



Australia's National
Science Agency

Bioinformatics Student Exchange Program

Start your career with a research project at one of the world's premier research organizations.

2027





BSEP 2027

The Bioinformatics Student Exchange Program (BSEP) is aimed at giving overseas students the opportunity to contribute to world-class research and gain experience in an international research environment. Master and Honours students are invited to conduct original research as part of their University Thesis. This is an exciting opportunity to forge new collaboration and build up a strong international network.

Why Australia?

Australia is the “most productive of all G20 nations” with respect to papers published [[nature Index](#)] and a recent [Nature article](#) says that “Scientists from across the world are attracted to the country, which competes internationally by focusing on its strengths”.

The Commonwealth Scientific and Industrial Research Organisation (CSIRO) is Australia’s Government Research Agency and one of the largest and most diverse scientific organisations in the world. By igniting the creative spirit of our people, CSIRO deliver great science and innovative solutions that benefit industry, society and the environment.

COVID-19: International travel has been opened back up in Nov 2021. However, with volatility in cases future restrictions might prevent all or any part of their studies in Australia. Catering for this, we also offer BSEP remotely.

University	Contact Person
Freie Universität Berlin	Ulrike Seyferth Studiengangskoordination Bioinformatik Tel.: +49-(0)30/838-75336 Email: ulrike.seyferth@fu-berlin.de
Eberhard Karls University Tübingen	Prof. Dr. Daniel Huson Algorithms in Bioinformatics Tel.: +49-7071-29-70450 Email.: daniel.huson@uni-tuebingen.de
Justus-Liebig-University Giessen	Prof. Dr. Alexander Goesmann Bioinformatik und Systembiologie Tel. +49 (0)641 99-35801 Email: Gwyneth.schulz@computational.bio.uni-giessen.de
CSIRO	Prof. Dr. Denis Bauer Biosecurity Intelligence, CSIRO Phone: +61 2 9325 3174 Email: denis.bauer@csiro.au

Key dates

Dates Europe	
June	CSIRO calls for project proposals
31 st July	Program Booklet sent to the Universities
Early August to early November	Students choose proposals and get in contact with CSIRO
Early August to early November	Deadline for PROMOS or equivalent funding application
Dec	Thesis committee assesses suitability of projects and identifies appropriate co-supervisor amongst the <u>faculty</u> .
Jan	CSIRO starts recruitment process (visa)
May*	Students commence research in Australia
Oct*	Students return home
Nov*	Students finalise reports and write master thesis with input from CSIRO researchers

* Times for visit can be flexible

How to apply

Please choose the project you are interested in and get in touch with your contact person listed above. Your first step will be to organize funding by applying (see below). After a successful interview, CSIRO will issue a contract with a visa sponsorship number in January. It is crucial to apply for the Australian Visa quickly as it can take up to 3 months to be approved. CSIRO will guide you through the process but please have a look at:

Funding

Students are encouraged to apply for funding. Unless stated otherwise, the projects will not provide funding.

PROMOS

German funding through PROMOS (Deadline Early October to early November), which will cover from 300 to 500 EUR per month or traveling costs up to 1950 EUR

Note, PROMOS is not explicitly paying a health insurance, this hence needs to be covered by the student.

DAAD

The DAAD offers Study Scholarships, such as Master Studies for All Academic Disciplines <https://www2.daad.de/deutschland/stipendium/datenbank/en/21148-scholarship-database/?detail=50026200> with funding between 10 and 24 months.

There are also other funding sources available such as <http://www.ranke-heinemann.de>.

Other resources

Please choose the project you are interested in and get in touch with your contact person listed above. Your first step will be to organize funding by applying for PROMOS or equivalent sources (DAAD). After a successful interview in January, CSIRO will issue a contract with a visa sponsorship number. It is crucial to apply for the Australian Visa quickly as it can take up to 3 months to be approved. CSIRO will guide you through the process but please have a look at:

VISA:

Depending on what your plans are before and after the internship you will have to choose the right visa, with Subclass 407 being an option. For the most up-to-date information use the <https://immi.homeaffairs.gov.au/visas/getting-a-visa/visa-finder> tool.

