinder: Unraveling Networks of Gene Co-occurrence and Avoidance in Bacterial Pangenomes. :2024.04.29.591652. Available from: https://www.biorxiv.org/content/10.1101/2024.04.29.591652v1

Georjon H, Bernheim A. 2023. The highly diverse antiphage defence systems of bacteria. Nat Rev Microbiol 21:686–700.

Tesson F, Bernheim A. 2023. Synergy and regulation of antiphage systems: toward the existence of a bacterial immune system? Current Opinion in Microbiology 71:102238.

Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, Bernheim A. 2022. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun 13:2561.

Wu Y, Garushyants SK, van den Hurk A, Aparicio-Maldonado C, Kushwaha SK, King CM, Ou Y, Todeschini TC, Clokie MRJ, Millard AD, et al. 2024. Bacterial defense systems exhibit synergistic anti-phage activity. Cell Host & Microbe [Internet].

Analyzing regions of interest for temperature tolerance in a cauliflower pangenome

Supervisor	Cauliflower-PhD, Sandra Smit
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Programming, Pangenomics, Visualization
Timestamp	June 2024

Description

Cauliflower is an important vegetable that can grow in various climatic regions. However, it is sensitive to high ambient temperatures during development, which might result in delayed curd induction and poor curd quality [1]. Some cauliflower accessions are more sensitive than others. The genetic basis underlying temperature tolerance is not well understood yet.

A large data set on the effects of temperature in ~200 different cauliflower genotypes will be generated, and genome-wide association studies will be performed, ideally leading to genomic regions associated with the trait. We already know temperature tolerance is a complex trait with multiple genes involved. In the past a single reference genome was not sufficient to find the causal gene(s). Therefore, a pangenomic approach is needed to study genetic variation across multiple genomes.

In this MSc project you will collect cauliflower genomes [3,4] (assemblies and resequenced accessions), construct and annotate a cauliflower pangenome using PanTools [2] and study genetic variation in genes/regions of interests that were identified in earlier work. This will result in a strategy to analyze the novel GWAS regions once they are identified. It will also provide essential feedback to current features in PanTools that are being developed. And it will start to shine light to the complexity underlying temperature tolerance in cauliflower.

You will work as part of the Comparative PanGenomics team, exploring state-of-the-art software and visualization, while working on your own data set.

References

[1] XiaoXue Sun, Johan Bucher, Yongran Ji, Aalt D.J. van Dijk, Richard G.H. Immink, Guusje Bonnema (2018) Effect of ambient temperature fluctuation on the timing of the transition to the generative stage in cauliflower. *Environmental and Experimental Botany* 155: 742-750.

[2] Jonkheer, E. M., van Workum, D.-J. M., Sheikhizadeh Anari, S., Brankovics, B., de Haan, J. R., Berke, L., J van der Lee, T. A., de Ridder, D., & Smit, S. (2022). PanTools v3: functional annotation, classification and phylogenomics. *Bioinformatics*, *38*(18), 4403–4405.

[3] Wang, Y., Ji, J., Fang, Z., et al. BoGDB: An integrative genomic database for Brassica oleracea L.. Front. Plant Sci 13,852291 (2022). http://www.bogdb.com/

[4] Guo N, et al. A graph-based pan-genome of Brassica oleracea provides new insights into its domestication and morphotype diversification. Plant Commun. 2024 Feb 12;5(2):100791.

The evolution of defense and anti-defense genes in bacteria and bacteriophages

Supervisor	Anne Kupczok
Type	Data analysis
Requirements	Advanced Bioinformatics, Programming in Python
Skills	Comparative genomics, databases, phylogenetics, python, statistics
Organism	Viruses, bacteria/archaea
Timestamp	February 2024

Description

Bacteriophages (short: phages) are viruses that infect bacteria. In the arms race of bacteria and phages, bacteria have evolved a huge diversity of defense strategies against invading phages (Georjon and Bernheim 2023). One prominent member is the adaptive immune system CRISPR/Cas named after its repeat structure containing sequences specific for immunity (CRISPR) and the associated genes (Cas). Phages, in turn, came up with a variety of counter-defense strategies to circumvent these defense systems. In particular, various anti-CRISPR proteins encoded by phages have been described (Pawluk et al. 2018, Peng et al. 2020). Interestingly, also homologs of the Cas proteins that function in the CRISPR system can be encoded on viral genomes where they are likely involved in counter-defense (Faure et al. 2019).

Thus, homologous proteins are encoded in bacteria and phage genomes that are involved in defense and anti-defense mechanisms. This gives rise to the intriguing research question where these genes originated and how they evolve. Phages evolve rapidly (Dion et al. 2020), which gives rise to interesting evolutionary scenarios, where gene versions that evolved in phages could be transferred to bacteria.

This project will study defense systems and their occurrences in phages, where they could function as defense or anti-defense mechanisms. To this end, publicly available genomes of phages will be scanned for defense genes (Tesson et al. 2022). For selected groups, phylogenetic analysis will be performed to study their evolutionary history in detail and to potentially identify gene transfer between phages and bacteria.

References

Dion MB, Oechslin F, Moineau S. 2020. Phage diversity, genomics and phylogeny. Nat Rev Microbiol 18:125–138.

Faure G, Shmakov SA, Yan WX, Cheng DR, Scott DA, Peters JE, Makarova KS, Koonin EV. 2019. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. Nature Reviews Microbiology 17:513–525.

Georjon H, Bernheim A. 2023. The highly diverse antiphage defence systems of bacteria. Nat Rev Microbiol 21:686–700.

Pawluk A, Davidson AR, Maxwell KL. 2018. Anti-CRISPR: discovery, mechanism and function. Nature Reviews Microbiology 16:12.

Peng X, Mayo-Muñoz D, Bhoobalan-Chitty Y, Martínez-Álvarez L. 2020. Anti-CRISPR Proteins in Archaea. Trends in Microbiology 28:913–921.

Tesson F, Hervé A, Mordret E, Touchon M, d'Humières C, Cury J, Bernheim A. 2022. Systematic and quantitative view of the antiviral arsenal of prokaryotes. Nat Commun 13:2561.

Tracing the evolution of plant cell types

Supervisors: Kelvin Adema (Molecular Biology) & Rens Holmer (Bioinformatics)
Type: Comparative transcriptomics, single-cell transcriptomics
Requirements: Advanced bioinformatics, Genomics & Machine learning / Statistics for data scientists
Skills: Programming (R/Python), data analysis, knowledge on plant biology is a pre
Timestamp: February 2024

Complex multicellular organisms consist of distinct cell types expressing unique combinations of genes to fulfill specialized functions. Cell types, like their gene counterparts, are subject to evolutionary changes. For a novel cell type to arise over evolutionary time it must accumulate genomic alterations that enable the expression and maintenance of distinct genes(Arendt, et al., 2016).

The advent of single-cell RNA-sequencing has allowed us to identify the transcriptional program defining unique cell types and enabled cross-species cell type comparisons of gene regulatory networks (Boroviak, et al., 2018; Geirsdottir, et al., 2019). This potential can also be harnessed in plants. A primary cell type of interest would be the cortex. *Arabidopsis thaliana* has a single cortex layer, whereas *Medicago truncatula* has five cortex layers likely with distinct functions. Our objective is to unravel the evolutionary relationship between these cortex layers across species. Your goal would be to build a foundation for cross-species comparisons of plant single-cell RNA sequencing datasets. Ultimately, we aim to trace the evolution of plant cell types.

Comparative single-cell genomics has advanced primarily in mammalian systems, utilizing the concepts of orthology (Moravec et al, 2023; Mah & Dunn, 2023). However, applying these analyses in plants is more complex due to wide-spread whole-genome duplications, complicating the identification of 1:1 orthologous gene relationship. Key challenges in this thesis will be: 1) Linking orthology information to single-cell RNA datasets, 2) Using orthology information to integrate cross-species datasets and 3) to visualize relationships between cell types across species.

Arendt, D., et al. (2016). The origin and evolution of cell types. *Nature reviews genetics, 17*, 747-757. doi:https://doi.org/10.1038/nrg.2016.127

Boroviak, T., et al. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development, 145*(12). doi:10.1242/dev.167833

Geirsdottir, L., et al. (2019). Cross-Species Single-Cell Analysis Reveals Divergence of the Primate Microglia Program. *Cell*, 1609-1622. doi:https://doi.org/10.1016/j.cell.2019.11.010

Mah, J. L., & Dunn, C. W. (2023). Reconstructing cell type evolution across species through cell phylogenies of single-cell RNAseq data. *BioRxiv*. doi:https://doi.org/10.1101/2023.05.18.541372

Moravec, J. C., et al. (2023). Testing for Phylogenetic Signal in Single-Cell RNA-Seq Data. *Journal Of Computational Biology, 30*(4), 518–537. doi:10.1089/cmb.2022.0357

Integrating data to find heat and drought resilience gene modules in plants

Supervisors: Ben Noordijk, Dick de Ridder
Type: Workflow development, Data analysis
Requirements: Programming in Python
Skills: Transcriptomics, Machine Learning, Programming, Python
Timestamp: November 2023

Climate change poses significant challenges to global food security due to its negative impact on crop growth and productivity. With a growing world population, especially in climate-vulnerable areas, it becomes essential to develop crop varieties that are resilient to combinatorial stresses such as heat and drought. However, our current understanding of plant responses to combinations of stresses is limited.