Health insurance:

<https://www.studyaustralia.gov.au/en/plan-your-move/overseas-student-health-cover-oshc>

German information on going to Australia:

<http://www.reisebine.de/>

Official government website with information about studying and living in Australia

<https://www.studyaustralia.gov.au/>



Experience Reports from previous students

2026 – Janine Ponzer



Thanks to the Bioinformatics Student Exchange Program, I had the opportunity to conduct my Master's thesis at CSIRO in Sydney. The project addressed zoonotic evolution and the pandemic potential of influenza A virus. Within this project, I developed skills in genomic analysis and the application of machine learning algorithms, especially in data mining. Working at the intersection of bioinformatics and infectious disease research also strengthened my understanding of global public health challenges, such as pandemic preparedness, which require international cooperation in science, as exemplified by this program.

One of the most valuable aspects of the experience was the collaboration with my experienced supervisors at CSIRO. They provided guidance and support while also encouraging me to think more independently as a researcher. I learned that bioinformatics is not just about writing code to address a biological problem, but also about comparing different approaches and their advantages and limitations, making informed decisions about which methods to use, and critically evaluating and communicating the findings. The discussions with my supervisors were particularly important in shaping my project and helped me improve not only my analysis, but also the way I approach scientific thinking and writing.

Living in Sydney during my project was also a special experience. I enjoyed living in the city and discovering what it has to offer: lots of nature, beaches, sports and coffee! I met great people both at work and outside of work, while growing as a young researcher. I am very thankful for having had the opportunity to spend this time in Australia and to be part of the BSEP.

Project

Projects can be altered to fit the students interests and skills. The Transformational Bioinformatics group at CSIRO has a very broad spectrum of activities, ranging from human health to biosecurity; from basic science to real-world applications. We highly encourage you to check our webpage (<https://bioinformatics.csiro.au/>) for our activities and approach us with your **own project ideas**.

PROJECT.....	7
<i>BSEP01 Fine Tuning AlphaGenome for Wild Mice</i>	<i>8</i>
<i>BSEP02 An agent-based framework for ETL of health records and analytics.....</i>	<i>9</i>
<i>BSEP03 Automated Classification of Influenza Strains Using NLP and Machine Learning.....</i>	<i>10</i>
<i>BSEP04 Developing Antiviral CRISPR/Cas13 Guides Robust to Escape Mutants</i>	<i>11</i>
<i>BSEP05 Develop a new approach for using random forest machine learning to identify interactions</i>	<i>12</i>
<i>BSEP06 Improving Flu and Dengue Analysis with Detailed Climate Data</i>	<i>13</i>
<i>BSEP07 Extending BacXGen for viral pathogens (H5N1)</i>	<i>15</i>
<i>BSEP08: Comparative analysis of random forest in multi-omics analysis</i>	<i>16</i>
<i>BSEP09: New mathematical algorithms for epistasis</i>	<i>17</i>
<i>BSEP10: Automated input method for MUTEXA</i>	<i>18</i>
<i>BSEP11: Evolution-aware CRISPR guide design</i>	<i>19</i>
<i>BSEP12: Website Development Cas13 Guides Design Platform.....</i>	<i>20</i>
<i>BSEP13: Feature Co-occurrence Matrices in VariantSpark</i>	<i>21</i>
<i>BSEP14: AskTheSheeps -- making livestock data interactive</i>	<i>22</i>
.....	25

Project Title	BSEP01 Fine Tuning AlphaGenome for Wild Mice
Brief description of the project highlighting expected outcomes	AlphaGenome is a foundation model for predicting the functional effects of genetic variation from DNA sequence, with applications across a range of genomic and regulatory task. However, predictive performance varies between species, including mice and non-human animals. This project will extend AlphaGenome through species-specific fine tuning using available multi-omics data, focussing on wild mice populations. Improved prediction of naturally occurring genetic variations and their effects could have potential applications in developing more targeted and sustainable biocontrol strategies for wild mouse populations. <u>The intended outcome of this project is to develop an approach for improving AlphaGenome's predictive performance in mice, particularly wild-bred, while establishing the technical foundations for multi-species genome prediction.</u>
Duties/Tasks	The student will perform • Identification and curation of applicable mouse omics datasets • Fine tuning of AlphaGenome heads while retaining the generalisable representations learned in the foundational trunk • Model evaluation • Incorporation of alternative genome assemblies and species-specific sequence information while maintaining compatibility with existing architecture
Relevant field/s of study	• Machine learning • Bioinformatics
Supervisor	Letitia Sng Callum Macphillarmy
Contact Details	letitia.sng@csiro.au callum.macphillarmy@csiro.au
Location	Remote or in person (Sydney/Adelaide)

Project Title	BSEP02 An agent-based framework for ETL of health records and analytics
Brief description of the project highlighting expected outcomes	An agent-based framework for health data ETL (Extract, Transform, Load) using natural language streamlines data processing with advanced extractors, joiners, and formatters. Users can execute these to manipulate data frames through natural language commands, facilitating seamless integration and analysis of diverse health data sources. This approach simplifies complex data operations, allowing non-technical users to interact with data conversationally. Extracted data is transformed and loaded into a structured format for analysis. Code for data analysis is generated from natural language descriptions. <u>Student will implement a pipeline with agent-based extractors, transformers and analytical script executors that will be executed using LLM based function calling.</u>
Duties/Tasks	The student will perform • Design basic extractor agents (from JSON and VCF formats to tabular form) • Data joiners/transformer agents (to combine results of extractors using table fields) • Prompt engineering LLMs to generate code to perform analytics on joined data frames (executor agents) • Basic visualisations (frequency plots, heatmaps, etc)
Relevant field/s of study	• Machine learning (Scikit-learn, matplotlib) • Bioinformatics (Familiarity with VCF format) • Python programming (analytics, and plots)
Supervisor	Yatish Jain
Contact Details	Yatish.Jain@CSIRO.au
Location	Remote or in person in Sydney

Project Title	BSEP03 Automated Classification of Influenza Strains Using NLP and Machine Learning
Brief description of the project highlighting expected outcomes	This project aims to develop a text mining system that automatically assigns host-specific labels—human, swine, and avian—to various influenza strains mentioned in a corpus of 150 academic papers. The system will analyze contextual information to determine the appropriate categories, improving the efficiency and accuracy of influenza strain classification. Influenza viruses exhibit significant diversity and cross-species transmission potential, necessitating accurate classification and tracking of strains across different hosts. Current manual methods for labelling strains are time-consuming and prone to errors. By leveraging natural language processing (NLP) and machine learning techniques, <u>this project aims to automate the extraction and classification process of influenza host labels, thus improving efficiency and accuracy.</u>
Duties/Tasks	The student will perform • Data Preprocessing • Use NLP techniques to identify and extract contextual clues about host species (human, swine, avian). • Develop a Python script that integrates NLP models for contextual analysis. • Train a machine learning classifier to assign appropriate labels based on extracted context. • Validate the system using a separate dataset to ensure accuracy.
Relevant field/s of study	· Natural Language Processing Libraries (spaCy, NLTK, ...) · Machine Learning · Bioinformatics
Supervisor	Carol Lee
Contact Details	carol.lee@csiro.au
Location	Remote or in person in Sydney