A plant's environmental response is governed by gene regulatory networks (GRNs); sets of genes which influence each other's expression. Stress-response GRNs typically involve a vast number of genes and an even larger number of regulatory interactions, making detailed modelling of the entire network impractical. In this research (part of the PlantXR consortium [1]), we intend to address this issue by summarising the expressions of groups of genes through gene modules: groups of genes that tend to be co-expressed and functionally related — potentially working together to adapt to environmental stressors. By focusing on gene module behaviour instead of individual genes, we significantly reduce the overwhelming number of interactions that require analysis.

Gene module extraction methods benefit from large amounts of data, which can be created by merging many public datasets. Nevertheless, the extraction of gene modules from extensive biological data is not a trivial task because differences in experimental methodology can have a severe effect on data characteristics. In this thesis project, you will explore different (machine learning) methods [2]–[4] to integrate and combine many existing large-scale transcriptomics datasets, ultimately deriving gene modules from this comprehensive data pool [5]. Potentially, you will also work on biological interpretation of the modules you find [6], [7]. Taking into account your preferences, we will select the specific approaches to tackle this project. Eventually, your methodology might be used in ongoing research, where it will allow us to pinpoint key modules associated with resilience traits. Ultimately, this will aid the cultivation of more resilient crops and reduce failed crop yields due to climate change.

References

[1] 'CropXR | What we do', CropXR. https://cropxr.org/what-we-do/
[2] P.-H. Hsieh et al., 'Adjustment of spurious correlations in co-expression measurements from RNA-Sequencing data', Bioinformatics, 2023, doi: 10.1093/bioinformatics/btad610.
[3] S. M. Foltz et al., 'Cross-platform normalization enables machine learning model training on microarray and RNA-seq data simultaneously', Commun. Biol. 6(1), 2023, doi: 10.1038/s42003-023-04588-6.
[4] K. Tang et al., 'Rank-in: enabling integrative analysis across microarray and RNA-seq for cancer', Nucleic Acids Res. 49(17):e99, 2021, doi: 10.1093/nar/gkab554.
[5] W. Saelens et al., 'A comprehensive evaluation of module detection methods for gene expression data', Nat. Commun., 9(1), 2018, doi: 10.1038/s41467-018-03424-4.
[6] E. Segal et al., 'A module map showing conditional activity of expression modules in cancer', Nat. Genet. 36(10), 2004, doi: 10.1038/ng1434.
[7] W. Mao et al., 'Pathway-Level Information ExtractoR (PLIER) for gene expression data', Nat. Methods 16(7):607–610, 2019, doi: 10.1038/s41592-019-0456-1.

FERMO: Empowering Prioritization of Specialized Metabolites through an Intuitive Mass Spectrometry Metabolomics Dashboard

Supervisors	Mitja Zdouc, Justin van der Hooft
Type	Software engineering, method development
Requirements	Advanced Bioinformatics, Software Engineering or equivalent
Skills	Visualization, Python, Mass Spectrometry, Metabolomics
Timestamp	March, 2024

Description

Specialized metabolites are ubiquitous in nature, playing diverse roles such as communication, defense against predation, and offensive purposes. Many specialized metabolites have been adapted to serve as drugs and crop protection agents, and are therefore of high commercial importance. Liquid chromatography tandem mass spectrometry (LC-MS/MS) is often used to identify and annotate metabolites. However, the wealth of metabolites detected by LC-MS/MS makes it challenging to prioritize the most interesting metabolites for downstream analysis. Integrating additional phenotypic data can aid in this process.

To address these challenges and facilitate qualitative metabolomics data processing, we have developed FERMO [1], a user-friendly dashboard application. FERMO automatically combines LC-MS/MS data with phenotypic data and other metadata, enabling comprehensive analysis. FERMO annotates detected metabolites and correlates them with biological activity. Its aim is to make qualitative metabolomics data processing broadly accessible to the community.

We have released an online demo version of FERMO (https://fermo.bioinformatics.nl/) which has been received very positively by the community, e.g., with over 10,000 views on Twitter. Building on this success, we want to continue the development of FERMO. In this project, we want to further refine the data processing pipeline of FERMO to extend the types of data input and accuracy and reliability of metabolite annotation and correlation with biological activity.

References

[1] Zdouc, M. M. et al. FERMO: A dashboard for streamlined rationalized prioritization of molecular features from mass spectrometry data. bioRxiv (2022) doi:10.1101/2022.12.21.521422.

Using genome scale metabolic modelling to design a growth medium for unculturable plant pathogens

Supervisors	Chrats Melkonian, Florien Gorter, Jan van der Wolf, Marnix Medema
Type	Development and Simulation of Metabolic Models
Requirements	Programming in Python, Advanced Bioinformatics
Additional Skills	Biochemistry
Timestamp	January 2024

Description

The genus *Liberibacter*, which is part of the *Rhizobiaceae* family, contains multiple bacterial species that cause disease on a number of different host plants. *Candidatus Liberibacter (Ca. L.) asiaticus*, *Ca. L. africanus* and *Ca. L. americanus* are all associated with Huanglongbing (HLB) disease, also known as citrus greening disease. HLB causes a number of different symptoms, including the formation of small, bitter fruits. HLB has a major impact across the world; e.g. over the past few decades, the disease has cut orange tree production in the US in half. *Ca. L. solanacearum* is a closely related species but has a very different host range. It infects carrots and other *Apiaceae*, as well as potato. In potato it causes a disease called Zebra Chip. Potatoes infected with the bacterium develop “zebra stripes” when they are fried, which renders the crop unsellable. All the above species have one thing in common: they are unculturable [1]. This unculturability is probably the result of their host-associated lifestyle: the bacteria inhabit the living cells of the plant phloem and are transmitted from one plant to another via the haemolymph of psyllid insects. Over evolutionary time, they have lost several genetic elements that are required for independent growth, and they now have a very small genome of only around 1.2 Mb. Several previous studies have attempted to design a growth medium for the *Ca.L.* species, but none of these have yielded a medium suitable for the continuous propagation of the bacterium. Possibly, helper bacteria and specific physical as well as chemical conditions are required but reports are inconsistent.

In this project, you will develop a genome-scale metabolic model (GEM) for *Ca. L. solanacearum* and, time permitting, for *Ca. L. asiaticus*. The primary objective is to design an artificial medium for cultivating these organisms in the laboratory [2]. To achieve this, we will employ high-quality publicly available genomes and enhance their annotations using various tools and databases. The resulting biochemical reactions will form the preliminary metabolic model, which will undergo further refinement through manual curation and gap-filling to produce the finalized version. We will create several hypothetical growth media based on our knowledge of the metabolic profile of the organism's natural habitat. Subsequently, we will conduct simulations using flux balance analysis, utilizing tools like COBRA or alternative Python libraries, to identify potential compounds required for *Ca. L. solanacearum* growth. To visualize and analyze the associated reaction networks, we will utilize tools such as Escher and Cytoscape. In addition, we may integrate genomic information from putative helper bacteria to enhance our understanding of metabolic requirements. Furthermore, we will leverage insights from the closely related species, Liberibacter crescens, which can be cultured in the laboratory and was isolated from papaya, to provide valuable comparisons and guidance for this project. Your findings will be tested *in vivo* by researchers from WUR Biointeractions and Plant Health using *Ca. L. solanacearum*.

Key references

1. Fagen, J.R., Leonard, M.T., McCullough, C.M., et al., 2014. Comparative genomics of cultured and uncultured strains suggests genes essential for free-living growth of *Liberibacter*. PLOS ONE 9(1): e84469.
2. Tejera, N., Crossmann, L., Pearson, B., et al., 2020. Genome-scale metabolic model driven design of a defined medium for *Campylobacter jejuni* M1cam. Front. Microbiol. 11:1072.