Project Title	BSEP04 Developing Antiviral CRISPR/Cas13 Guides Robust to Escape Mutants
Brief description of the project highlighting expected outcomes	Objective This project focuses on developing predictive models to design antiviral Cas13 guides that are robust against escape mutations in RNA viruses. By leveraging artificial intelligence (AI), the aim is to analyze large genomic datasets to identify patterns in guide efficiency, particularly focusing on off-target effects. This will result in the creation of adaptable antiviral solutions to mitigate the impact of viral outbreaks in livestock, thereby advancing RNA therapies and improving the resilience of livestock to viral threats. In the second phase of this project, the student will investigate how RNA viruses evolve under selective pressure from antiviral interventions, using Cas13 as a model system. <u>The student will help develop anti-viral capabilities that are robust against viral evolution.</u>
Duties/Tasks	The student will perform: • Data Mining and Integration: Gather and curate viral genome datasets from public databases that track viral evolution under immune pressure. • Escape Mutation Profiling: Identify mutations associated with known immune evasion, focusing on conserved versus variable regions in viral genomes. • Comparative Analysis: Use insights from immune escape tools (e.g., epitope variability) to infer likely escape pathways in response to Cas13-based targeting. • Simulation of Cas13 Pressure: Develop in silico models to simulate guide-target interactions and predict how viral sequences might adapt to evade Cas13 interference. • Model Refinement: Integrate findings into our guide design tool to improve the prediction of guide durability and specificity in evolving viral populations.
Relevant field/s of study	• Machine Learning • Deep Learning • Bioinformatics
Supervisor	Emiliana Weiss
Contact Details	Emiliana.weiss@csiro.au
Location	Remote or in person in Canberra

Project Title	BSEP05 Develop a new approach for using random forest machine learning to identify interactions
Brief description of the project highlighting expected outcomes	Searching for genetic interactions using random forests involves a very large search space for each node of each tree. Because random forests only include a small subset of features when picking splitting features for a node, it is likely that interacting features will be missed. In addition, features with low or no marginal effects are unlikely to be selected as the most discriminating feature in a node and are therefore unable to appear in potential interactions. The aim of this project is to make alterations to the random forest machine learning algorithm to improve its sensitivity to interactions between features. <u>The expected outcome is an improved random forest algorithm that can more precisely identify interacting features in feature-rich datasets such as GWAS, including features that have a low marginal effect.</u>
Duties/Tasks	The student will • Design a method for improving random forest sensitivity to interacting features (e.g. variable mTry, stochastic or forced feature selection) • Tune the specific hyperparameters of the method to maximise recall. • Evaluate the new method against baseline RF performance with respect to interaction sensitivity and run time, using both real and synthetic GWAS data. • Integrate the improved algorithm into the VariantSpark random forest implementation.
Relevant field/s of study	• Machine Learning • Distributed Computing • Programming (Java and Python)
Supervisor	Brendan Hosking
Contact Details	brendan.hosking@csiro.au
Location	Remote or in person in Sydney

Project Title	BSEP06 Improving Flu and Dengue Analysis with Detailed Climate Data
Brief description of the project highlighting expected outcomes	The relationship between climate variables and the incidence of diseases such as flu and dengue is a critical area of research. Accurate and localized climate data can significantly enhance the quality of analysis and improve our understanding of how environmental and natural factors influence these diseases. This proposal aims to refine the current methodology by utilizing precise climate data for specific locations and dates associated with flu and dengue sequences. Data Extraction: Collect environmental factors such as minimum, maximum, and average temperatures, as well as average precipitation for the exact locations (latitude and longitude) and dates corresponding to flu and dengue sequences. Data Analysis: Apply machine learning techniques to analyse the refined climate data, investigating potential links between changing climate conditions and emerging mutations in influenza and dengue viruses. <u>The student will help understand how climate influences viral evolution.</u>
Duties/Tasks	The student will perform • Gathering and organizing accurate climate data. • Combining climate data with disease data to create a complete dataset. • Using machine learning models to study the data, improving analysis and problem-solving skills.
Relevant field/s of study	• Machine learning • Bioinformatics (alignment) • Programming (python)
Supervisor	Laurence Wilson
Contact Details	laurence.wilson@csiro.au
Location	Remote or in person in Sydney