Unraveling Microbial Dynamics in Spontaneous Wine Fermentation

Supervisors	Chrats Melkonian (WUR&UU) & Richard Notebaart (WUR)
Type	Metagenomics, Metabolomics, Multi-omics
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Bioinformatics, Genomics, Metabolism (Microbial ecology is a plus)
Timestamp	October 25, 2023

Microbes play a pivotal role in winemaking, contributing to the diverse flavors that define wine.[1] This project delves into the fascinating world of microbial interactions in winemaking, employing cutting-edge high-throughput sequencing techniques such as shotgun metagenomic sequencing alongside metabolomics during wine fermentation. Sensory data generated by a panel of professional judges will provide the final link from grape must to flavor through microbial activity, which alters the metabolic profiles. Our aim is to enhance our understanding of these microscopic players in the winemaking process.

Key Objectives:

- Microbial Diversity Exploration: Investigate how the composition of microbial communities in spontaneously fermented wine must vary based on the vineyard of origin, shedding light on the influence of terroir.
- Integration of Multi-Omics Data: Utilize metagenomic and metabolomic data collected during wine fermentation, combined with sensory data from the final wine bottles, to create a comprehensive understanding of the relationship between microbial dynamics and wine quality.
- Wineries/Cellar Microbiota Influence: Assess the impact of wineries/cellar microbiota on wine fermentation and aroma. Explore how the winery environment shapes the microbial profile and the resulting wine characteristics.

This study focuses on spontaneously fermented must samples sourced from seven distinct Riesling vineyards in Pfalz, Germany.[1] By combining metagenomic sequencing, we aim to gain a holistic understanding of the microbial intricacies at play during wine fermentation. Additionally, sensory data collected from the final wine bottles will provide valuable insights into the sensory aspects of wine quality.

This project aims to not only advance the field of wine microbiology but also holds practical implications for the winemaking industry. By unraveling microbial interactions, we can contribute to the production of wines with enhanced complexity and unique flavors.

This project offers an opportunity to delve into the world of wine microbiology, applying state-of-the-art sequencing technologies to unlock the secrets of microbial interactions in winemaking. You will sharpen your bioinformatics and statistics and participate in our group meetings and Journal club. Previous experience with coding is a requirement (R, Python and Linux).

[1] Sirén K, Mak SST, Melkonian C, Carøe C, Swiegers JH, Molenaar D, Fischer U and Gilbert MTP (2019) Taxonomic and Functional Characterization of the Microbial Community During Spontaneous *in vitro* Fermentation of Riesling Must. *Front. Microbiol.* 10:697. doi: 10.3389/fmicb.2019.00697

Small RNA-based communication in virus-vector-plant interactions and its possible impact on the spread of plant viral diseases

Supervisor	Anne Kupczok, Emilyn Matsumura (VIR)
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Transcriptomics
Organisms	Viruses, insects, plants
Timestamp	May 2023

Description

Plant viral diseases are a major threat to agriculture and its spread relies mostly on the transmission by insect vectors. A still unresolved question is: how is insect behavior affected by the interactions with virus-infected plants? Small RNAs (sRNAs) are important regulators in plant pathosystems. First, they can act as regulators of gene expression and antiviral response within the plant. Second, they can also travel to interacting organisms that are unrelated (such as pathogens) to target their messenger RNAs (mRNAs) via the RNA interference (RNAi) mechanism. This striking phenomenon is known as “trans-kingdom RNAi”. In plant viral diseases, virus infection is known for altering the composition and abundance of sRNAs in plant cells. These sRNAs might reach the insect vector and hence contribute to modulation of the insect transcriptome and behavior.

In this project, we will study the influence of virus infection on the sRNA-based interaction between plant hosts and their vectors. Small RNA and transcriptome sequencing datasets from plants and insects of three plant-vector-virus systems are available and will be analyzed for:

I) identifying the changes in expression profiles of plant sRNAs that are induced by different types of plant viruses, and

II) identifying the virus-regulated plant sRNAs that act as trans-kingdom regulators in insect vectors.

The outcomes of this project will give new insights on the complexity and dynamics of sRNA-based crosstalk within this tritrophic pathosystem.

References

Matsumura, E.E.; Kormelink, R. Small Talk: On the Possible Role of Trans-Kingdom Small RNAs during Plant–Virus–Vector Tritrophic Communication. Plants 2023, 12, 1411. https://doi.org/10.3390/plants12061411

Collin Hudzik, Yingnan Hou, Wenbo Ma, Michael J. Axtell, Exchange of Small Regulatory RNAs between Plants and Their Pests, Plant Physiology, Volume 182, Issue 1, January 2020, Pages 51–62, https://doi.org/10.1104/pp.19.00931

Huang CY, Wang H, Hu P, Hamby R, Jin H. Small RNAs - Big Players in Plant-Microbe Interactions. Cell Host Microbe. 2019 Aug 14;26(2):173-182. doi: 10.1016/j.chom.2019.07.021

Using large language models (LLMs) for QTL causal gene prioritisation

Supervisors	Harm Nijveen, Dick de Ridder
Type	Algorithm development
Requirements	Adv. Bioinformatics, Machine Learning, Deep Learning
Skills	Genomics, Programming, Statistics, Machine/Deep learning
Timestamp	March 13, 2024

Description

Linking traits to genes is an important problem in biological research and particularly relevant for plant breeding. QTL mapping which links traits to genomic regions, is a potentially powerful approach for this. However, QTL mapping has a resolution problem: it typically yields genomic regions with dozens if not hundreds of candidate causal genes. Prioritising the most likely causal genes can be done using biological knowledge [1], for instance from publicly available databases. But these databases contain only a fraction of the current biological knowledge. A large body of biological knowledge is contained in scientific publications. Until recently, extracting meaningful gene-trait interactions from literature was not very sophisticated, but recent developments in transformer based large language models (ChatGPT/GPT-4, LLaMA, PaLM) have opened up new opportunities.

In this project, you will explore the use of large language models for linking genes to traits, to prioritise causal genes in trait associated genomic regions (QTLs). This will involve selecting the most appropriate LLM architecture, fine-tuning the model with domain-specific training texts (literature, gene annotation, etc.) and evaluating the performance. Issues to address in the project include the tendency of LLMs to hallucinate (giving convincing but wrong answers) and the requirement of large computational resources. Our AraQTL web platform can provide the QTL regions for the prioritisation [2].

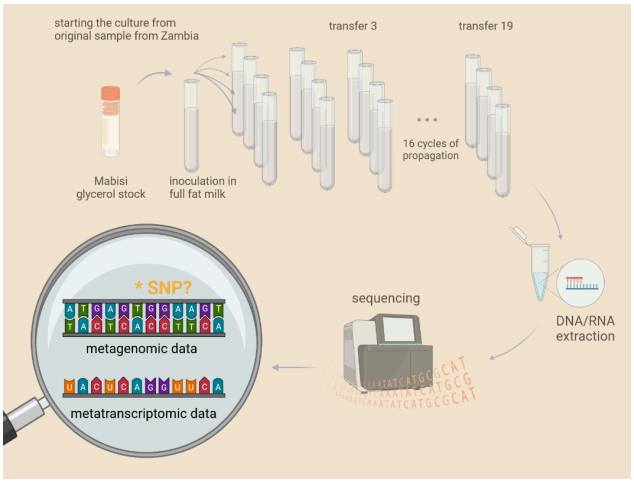
References

[1] Margi Hartanto, Asif Ahmed Sami, Dick de Ridder, Harm Nijveen (2022). Prioritizing candidate eQTL causal genes in Arabidopsis using RANDOM FORESTS. G3 Genes|Genomes|Genetics, Volume 12, Issue 11, https://doi.org/10.1093/g3journal/jkac255

[2] Harm Nijveen et al. (2017) AraQTL – workbench and archive for systems genetics in Arabidopsis thaliana. Plant J, 89: 1225–1235. doi:10.1111/tpj.13457

Eco-evolutionary dynamics of fermented milk communities

Supervisor	Anne Kupczok, Anna Alekseeva & Sijmen Schoustra (Genetics)
Type	Data analysis
Requirements	Advanced Bioinformatics, Programming in Python
Skills	Metagenomics, Transcriptomics, Comparative genomics
Organisms	Bacteria/archaea
Timestamp	September 2023, available April 2024



Description

All microorganisms interact with others in microbial communities (Pacheco and Segrè 2019). Such interactions are crucial for the functions of microbial communities, e.g., promoting or compromising host health in animal and plant microbiomes or – in industrial environments – biodegradation or fermentation in food production. Communities of fermenting microbes provide excellent model systems to study microbial interactions and community function (Smid and Lacroix 2013). For example, Mabisi, a traditional fermented milk product from Zambia, contains a relatively simple microbial community that is responsible for the production of aroma compounds specific for Mabisi. In the traditional Mabisi production, a part of a previous batch is used as a starter culture for a new batch (Groenenboom et al 2022). This resembles the well-known lab technique experimental evolution, where a single population is maintained over time under certain selection pressures (Lenski, 2017). In contrast, during propagation of communities such as Mabisi, the whole community of diverse members is maintained, allowing the study of genetic adaptations (evolution) as well as ecological effects (i.e., eco-evolutionary dynamics).