Project Title	BSEP07 Automated Reporting for BacXgen
Brief description of the project highlighting expected outcomes	BacXgen is an end-to-end bacterial genomics pipeline that takes sequencing data through assembly, typing, antimicrobial resistance detection and phylogenetic placement. Interpreting those outputs and writing them up is currently a manual step, which is slow and inconsistent between analysts. This project will develop an automated reporting layer that assembles pipeline outputs into structured diagnostic reports using large language models constrained by validated output schemas. Each statement in a generated report will be traceable back to the pipeline result that produced it, so the reasoning can be audited rather than taken on trust. The intended outcome of this project is a reporting component that produces reliable draft interpretations of bacterial genomic analyses, along with an evaluation framework for measuring factual accuracy and detecting claims the underlying data does not support.
Duties/Tasks	The student will perform: • Curation of a representative set of pipeline outputs and matched reference reports for development and evaluation • Design and implementation of structured output schemas that constrain what the model is permitted to report • Development of the report generation component, including provenance tracking from each reported statement back to its source output • Evaluation of generated reports for factual accuracy, completeness and unsupported claims • Integration of the component with the existing pipeline and its user interface
Relevant field/s of study	• Machine learning • Bioinformatics • Software engineering
Supervisor	Yatish Jain
Contact Details	Yatish.jain@csiro.au
Location	Remote or in person in Brisbane

Project Title	BSEP07 Extending BacXGen for viral pathogens (H5N1)
Brief description of the project highlighting expected outcomes	This project will extend BacXGen, an end-to-end workflow developed for target identification in bacterial and fungal genomes, to support viral genomes. BacXGen currently uses modular Snakemake pipeline and user-friendly GUI to lower the technical barrier to genomic analysis for non-bioinformaticians and improve reporting efficiency. However, the existing tools and workflow are primarily designed for bacterial and fungal genomes and require adaptation and additional virus-specific tools to support viral genomic analysis. Working with collaborators from the Australian Centre for Disease Preparedness (ACRP), the project will incorporate appropriate viral genome analysis tools to ensure that BacXGen is an automated, end-to-end workflow for processing viral sequencing data, including H5N1. <u>The expected outcome is a robust and reproducible viral genomics pipeline that automates the process from raw viral sequencing data through analysis and reporting.</u> intended to augment the time and technical expertise required for biologists to interpret and report on genomic data. This project will establish a foundation for extending BacXGen to additional viral pathogens and supporting more efficient genomic surveillance and research workflows.
Duties/Tasks	The student will perform: 1. Working with ACDP to determine appropriate tools and pipelines for viral genomes, including currently implemented tools that require extension. 2. Integrate identified tools into the workflow while ensuring modularity, scalability and reproducibility 3. Convert pipeline to a cloud-based workflow (nextflow) 4. Validate and refine workflow based on feedback from users (e.g., ACDP).
Relevant field/s of study	Pathogen genome assembly and analysis
Supervisor	Letitia Sng
Contact Details	Letitia.Sng@csiro.au
Location	Remote or in person in Sydney

Project Title	BSEP08: Comparative analysis of random forest in multi-omics analysis
Brief description of the project highlighting expected outcomes	Multi-omics data integrates diverse molecular measurements, such as gene expression, protein abundance, and chromatin accessibility, to uncover cellular states or pathway activities underlying phenotypes. PrecisionGO represents one of the leading services in Australia for generating multi-omics datasets, bringing together wet lab and bioinformatics expertise to help provide insights on how phenotypes are regulated. With the cost of generating these datasets decreasing, we can now apply these multi-omic techniques to non-model organisms like those that exhibit a biosecurity risk to Australia. Many of the bioinformatics pipelines needed to process these multi-omics data have varying levels of requirements to run successfully and some of these requirements cannot be met for biosecurity-relevant species. Given this, there is substantial work that needs to be done making these pipelines usable in a biosecurity context. By enabling researchers to apply multi-omics analyses to pest species, they can better understand what factors influence pest phenotypes like resistance to a specific disease. In this project, we will develop and adapt existing bioinformatics pipelines and apply them to biosecurity-relevant species so that the biosecurity community can begin generating rich multi-omics datasets for pest species. <u>The student will adapt Nextflow pipelines to work with biosecurity-relevant species with the option to make it consumable to an agentic AI tool.</u>
Duties/Tasks	Tasks include: - Successfully set up and run Nextflow pipelines on CSIRO's Petrichor HPC - Identify input requirements for each pipeline and species - Adapt existing Nextflow pipelines and writing new ones if the required pipeline doesn't exist. - Make the pipelines consumable to agentic AI.
Relevant field/s of study	Bioinformatics, Nextflow, Genomics
Supervisor	Callum MacPhillamy
Contact Details	callum.macphillamy@csiro.au
Location	Remote or in-person in Adelaide

Project Title	BSEP09: New mathematical algorithms for epistasis
Brief description of the project highlighting expected outcomes	Identifying complex interactions between genetic variants (i.e., epistasis) associated to a biological outcome or phenotype remains challenging as these relationships cannot be adequately described by additive models. The detection of epistasis requires methods that can distinguish complex patterns from those expected under simpler models including those with linearity assumptions. Current approaches to epistasis detection typically rely on assumptions about the underlying relationship between genetic variation and outcome (e.g., case/control status). This project will investigate alternative approaches that are less dependent on a specific functional form, with the aim of improving the ability to detect a broader range of genetic interactions. <u>The expected outcome of this project is a coded and optimised concept and method for epistasis that can be applied across a broad range of genetic datasets and biological outcomes.</u>
Duties/Tasks	1. Review existing literature and methods 2. Run and compare code (including Boost and BitEpi) on example data and think about how the epistatic concepts in these tools could be generalised and extended. 3. Code and optimise (for practical speed) an alternative method of epistasis detection for running on real data. 4. Extend to continuous phenotype (rather than just binary case/control)
Relevant field/s of study	Coding, and knowledge of optimisation / data science / statistics.
Supervisor	Letitia Sng Brendan Hosking
Contact Details	Letitia.sng@csiro.au Brendan.hosking@csiro.au
Location	Remote or In-person in Sydney

Project Title	BSEP10: Automated input method for MUTEXA
Brief description of the project highlighting expected outcomes	Pathogen mutation surveillance is crucial to understanding the functional changes in a pathogen (virus, bacteria, etc) and how they evolve to avoid detection by the host immune system. This project is to enhance our existing pathogen tracking pipeline with an additional robust mutation method to automate the process of inputting mutations of interest. Machine learning or other statistical methods should be implemented to extract mutations that may be of interest based on the region/gene of a pathogen genome or if the frequency of specific mutations have spiked over a short period of time. Degenerate codons should be included in the consideration of this method. Addition of this will create a more dynamic pathogen mutation tracking pipeline that can be used to inform pathogen research and public health outcomes. <u>The expected outcome is to develop a ML or statistical method to automate mutation information input into the existing pathogen tracking pipeline.</u>
Duties/Tasks	The student will perform • Review literature to assess the most suitable approach (ML or statistical) for mutation extraction • Data preprocessing for relevant contextual information (COVID-19 or influenza) • Development a python script to integrate with the current pipeline to extract mutations of interest from genomic data • Validation and optimisation of the additional method with the current existing pipeline
Relevant field/s of study	• Machine Learning • Bioinformatics • Pathogen genomics
Supervisor	Carol Lee
Contact Details	Carol.lee@csiro.au
Location	Remote or in-person (Sydney)