In our lab, we have obtained metagenomics data, metatranscriptomics data, and aroma profiles from Mabisi communities (Groenenboom et al 2022). In this project, we aim to focus on different Mabisi metagenomes (natural communities from Zambia, shortly propagated in the lab and long-term propagated in the lab). Specifically, we aim to (1) track variation within species that is associated with the propagation by analyzing single nucleotide polymorphisms (SNPs), (2) study the expression of the different variants, and (3) determine the metabolic pathways which show signatures of conservation or adaptation. This analysis will help to understand the impact of genetic changes during long-term co-adaptation of Mabisi communities.

References

Groenenboom AE, van den Heuvel J, Zwaan BJ, Smid EJ, Schoustra SE. 2022. Species dynamics in natural bacterial communities over multiple rounds of propagation. Evolutionary Applications 15:766–1775.

Lenski RE. 2017. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. The ISME journal, 11(10), 2181-2194

Pacheco AR, Segrè D. 2019. A multidimensional perspective on microbial interactions. FEMS Microbiol Lett 366:fnz125.

Smid EJ, Lacroix C. 2013. Microbe–microbe interactions in mixed culture food fermentations. Current Opinion in Biotechnology 24:148–154.

Detecting structural variation after horizontal chromosome transfer from low-coverage Nanopore data

Supervisor	Like Fokkens (PHP), Anne Kupczok
Type	Data analysis
Requirements	Advanced Bioinformatics, Programming in Python
Skills	Script programming, Biological sequence analysis
Organisms	Fungi
Timestamp	March 2024

Description

Fusarium oxysporum is a soil-dwelling fungus that can be harmless or even beneficial to a plant host, but can also cause wilting disease. The lifestyle of this fungus is at least partially dependent on the presence of specific accessory chromosomes that are enriched in genes involved in virulence and in transposable elements. Horizontal transfer of these chromosomes can turn a previously harmless strain into a pathogen.
To study the effect of receiving a transposon-rich chromosome on the rest of the genome, we performed two evolution/mutation accumulation experiments with strains that received one or two chromosomes, and of several natural strains that have received a transposon-rich chromosome some time in their past. One experiment is performed in planta: we iteratively infected tomato plants and reisolated the fungus, and one experiment is performed in vitro: we iteratively transferred small amounts of fungal material to new plates, where we used plates that were either rich or poor in glucose. We obtained low-coverage nanopore reads for 158 progeny strains.
The objective of the project is to infer structural variations (SVs), such as rearrangements, insertions and deletions, by mapping the long-reads to the ancestral genomes. We aim to answer the following questions:

1. Can we reliably detect SVs?
2. Are certain genomic regions enriched for SVs? Do SVs typically take place within a genomic compartment (i.e. within the conserved core or within the accessory genome), or is compartmentalization disrupted?
3. Do we find more SVs for ancestral strains that have more transposable elements?
4. Do we find more SVs in strains that only just received a transposon-rich chromosome compared to the strains that obtained that chromosome some time ago?
5. Do we find more, or different SVs in one or more experimental conditions (in planta, high-glucose, low-glucose)?

We will use existing software to detect structural variants (e.g. SENSV or NanoVar) for each of the progeny strains with respect to its ancestor. We will map these SVs back to the ancestral genome to determine whether these are clustered and whether these preferentially occur in specific genomic regions. We will compare the number of SVs between experimental conditions and between ancestral strains.

References

Leung, H.C.M., et al. Detecting structural variations with precise breakpoints using low-depth WGS data from a single oxford nanopore MinION flowcell. Sci Rep 12, 4519. https://doi.org/10.1038/s41598-022-08576-4 (2022)

Lorrain C, et al. Dynamics of transposable elements in recently diverged fungal pathogens: lineage-specific transposable element content and efficiency of genome defenses. G3 (2021)

Simone Fouché, et al., A devil's bargain with transposable elements in plant pathogens, Trends in Genetics, (2022)

Tham, C. Y. et al. NanoVar: accurate characterization of patients' genomic structural variants using low-depth nanopore sequencing. Genome Biol. 21, 56. https://doi.org/10.1186/s13059-020-01968-7 (2020).

Torres D.E. et al., Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biology Reviews (2021)

Vanheule A, et al.. Living apart together: crosstalk between the core and supernumerary genomes in a fungal plant pathogen. BMC Genomics. (2016)

Computational Metabolomics Tool Development for Structure Annotation

Supervisor	Justin van der Hooft
Type	Workflow development and/or Data Analysis
Requirements	Advanced Bioinformatics (BIF30806)
Skills	Programming, Metabolomics, Mass spectrometry, Machine learning
Organisms	Bacteria/Archaea, Fungi, possibly various other sources
Timestamp	March 2024

Description

Metabolomics has been coined as the ultimate phenotyping tool. Whilst this could be the topic of an interesting debate, it has become clear that analytical chemistry-based measurements of complex metabolite mixtures yield ever complex and information-dense profiles. Often, this results in liquid chromatography tandem mass spectrometry (LC-MS/MS) datasets that contain information about the molecules in the sample. However, the structural annotation of these molecules is not a trivial task: only a fraction of known molecules can be annotated by comparison against databases, and de-novo annotation is still in its infancy. In the Bioinformatics Group, we are working on novel computational metabolomics approaches based on networking and machine learning [1,2] to break the barriers between researchers and their data at one end, and chemical and functional insights at the other end: i) the organization of large-scale datasets, ii) the prioritization of relevant spectral data, and iii) the structural and functional annotation of the metabolites in a mixture. Recent examples of tools that were developed within the Bioinformatics Group that contribute to breaking these barriers are MS2Query [3], FERMO [4], and Spec2Vec [5].

If you are as excited as we are to contribute to this field, please do get in touch and we can discuss the options that are available at that time. Both mass spectral networking based as well as machine learning based approaches are currently under much development both here in Wageningen as well as around the globe, and you would be welcome to contribute to these developments. There is also room to bring in your own ideas and creativity to support the development and application of computational metabolomics workflows.

References

[1] Bennidir et al., Natural Products Reports (2021):
[2] de Jonge et al., Metabolomics (2022):
[3] de Jonge et al, Nature Communications (2023):
[4] Zdouc, Mitja M., et al. *bioRxiv* (2022): 2022-12.
[5] Huber, Florian, et al. *PLoS computational biology* 17.2 (2021): e1008724.

BiG-MAP v2: an automated pipeline to profile gene cluster abundance and expression in microbiomes

Supervisor	Hannah Augustijn, Robert Koetsier, Marnix Medema
Type	Workflow development
Requirements	Advanced Bioinformatics (BIF30806)
Skills	Comparative genomics, Metagenomics, Metatranscriptomics
Organisms	Bacteria/Archaea, Humans
Timestamp	March 2024

Description

Microbes play an increasingly recognized role in determining their host's health. For example, studies on diseases like Alzheimer's and type 2 diabetes found that small molecules produced by microbes play a role in the progression of these conditions[1,2]. These molecules are often synthesized by genes that are grouped on the genome in biosynthetic gene clusters (BGCs). As the availability of metagenomic and metatranscriptomic datasets continues to grow, there is a need for tools that can automate the analysis of these BGCs across omics data sets. To address this need, we have developed a mapping pipeline called BiG-MAP (Biosynthetic Gene cluster Meta'omics Abundance Profiler)[3], which enables the assessment of gene cluster abundance and expression in microbiome samples using metagenomic and metatranscriptomic data. Unlike existing functional annotation and pathway inference tools, BiG-MAP does not rely on reference databases. Instead, it utilizes gene clusters predicted by antiSMASH[4] or gutSMASH[5] and sets of reference genomes. This distinction sets BiG-MAP apart and enhances its capabilities in analyzing and understanding the complex relationships between BGCs (both of known and of unknown function) and host-associated phenotypes.

In this project, you will further expand and optimize the functionalities of BiG-MAP to enhance multi-omic data exploration. This will involve optimizing the analysis module for metatranscriptomics data, which currently visualizes housekeeping genes for manual interpretation of expression by the user. However, housekeeping genes are often undetected and cannot be visualized. Therefore, you will explore better methods for establishing an expression baseline and performing data normalization. Additionally, we aim to streamline the various separate modules that make up BiG-MAP by combining these modules into one user-friendly command-line package. With these expanded functions, you will analyze the abundance and expression of BGCs in relevant metagenomic or metatranscriptomic datasets derived from human microbiomes.

Key references

1. Jiang, C., Li, G., Huang, P., Liu, Z. & Zhao, B. The Gut Microbiota and Alzheimer's Disease. *J. Alzheimers. Dis.* **58**, 1–15 (2017).
2. Koh, A. *et al.* Microbially Produced Imidazole Propionate Impairs Insulin Signaling through mTORC1. *Cell* **175**, 947–961.e17 (2018).
3. Pascal Andreu, V. *et al.* BiG-MAP: an Automated Pipeline To Profile Metabolic Gene Cluster Abundance and Expression in Microbiomes. *mSystems* **6**, e0093721 (2021).
4. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* (2023) doi:10.1093/nar/gkad344.
5. Pascal Andreu, V. *et al.* gutSMASH predicts specialized primary metabolic pathways from the human gut microbiota. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01675-1.

Contextualized Spectral Pattern Annotation using Computational Metabolomics Network-based Approaches

Supervisor	Justin van der Hooft
Type	Data Analysis & some Workflow development
Requirements	Advanced Bioinformatics (BIF30806)
Skills	Programming, Metabolomics, Mass spectrometry, Machine learning
Organisms	Bacteria/Archaea, Fungi, possibly various other sources
Timestamp	March 2024

Description

Metabolomics is the research discipline that can provide detailed phenotypic information about natural systems to gain insight into active metabolic pathways and metabolic differences related to the biological question. As it is impossible to directly see metabolites, biological samples are extracted, and these extracts are then typically measured with mass spectrometry. Such a workflow generates information-rich metabolomics profiles. However, many researchers perceive barriers to get from such metabolomics profiles to structure and function information. Recent computational metabolomics approaches based on mass spectral networking and machine learning have increased our understanding of metabolite mixtures [1,2,3]. In these approaches, the hypothesis that similar molecules generate similar mass spectra is key [4]. Furthermore, it could be deducted that, as a logical consequence, similar substructures, or scaffolds (i.e., part of a molecule), do result in similar mass features in mass spectra. Indeed, it was demonstrated that substructure patterns representing biochemical building blocks can be mined from metabolomics profiles [5].

In this project, you will explore if a combination of two existing computational metabolomics strategies, i.e., MS2LDA (un)supervised substructure finding [5] and SNAP-MS [6] (building onto GNPS molecular networking [7]) can result in improved annotation of substructure patterns. The suggested workflow could work due to the concept of molecular families that group mass spectra generated of similar molecules; and hence, these molecular families are expected to be enriched for specific elemental formulas (of molecular analogues) [this is the concept behind SNAP-MS] as well as enriched for the presence of specific substructure patterns. It may then yield improved annotations by linking the enriched elemental formula annotations from SNAP-MS to the substructure patterns discovered by MS2LDA.

You will take public mass spectrometry data run these through Molecular Networking, MS2LDA substructure finding, and SNAP-MS. Then you will link the various types of information: i.e., enriched substructures from annotated candidate structures (i.e., through graph-based mining), with presence/absence of substructure patterns and their corresponding mass fragment & neutral loss features, and the occurring molecular formulas. Then, you will designand test contextualized (Mass2Motif) molecular formula assignment and Mass2Motif-mass fragment and neutral loss annotation. If there is time left, the critical component of mass spectral groupings could be explored: what if these groupings are based on different parameters or a different mass spectral similarity score, i.e., machine learning based versus mass fragmental overlap based [2].

References

[1] Bennidir et al., Natural Products Reports (2021)
[2] de Jonge et al., Metabolomics (2022)
[3] de Jonge et al, Nature Communications (2023)
[4] Huber, Florian, et al. PLoS computational biology (2021)
[5] van der Hooft et al., PNAS (2016)
[6] Morehouse et al. Nature Communications (2023)
[7] Wang et al., Nature Biotechnology (2016)

Phylogeny of Acidobacteria originating from marine sponges and beyond: identifying their genetic and biosynthetic diversity

Supervisors	Pavlo Hrab, Michelle Schorn, Detmer Sipkema, Marnix Medema
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Programming, Data analysis
Timestamp	March 2024

Description

For decades, Actinomycetota remained the main phylum of interest for antibiotic discovery. However, despite great biosynthetic potential, high rediscovery rates made the focus shift to more understudied species [1]. Furthermore, even greater diversity is just about to be brought into the lab, as many microbes remain yet to be cultivated. In that sense, cultivation-free techniques, such as metagenomics, serve a helpful purpose to evaluate how to culture these as-yet-uncultured microbes.

As the amount of data has increased, we have observed that bacteria from the phylum Acidobacteriota are both broadly distributed in the environment and their genomes show a large potential to produce natural products. Since only a handful of reports of their successful cultivation exists, this phylum represents a low-hanging fruit in uncovering natural products' unseen chemistry. Interestingly, we have identified several lineages of Acidobacteriota in multiple marine sponges, harboring unique biosynthetic potential.

But, despite this potential of Acidobacteriota, no whole-genome phylogenetic studies have been conducted. As a future guidance, the field will greatly benefit from linking groups of organisms with their Biosynthetic Gene Clusters (BGCs). This will answer the long-standing question of which microbes to pursue for isolation efforts [2]. Moreover, placing genomes on a phylogenetic tree should help in defining their lifestyle as well [3], providing a reference for future studies.

The aim of this project is to construct a global Acidobacteriota phylogeny and annotate it with biosynthetic diversity, using both in-house and publicly available acidobacterial Metagenome-Assembled genomes (MAGs). Hence, the first challenge is to find and parse the data from different genome repositories and define a set of markers to construct the tree on. Then, after annotating BGCs, you will assess whether any associations exist between BGC composition/amount/novelty and environmental metadata. Based on this, you will construct a short summary of phyllum *Acidobacteriota* as a potential source of natural products.

References

[1] M. I. Hutchings, A. W. Truman, and B. Wilkinson, "Antibiotics: past, present and future," *Curr. Opin. Microbiol.*, vol. 51, pp. 72–80, Oct. 2019, doi: 10.1016/j.mib.2019.10.008.

[2] A. Gavriilidou *et al.*, "Compendium of secondary metabolite biosynthetic diversity encoded in bacterial genomes," Genomics, preprint, Aug. 2021. doi: 10.1101/2021.08.11.455920.

[3] J. Sikorski *et al.*, "The Evolution of Ecological Diversity in Acidobacteria," *Front. Microbiol.*, vol. 13, 2022, Accessed: Jun. 28, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fmicb.2022.715637

Chemical warfare in attack and defense of *Fusarium oxysporum*

Supervisors Like Fokkens (Phytopathology), Marnix Medema
Type Data analysis
Requirements Programming in Python, Advanced Bioinformatics
Skills Genomics, programming
Timestamp March 2024

Description

The *Fusarium oxysporum* (*FOSC*) species complex contains many soil fungi that are pathogenic towards different hosts. Its genome consists of 11 core chromosomes that are shared by almost all strains and one or more accessory chromosomes that are lineage specific, or shared by isolates that infect the same host (host-specific accessory chromosomes or pathogenicity chromosomes). The species produces many different secondary metabolites, most of which presumably serve as toxins to weaken its host and compete with or defend against other microbes in the soil. Genes that encode the proteins involved in the metabolic pathways that produce these secondary metabolites, are often clustered on the genome in secondary metabolite gene clusters. We hypothesize that gene clusters that encode for secondary metabolites involved in infection are mostly located on pathogenicity chromosomes, while gene clusters that encode secondary metabolites that are involved in microbe-microbe interactions are mostly located on core chromosomes or on lineage-specific accessory chromosomes.
In this project we will use existing advanced bioinformatic tools to identify these metabolic gene clusters in the pangenome of ~600 publicly available FOSC genome sequences. We will study and compare the following characteristics:

- Whether/which secondary metabolite gene clusters occur on core chromosomes, lineage-specific accessory chromosomes or pathogenicity chromosomes. *This may give us an indication on whether these metabolites are typically used in the context of infection or not.*
- Presence-absence polymorphisms within the species complex (and/or within Fusaria) and deletion, loss and horizontal transfer events along the phylogenetic tree. *Which metabolite gene clusters are often horizontally transferred? For which gene clusters do we observe partial losses/gains in evolution?*

This project will be the first comprehensive survey of secondary metabolite gene clusters and their location on the genome in this important pathogen. This will provide a strong basis for future studies on metabolite-mediated host-microbe and microbe-microbe interactions. Based on our findings, we can choose to expand our research and identify genes/clusters under positive/diversifying selection, study the occurrence patterns of gene clusters in other Fusarium species, and compare expression levels within and between clusters in different conditions using public RNA-seq datasets.