Project Title	BSEP11: Evolution-aware CRISPR guide design
Brief description of the project highlighting expected outcomes	CRISPR and other targetable endonucleases are continuing to be explored as next generation technologies for genetic biocontrol. Being able to directly influence the genotype of invasive species is a powerful method for curbing the damage to the Australian ecosystem. However, such approaches need to contend with the genetic diversity present within wild populations. While methods exist for designing guides across heterogeneous genotypes, including the CSIRO tools VARSCOT and SAUTE, these methods do not take into account evolution. The aim of this project, will be to integrate existing methods for heterogenous guide design with models of evolution to develop a framework for designing guides that are resistant to future genetic changes. <u>The student will explore models to understand how a guide site evolves and how this can be leveraged to increase the overall effectiveness of intervention strategies.</u>
Duties/Tasks	The student will: - Perform a literature search around guide-design and evolution - Implement models of evolution to simulate target-site evolution - Integrate these models with existing tools (e.g. VARSOCT and SAUTE) to develop a framework for evolution-resistant guide design - Investigate how this platform could be applied to genetic biocontrol or to combat pathogens
Relevant field/s of study	- Bioinformatics - Programming in at least one relevant language (e.g. Python, R) - Familiarity with genetic concepts - Familiarity with machine learning
Supervisor	Laurence Wilson
Contact Details	Laurence.wilson@csiro.au
Location	Remote or In-person in Sydney

Project Title	BSEP12: Website Development Cas13 Guides Design Platform
Brief description of the project highlighting expected outcomes	This project focuses on developing a user-friendly website to host an AI model that predicts the efficiency of CRISPR-Cas13 guides targeting RNA sequences. The current model, already developed, identifies guides with high knockdown efficiency. The website will enable researchers to input genomic sequences, run predictions, and visualise guide performance. <u>Expected outcomes include a fully deployed, scalable platform that makes advanced guide design tools accessible to the scientific community.</u>
Duties/Tasks	• Design and implement a user-friendly web interface for a CRISPR-Cas13 guide design tool • Integrate backend prediction models and genomic data input options • Deploy the platform on cloud services (e.g., AWS) for accessibility and scalability • Evaluate performance and usability with test datasets and user feedback
Relevant field/s of study	• Web development (HTML/CSS/JavaScript) • Python • Bioinformatics background • Experience with cloud deployment (AWS)
Supervisor	Emiliana Weiss
Contact Details	Emiliana.weiss@csiro.au
Location	Remote or in person in Canberra

Project Title	BSEP13: Feature Co-occurrence Matrices in VariantSpark
Brief description of the project highlighting expected outcomes	VariantSpark is a scalable machine learning platform for genome-wide analysis that uses random forests to capture higher-order genetic interactions, unlike traditional GWAS methods that evaluate variants independently. In this project, you will implement feature co-occurrence matrices in VariantSpark to identify genetic variants that frequently appear together in trained decision trees. The insights derived from these matrices can be used to reveal correlated features and guide supervised dimensionality reduction by identifying variant subsets that jointly contribute to phenotype prediction. <u>You will contribute a new feature to our flagship software, VariantSpark, with option for a peer-reviewed publication.</u>
Duties/Tasks	The student will perform: • Investigate and select an appropriate data structure for storing and querying feature co-occurrence in random forests • Design and implement an efficient algorithm to generate co-occurrence matrices from trained VariantSpark models • Benchmark the performance and scalability of the implemented method on large genomic datasets • Apply supervised dimensionality reduction using RF-PHATE guided by co-occurrence patterns (if time permits) • Document the methodology and contribute code to the VariantSpark codebase
Relevant field/s of study	Software engineering (Python, Scala), distributed computing (Spark), some knowledge of data science
Supervisor	Brendan Hosking
Contact Details	Brendan.Hosking@CSIRO.au
Location	Remote or in person in Brisbane