References

Wisecaver JH and Rokas A (2015) Fungal metabolic gene clusters—caravans traveling across genomes and environments. *Front. Microbiol*. doi: 10.3389/fmicb.2015.00161

Hoogendoorn K, Barra L, Waalwijk C, Dickschat JS, van der Lee TAJ and Medema MH (2018) Evolution and Diversity of Biosynthetic Gene Clusters in *Fusarium*. *Front. Microbiol*. doi:10.3389/fmicb.2018.01158

Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W. *et al.* (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat Chem Biol*. doi: 10.1038/s41589-019-0400-9

A phylogenetic framework for linking genes to molecules in large-scale genomic/metabolomic datasets

Supervisors	Marnix Medema, Justin van der Hooft
Type	Data analysis & Method/algorithm development
Requirements	Programming in Python, Advanced Bioinformatics
Skills	Genomics, programming, chemical biology
Timestamp	March 2024

Description

Specialized metabolites produced by microbes, fungi, and plants are used for various applications like antibiotics and anticancer drugs. The genes encoding the enzyme ensembles that produce those specialized molecules (a.k.a. natural products) are often physically clustered together in biosynthetic gene clusters (BGCs). Algorithms have been developed that can predict the presence of such BGCs in whole genome sequences [1]. These tools yield large numbers of putative BGCs with some known but predominantly yet unknown molecular products - of which the structural prediction remains difficult. Experimental validation of links between BGCs and the products which they encode has also led to the discovery of several subclusters: modules of co-evolving genes that are responsible for the production of specific molecular substructures across otherwise structurally diverse molecules [2].

Several methods have been devised that make use of large genomic and metabolomic datasets to identify, by means of correlation, which BGCs are responsible for the production of which molecules, either by correlating absence/presence patterns of full BGCs across strains with those of full molecules [3, 4] or by correlating absence/presence patterns of subclusters with substructures. However, these approaches suffer from the problem that many shared absence/presence patterns are due to the phylogenetic relatedness of the underlying strains (co-inheritance) rather than independent parallel acquisition of the same BGC. Hence, a species-phylogenetic framework is required to correct for (or at least interpret/adjust) such correlation metrics.

The main goal of this project is to develop an improved, phylogenetically aware, way of identifying how ‘special’ the co-occurrence of molecules and BGCs across strains is, in order to judge their statistical significance. Given the large datasets that are already available to apply this on, the expectation is that this will make it possible to link BGCs to molecules in high throughput and generate a concrete reduced list of candidates coming from correlation analyses that can be tested in the laboratory by collaborators. Thus, this will accelerate the discovery of natural products such as antibiotics and anticancer agents.

Key references

1. Blin et al., Nucleic Acids Research, 2017. 45(W1): p. W36-W41.
2. Del Carratore et al., Communications Biology, 2019, 2: 83.
3. Doroghazi et al., Nature Chemical Biology, 2014, 10: 963-968.
4. Navarro-Muñoz et al., Nature Chemical Biology, in press.

Integrative QTL analysis

Supervisors	Harm Nijveen, Dick de Ridder
Type	Algorithm development
Requirements	Adv. Bioinformatics, Adv. Statistics / Modern Statistics for the Life Sciences
Skills	Genomics, Programming, Statistics, Machine learning
Timestamp	March 13, 2024

Description

In a recent national study on maternal effects on seed quality, 165 homozygous recombinant lines of *Arabidopsis thaliana* grouped in a number of different growth conditions were genotyped based on 1059 markers and transcript levels were measured. These lines were also extensively phenotyped, with the goal of performing generalized genetical genomics [1] – correlating genotype with phenotype (expression) under a range of conditions. Levels of a number of primary metabolites were measured as well.

In this project, the goal is to develop methods to learn which genes influence which genotype, extending the QTL approach by incorporating expression and metabolic pathway information [2]. Prior knowledge on metabolic regulation and the relation between condition and metabolic activation can be used to refine the search and zoom in on possible mechanistic explanations of the observed phenotypes. The desired outcome is a method to optimally combine genetical genomics data with prior knowledge.

References

[1] Y. Li *et al*. (2008) Generalizing genetical genomics: getting added value from environmental perturbation. *Trends Genetics* 24(10):518-24.

[2] R.C. Jansen *et al.* (2009) Defining gene and QTL networks. *Current Opinion in Plant Biology* 2009, 12:1–6.

[3] Nijveen, H. *et al.* (2017) AraQTL – workbench and archive for systems genetics in *Arabidopsis thaliana*. Plant J, 89: 1225–1235. doi:10.1111/tpj.13457

Finding genes related to regeneration

Supervisors	Harm Nijveen, Renze Heidstra
Type	Data analysis
Requirements	Advanced Bioinformatics, Adv. statistics/Modern Statistics for the Life Sciences
Skills	Genomics, Programming, Statistics, Machine learning
Timestamp	April 3, 2024

Description

Introduction of transgenes in plants has been a force in molecular biology and biotechnology for decades. *Agrobacterium tumefaciens* is generally used as a vehicle to introduce genetic material in the plant cell. Unfortunately, not all plants (particularly agronomically important crops) have the ability to regenerate a complete plant from a single (transgenic) cell.
Therefore, identification and knowledge on the molecular factors involved in the regeneration process is required[1]. To gain more information on the regeneration process, an RNA-sequencing experiment was conducted using *Arabidopsis thaliana* regenerating tissue at multiple time-points up to the fully regenerated shoot.

In this project the goal is to uncover candidates whose function is currently not associated with the regeneration process, as well as characterize the differential expression of genes suspected to be involved in regeneration through time. This requires the analysis of the RNAseq dataset individually, comparing expression data from different time-points, and to existing datasets. Genes/transcripts can be clustered based on co-expression [2,3] measures to find genes that show expression patterns similar to regeneration related genes. Clustered genes can then be analysed for enriched Gene Ontology annotations or common transcriptional regulators[4] to learn more about the underlying regulatory mechanisms.

References

1. Radhakrishnan D, Kareem A, Durgaprasad K, Sreeraj E, Sugimoto K, Prasad K: Shoot regeneration: a journey from acquisition of competence to completion. *Current Opinion in Plant Biology* 2018, 41:23-31.
2. Langfelder P, Horvath S: WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008, 9:559.
3. Serin EAR, Nijveen H, Hilhorst HWM, Ligterink W: Learning from Co-expression Networks: Possibilities and Challenges. *Frontiers in Plant Science* 2016, 7:444.
4. Kulkarni SR, Vaneechoutte D, Van de Velde J, Vandepoele K: TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information. *Nucleic Acids Research* 2018, 46:e31-e31.

Prediction in pangenome graphs

Supervisors	Dick de Ridder
Type	Algorithm development
Requirements	Machine learning, Algorithms in Bioinformatics
Skills	Programming, Machine learning
Timestamp	April 1, 2023

Description

In recent years, the number of genomes has grown rapidly. Many species are no longer represented by a single reference genome but by numerous related genomes. To capitalize on the genomic diversity in large collections of genomes, we need to transition from a reference-centric approach to a pangenome approach. Computational pangenomics is currently a hot topic and a challenging field of research [1]. We have developed a pangenome solution called PanTools which compresses multiple annotated sequences into a single graph representation, constructed, stored, and annotated in a Neo4j graph database [2].

There are numerous applications for such a pangenome representation at different levels, containing sequences, variants, genes, expression levels, orthologous groups, functional annotations, phenotypes etc. of a set of genomes. Whereas traditional studies mostly relate variation at one individual level to a phenotype, for example SNPs in GWAS or expression differences in transcriptomics studies, a particularly intriguing application of a pangenome would be to combine various levels of annotation to predict such a phenotype. This could help learn whether nonlinear combinations of annotations predict a phenotype; for example, whether a certain SNP in a gene with a certain function is predictive of the phenotype only in genomes in which it is expressed. A long term goal of such an approach would be to use machine learning to infer novel relations and add these as computationally derived edges to the graph.

In this project, we will further develop a prototype machine learning pipeline based on PanTools. A basic framework is available to extract sets of relevant features from a pangenome graph, to predict phenotypes and functions in a machine learning setting. Neo4j offers Cypher, a query language to formulate complex graph structure and property queries; extracted features are then converted to training/test sets for a machine learning application in Python or R. This approach has been developed and validated on a set of bacterial genomes. Open questions are what gene function prediction performance is possible on larger, less well annotated genomes (plants), what computational bottlenecks arise in queries (and whether the underlying datastructure limits this), how predicted annotations should best be added to the pangenome graph and whether prediction “on the fly” when querying can be implemented.

References

[1] Computational Pan-Genomics Consortium (2018) Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics* 19(1):118-135.

[2] Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics* 32(17):i487-i494.

DeepVariant for non-human species

Supervisors	Dick de Ridder, Richard Finkers
Type	Algorithm development
Requirements	Programming in Python, Machine Learning, Deep Learning
Skills	Programming, machine learning
Timestamp	November 6, 2023

DeepVariant is a deep learning-based variant caller that takes aligned reads (in BAM or CRAM format), produces pileup image tensors from them, classifies each tensor using a convolutional neural network, and finally reports the results in a standard VCF or gVCF file. DeepTrio [1] is built on top of DeepVariant. It is intended for variant calling of trios or duos. The models included with DeepVariant are only trained on human data. For other organisms, see the blog post on non-human variant-calling [2] for some possible pitfalls and how to handle them.

The focus of this thesis will be to fine-tune and/or retrain DeepVariant on plant species [3] and investigate the performance of DeepVariant and DeepTrio in terms of speed and accuracy, and compare these to widely used approaches, such as freebayes [4] and GATK [5]. If necessary, simulated data can help to learn what determines performance/speed of various algorithms. Subsequently, the performance of DeepVariant can be assessed in plant species for which it did not undergo specific training or on low-depth population re-sequencing data, in order to determine its robustness.

References

[1] https://github.com/google/deepvariant/blob/r1.5/docs/deeptrio-details.md
[2] https://google.github.io/deepvariant/posts/2018-12-05-improved-non-human-variant-calling-using-species-specific-deepvariant-models/
[3] https://cloud.google.com/blog/products/data-analytics/analyzing-3024-rice-genomes-characterized-by-deepvariant
[4] https://github.com/freebayes/freebayes
[5] https://gatk.broadinstitute.org/hc/en-us

Sequencing-based high density marker development

Supervisor	Dick de Ridder, Richard Finkers
Type	Workflow development
Requirements	Advanced Bioinformatics, Machine Learning
Skills	Comparative genomics, programming
Timestamp	April 1, 2023

Next generation sequencing technologies, such as Illumina, have become cheap enough that large experimental plant breeding populations can now be screened via whole genome sequencing (WGS) approaches. As breeding populations have a degree of structure (e.g., the number of different alleles expected in the population is limited), individuals do not have to be sequenced at large depth, as intermediate positions are linked on the same chromosome and can be imputed. However, short read lengths hamper the detection of which markers share a genome and strategies using reference genomes can introduce noise (e.g., because of repetitive sequences in the genome). However, presence/absence signals of markers in the individuals can be used to cluster alleles into their respective chromosomes, constrained by knowledge on the parental genomes and the possible forms of recombination.

In this project, we will assess strategies to efficiently analyze WGS data of an autotetraploid breeding population. Such populations are particularly challenging as variants can occur on one or more chromosomes. We will investigate different strategies to call variants (e.g., GATK vs. K-mer-based approaches), and (advanced) clustering techniques to assign groups of variants to their respective homologous chromosomes. The optimal strategy should be implemented into an easy-to-use workflow to be used for, for example, interactive scaffolding of assemblies into pseudomolecules or performing genetic analyses such as QTL mapping.

Predicting local genome similarity with genomic DL foundation models

Supervisor	Dick de Ridder
Type	Method development
Requirements	Deep Learning
Skills	Deep learning, machine learning, genomics
Timestamp	April 19, 2024

To make optimal use of the genomic diversity found in large collections of genomes, there is a transition ongoing from a reference-centric approach to one based on pangenomes: computational representations of multiple genomes that facilitate fast analyses. Computational pangenomics is currently a hot topic and a challenging field of research [1]. We have developed a pangenome solution called PanTools, that compresses multiple annotated sequences into a single graph data structure, constructed, stored, and annotated in a Neo4j graph database [2]. The representation stores genome sequences (as so-called compressed, colored De Bruijn Graphs) and links genomes by detecting which genes are similar, indicating they may have the same function. However, no such similarity measures are yet available for other regions in the genome other than from methods based on costly full whole-genome alignments.

An exciting trend in the analysis of sequence data has been the development of Transformer-based models that are pre-trained on large data volumes, which can subsequently be fine tuned to perform a range of specific tasks with relatively minor effort. A well-known example is BERT, which has had a major impact in natural language tasks. Inspired by BERT, similar approaches have been developed on genome sequence data, in particular with the DNABERT approach [3] (and similar, such as HyenaDNA). This model showed good performance in various specific prediction tasks related to genome annotation (for example, predicting where proteins can bind the DNA), after minor fine-tuning for those tasks.

In this project you will explore how we can exploit the DNABERT model to predict local genome similarity in plants. The available DNABERT model has been developed for the human genome; a first challenge is to see whether it can be readily applied or should be fine-tuned/retrained on plant genomes. Second, we want to extend the current DNABERT architecture to predict similarity of genomic regions, for example using a Siamese neural network; an interesting question is how sensitive this will be to the evolutionary distance between two genomes. Finally, we can try to add higher level information on the contents of the genome (genes, functions, pathways) to arrive at a measure that combines both sequence and functional similarity.

References

1. Computational Pan-Genomics Consortium. 2018. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics* 19(1):118-135.
2. Sheikhizadeh S, Schranz ME, Akdel M, de Ridder D, Smit S. 2016. PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics* 32(17):i487-i494.
3. Ji Y, Zhou Z, Liu H, Davuluri RV. 2021. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics* 37(15):2112-2120. doi: 10.1093/bioinformatics/btab083.

Purging polyploid plant genome assemblies

Supervisor	Dick de Ridder, Richard Finkers
Type	Algorithm development
Requirements	Advanced Bioinformatics
Skills	Programming
Timestamp	November 6, 2023

Description

Many plant species, such as economically interesting ornamental species such as phalaenopsis, chrysanthemum, etc. are polyploid. However, many downstream analyses (excluding pangenomics) still rely on a single haploid representation to investigate sequence variation. For diploids, strategies have been developed to purge the two genomes into a single haploid representation. Examples of these are purge_dups [1] and purge_haplotigs [2]. While these tools do not always work perfectly even in diploids, utilization in polyploids is even more challenging.

In this project, we would like to develop and utilize a novel purging strategy and apply this strategy on sequence contigs from a tetraploid ornamental species. Such a strategy will involve purging unique variation, the choice of which is normally one of the most difficult challenges in this type of algorithm. Ideally, a decision tree/ report is being retained showing which choices were made and assesses their impact. Quality of the purged assembly will be evaluated with independent datasets (e.g. segregation of markers developed in the purged assembly in a segregating population; incomplete purging usually leads to development of poorly performing markers).

References

[1] Dengfeng Guan, Shane A. McCarthy, Jonathan Wood, Kerstin Howe, Yadong Wang, Richard Durbin. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* 36(9):2896-2898, 2020.

[2] Michael J. Roach, Simon A. Schmidt, Anthony R. Borneman. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* 19:460, 2018.

Using deep learning for end-to-end optimisation of plant resilience models

Supervisors	Ben Noordijk, Dick de Ridder
Type	Method development
Requirements	Deep Learning
Skills	Deep learning, Transcriptomics, Machine learning, Programming, Python
Organism	Plants
Timestamp	April 2024

Climate change poses significant challenges to global food security due to its negative impact on crop growth and productivity. With a growing world population, especially in climate-vulnerable areas, it becomes essential to develop crop varieties that are resilient to combinatorial stresses such as heat and drought. However, our current understanding of plant responses to combinations of stresses is limited.

A plant's environmental response is governed by gene regulatory networks (GRNs); sets of genes which influence each other's expression. Stress-response GRNs typically involve a vast number of genes and an even larger number of regulatory interactions, making detailed modelling of the entire network impractical. In this research (part of the PlantXR consortium [1]), we intend to address this issue by summarising the expressions of groups of genes through gene modules: groups of genes that tend to be co-expressed and functionally related — potentially working together to adapt to environmental stressors.

To achieve this, we are developing a method called BADDADAN — a Bioinformatics Approach to Describe Dynamical Activations of Dimension-reduced A-priori-informed Network. It allows simulation of plant gene modules in response to external stresses. Currently it is comprised of various distinct steps, each of which is optimised individually. Plausibly, its performance can be boosted if all steps in the process are optimised simultaneously in an end-to-end procedure [2]. This can be done by designing a custom-made loss-function, which also opens up the possibility to make the modules adhere to prior biological knowledge. For this, deep-learning approaches such as Lagergren et al. [3] might be useful, too. Taking into account your preferences, we will select the specific approaches to tackle this project. Eventually, your newly developed end-to-end methodology might be used in ongoing research, where it will allow us to model plant response to stress. Ultimately, this will aid the cultivation of more resilient crops and reduce failed crop yields due to climate change.