Project Title	BSEP14: AskTheSheeps -- making livestock data interactive
Brief description of the project highlighting expected outcomes	CSIRO holds some of Australia's biggest genomic data resources in the livestock industry. However, the data is currently siloed and can only be queried by experts. For the human space we developed AskBeacon, a Large Language Model enable users to query genomic data simply by asking questions. AskBeacon is built on the Beacon protocol, which is the globally accepted standard for genomic data exchange and our 2023 Nature Biotechnology implementation of it in the cloud. The student will be setting up AskBeacon for a sheep dataset on cloud infrastructure and then work with biological researchers to draft prompts for <u>gaining insights of the data round wool quality and genetic determinants of traits to guide future breeding programs.</u>
Duties/Tasks	The student will perform: • Set up AskBeacon on AWS resources • Prepare the sheep genomic data for ingestion in beacon for both genomic and meta-data. • Draft prompts and follow up their validity with standard bioinformatics tools for genomic analysis • Prepare a document outlining the approach and the outcome
Relevant field/s of study	Software engineering (Python), cloud computing (Terraform & AWS), bioinformatics
Supervisor	Callum Macphillamy
Contact Details	Callum.Macphillamy@csiro.au
Location	Remote or in person in Adelaide

Project Title	BSEP15: Machine learning to predict DNA packaging efficiency in viral vectors
Brief description of the project highlighting expected outcomes	Efficient packaging of nucleic-acid cargo into viral vectors depends on multiple properties of the cargo and not solely on sequence length. This project will develop a machine-learning framework to predict DNA packaging performance from sequence-derived and physicochemical features. The student will compile suitable DNA packaging data, extract relevant sequence descriptors, and train and evaluate predictive models to identify features associated with efficient packaging. The expected outcome is a computational model capable of ranking candidate DNA sequences according to their predicted packaging compatibility and identifying sequence characteristics that may influence packaging efficiency.
Duties/Tasks	• Curate and preprocess available DNA packaging datasets. • Extract sequence-derived features such as sequence length, nucleotide composition, GC content and other relevant descriptors. • Explore additional structural or physicochemical features relevant to packaging. • Train and compare machine-learning models for predicting DNA packaging performance. • Evaluate model performance using appropriate cross-validation approaches. • Determine which features contribute most strongly to packaging predictions. • Develop visualisations and a simple workflow for ranking new DNA sequences.
Relevant field/s of study	• Machine learning (Scikit-learn) • Bioinformatics / Computational Biology • Genomics / Gene therapy • Python programming • Data analysis and visualisation
Supervisor	Anne Klein
Contact Details Location	anne.klein@csiro.au Remote or in-person (site to be confirmed)

Project Title	BSEP16: Expansion of CapBuild for modelling viral capsids and virus-like particles
Brief description of the project highlighting expected outcomes	CapBuild is a computational platform for constructing complete viral capsids from experimentally determined structural information. It was initially developed for adeno-associated virus (AAV), but the same framework could potentially be adapted to other symmetric viral capsids and virus-like particles (VLPs). This project will extend and validate CapBuild for selected non-AAV systems, such as non-AAV parvoviruses, bacteriophage VLPs, HBV core particles or plant-virus capsids. The expected outcome is an expanded CapBuild framework capable of reconstructing multiple capsid architectures and supporting future studies of viral structure and nucleic-acid packaging.
Duties/Tasks	- Identify suitable experimentally determined capsid structures from structural databases. - Process and standardise structural files for use within CapBuild. - Adapt the existing Python pipeline to support selected non-AAV capsids and VLPs. - Implement structural alignment and capsid reconstruction procedures where required. - Validate reconstructed capsids against reference structures. - Characterise capsid geometry and internal structural properties. - Document the expanded workflow and generate structural visualisations.
Relevant field/s of study	- Bioinformatics / Computational Biology - Structural Bioinformatics - Molecular Modelling - Virology - Python programming
Supervisor	Anne Klein
Contact Details	anne.klein@csiro.au Remote or in-person (site to be confirmed)



As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

Contact us

1300 363 400
+61 3 9545 2176
csiroenquiries@csiro.au
csiro.au

For further information

Biosecurity Research Unit
Denis Bauer
+61 2 9325 3174
Denis.Bauer@csiro.au
csiro.au

Image copyright Unsplash. Images generated with GPT