References

[1] 'CropXR | What we do', CropXR. Accessed: Aug. 25, 2023. https://cropxr.org/what-we-do/

[2] M. AlQuraishi and P. K. Sorger, 'Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms', Nat. Methods, vol. 18, no. 10, pp. 1169–1180, Oct. 2021, doi: 10.1038/s41592-021-01283-4.

[3] J. H. Lagergren, J. T. Nardini, R. E. Baker, M. J. Simpson, and K. B. Flores, 'Biologically-informed neural networks guide mechanistic modeling from sparse experimental data', PLOS Comput. Biol., vol. 16, no. 12, p. e1008462, Dec. 2020, doi: 10.1371/journal.pcbi.1008462.

Developing strategies to enhance gene cluster family assignment algorithms

Supervisors	Jorge Navarro, Marnix Medema, Justin van der Hooft
Type	Method/algorithm development
Requirements	Programming in Python; Advanced Bioinformatics;
Skills	Programming;
Timestamp	March 2024

Description

Biosynthetic gene clusters (BGCs) are groups of genes that code for enzymatic assembly lines responsible for the production of natural products. These natural products are small molecules that are often bioactive, which makes the natural product chemical space a lucrative search space for drug discovery. For example, microorganisms produce antibiotic natural products to combat competing organisms. Natural product classes include polyketides (PKs), nonribosomal peptides (NRPs), and ribosomally synthesized and post-translationally modified peptides (RiPPs). Regions containing candidate BGCs can readily be identified in genomic sequencing data using tools like antiSMASH [1], but often include genes unrelated to natural product biosynthesis.

The recent availability of large and complex (meta)genomics datasets has allowed the generation of a vast amount of BGCs from all biomes. In order to efficiently explore the diversity contained in these datasets, the BiG-SCAPE software has been developed. The biosynthetic gene similarity clustering and prospecting engine (BiG-SCAPE [2]) is a computational framework designed to compare biosynthetic gene clusters (BGCs), via command-line generation of sequence similarity metrics and networks which can then be interactively explored in a locally hosted GUI. It features a comprehensive algorithm that handles BGC complexity by accounting for differences in modes of evolution between BGC classes, grouping BGCs at multiple hierarchical levels, and alignment modes that account for BGC fragmentation. BGCs are grouped in gene cluster families (GCFs) based on user input similarity cutoffs, and the framework further allows for the elucidation of phylogenetic relationships within these families.

While BiG-SCAPE has already been used to successfully explore highly diverse sets of BGCs, it suffers from limitations regarding its ability to discern between the core regions of a BGC/GCF, and adjacent regions which do not contribute to the synthesis of a natural product. These adjacent regions thus have a higher impact in GCF generation than what is desirable. Here, you will update and expand the software by designing new strategies to enhance the algorithms in BiG-SCAPE that drive BGC comparison, with a focus on core biosynthetic domains. You will be contributing to the design of a training dataset, as well as a selection of features that can be used to aid in distinguishing biosynthetic from non-biosynthetic regions. Subsequently, you will use BiG-SCAPE to benchmark the developed features using a comprehensive set of (meta)genomic data, and evaluate its performance and accuracy.

In order to work on this project you will have to be comfortable with contributing to larger software programs in Python and you will need to have a good understanding of BGC evolution and diversity.

Key references

1. Kai Blin, Simon Shaw, Alexander M Kloosterman, Zach Charlop-Powers, Gilles P van Wezel, Marnix H Medema, Tilmann Weber, antiSMASH 6.0: improving cluster detection and comparison capabilities, Nucleic Acids Research, Volume 49, Issue W1, 2 July 2021, Pages W29–W35, https://doi.org/10.1093/nar/gkab335
2. Navarro-Muñoz, J.C., Selem-Mojica, N., Mullowney, M.W. et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Biol 16, 60–68 (2020). https://doi.org/10.1038/s41589-019-0400-9

Exploring the origin and evolution of desiccation tolerance in plants using comparative transcriptomics.

Supervisors	Harm Nijveen, Asif Ahmed Sami, Mariana Silva Artur
Type	Data analysis
Requirements	Adv. Bioinformatics, Plant Plasticity and Adaptation (optional)
Skills	Comparative genomics, RNA-seq analysis
Timestamp	March 13, 2024

Description:

Desiccation tolerance (DT) refers to the capacity of cells or tissues to experience and survive water losses of up to almost 90% (Oliver et al., 2020). Although DT originated as early as in the Streptophytic algal lineage, at present it is confined to seeds of most angiosperms known as orthodox seeds. Only a handful of land plants exhibit DT in their vegetative tissues (Costa et al., 2017; Leprince et al., 2017). Despite the similarities, there are clear differences in the DT mechanisms in seeds and vegetative tissues of resurrection plants. Unfortunately, the exact nature of these differences and how they evolved are not clearly understood (Lyall et al., 2020).

In this project, you will analyze DT related transcriptome data from streptophytic algae, orthodox seeds, and resurrection plants and simultaneously construct and implement comparative transcriptomics pipeline(s) to highlight the similarities and differences in the mechanism of how DT is established in these different lineages. You will follow a targeted approach using already known DT-related genes and possible candidates and determine when these genes or gene families became a part of the DT gene regulatory network (GRN). The output will contribute to our understanding of the multi-level regulation of DT and generate candidate genes for further analysis.

References:

Costa, M.-C. D., Cooper, K., Hilhorst, H. W. M., & Farrant, J. M. (2017). Orthodox Seeds and Resurrection Plants: Two of a Kind? *Plant Physiology*, *175*(2), 589–599. https://doi.org/10.1104/pp.17.00760

Leprince, O., Pellizzaro, A., Berriri, S., & Buitink, J. (2017). Late seed maturation: Drying without dying. *Journal of Experimental Botany*, *68*(4), 827–841. https://doi.org/10.1093/jxb/erw363

Lyall, R., Schlebusch, S. A., Proctor, J., Prag, M., Hussey, S. G., Ingle, R. A., & Illing, N. (2020). Vegetative desiccation tolerance in the resurrection plant Xerophyta humilis has not evolved through reactivation of the seed canonical LAFL regulatory network. *Plant Journal*, *101*(6), 1349–1367. https://doi.org/10.1111/tpj.14596

Oliver, M. J., Farrant, J. M., Hilhorst, H. W. M., Mundree, S., Williams, B., & Bewley, J. D. (2020). Desiccation Tolerance: Avoiding Cellular Damage during Drying and Rehydration. *Annual Review of Plant Biology*, *71*, 435–460. https://doi.org/10.1146/annurev-arplant-071219-105542

Mapping the salt-induced gene regulatory network that guides root branching

Supervisors	Yiyun Li (Plant Physiology), Aalt-Jan van Dijk (Bioinformatics)
Type	Data analysis
Requirements	Advanced Bioinformatics
Skills	Programming, data analysis
Timestamp	April 2023

Description

As a major abiotic stress, high soil salinity severely affects plant growth and crop productivity globally. Plants are unable to move from their location, and therefore require various effective mechanisms to cope with salinity. In presence of salt, the root architecture is reshaped, which is often characterized by changes in lateral root (LR) development [1]. Though several interactions of genes involved in LR development have been identified under control condition [2], genes and pathways involved in root branching in response to salt remain unclear.

In this project, you will use bioinformatics approaches to analyze salt-related root transcriptome data to identify candidate genes contributes to the salt-induced root branching. You will implement comparative analysis among multiple in-house and publicly available salt-induced root transcriptomic datasets to highlight the similarities and differences in the tissue-specific mechanism of root branching in salt. With the identified candidate genes, you will study the regulatory interactions using gene regulatory network (GRN) inference [3] to map the salt-induced transcriptional regulatory networks that are involved in root branching.

References

[1] Van Zelm, E., Zhang, Y., & Testerink, C. (2020). Salt Tolerance Mechanisms of Plants. Annual Review of Plant Biology, 71, 403–433. https://doi.org/10.1146/annurev-arplant-050718-100005.
[2] Lavenus, J., Goh, T., Guyomarc'H, S., Hill, K., Lucas, M., Voß, U., Kenobi, K., Wilson, M. H., Farcot, E., Hagen, G., Guilfoyle, T. J., Fukaki, H., Laplaze, L., & Bennett, M. J. (2015). Inference of the arabidopsis lateral root gene regulatory network suggests a bifurcation mechanism that defines primordia flanking and central zones. Plant Cell, 27(5), 1368–1388.
[3] Van den Broeck, L., Gordon, M., Inzé, D., Williams, C., & Sozzani, R. (2020). Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling. Frontiers in Genetics, 11(May), 1–12. https://doi.org/10.3389/fgene.2020.00457