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.

2026

BSEP 2026

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 Transformational Bioinformatics, eHealth, CSIRO Phone: +61 2 9325 3174 Email: denis.bauer@csiro.au

Key dates

Date	
June	CSIRO calls for project proposals
31st 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 faculty.
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.studyinaustralia.gov.au/english/live-in-australia/insurance

German information on going to Australia:

http://www.reisebine.de/

Official government website with information about studying and living in Australia

www.studyinaustralia.gov.au

Experience Reports from previous students

2024 – Liam Reoch

The time I spent working with CSIRO was invaluable in my last year of university, as it placed me in the centre of a real life programming operation and challenged me to navigate working in a team. In particular, having to ensure that I was consistently keeping our project supervisor informed on our progress pushed me to deepen my critical reflection skills, and quantify exactly what we'd achieved. This made me especially mindful of our priorities, and ensured that as team lead, I steered the group in a well-considered direction.

I enjoyed being in a position to help less experienced team members identify their strengths and assigning them tasks that built on those.

A critical aspect of this project was having to integrate into an existing codebase and develop a solution that would allow future teams to easily step in where we left off. This project gave us a unique opportunity to develop with a focus on extensibility and maintainability.

One of the key features that we explored was using infrastructure tools such as Terraform, which allowed us to ensure that when the project was handed over, all our AWS configurations could be easily rebuilt with minimal intervention. This was my first real exposure to infrastructure-as-code, and it opened my eyes to what was possible with these skills.

Being able to discuss our ideas with our project supervisor was extremely insightful, especially when he was able to identify an approach that we were not aware of. This guidance, combined with the skills I'd developed over the past years of study and industry work, provided me with a clearer vision of the kinds of directions I could take in the future.

Our project supervisor went above and beyond with regard to providing us helpful advice, both specific to the project as well as with regard to skills and technologies that we could investigate to improve ourselves as computer scientists and software engineers.

This project gave me a refined understanding of the kinds of tools and requirements that are used in real-world applications, and helped me identify skills that would be useful in my career.

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 Title	**BSEP01 Predicting disease-risk from multiethnic genomic data using machine learning**
Brief description of the project highlighting expected outcomes	There is a significant genetic component to disease risks in humans. Can we predict these risks early? Find out with machine learning (ML) and the world's largest genomic datasets. This project aims to explore predictive performance of existing genomic risk models against ML algorithms for complex diseases. The student will also be able to explore ethnic-specific PRS/genetic variants by analysing and interpreting the tree-splitting behaviours in random forests generated by VariantSpark. At the end of the project, the student will determine the most suitable algorithm for generating polygenic risk scores of a certain complex disease, considering computational efficiency and its applicability across different ethnicities.
Duties/Tasks	The student will perform • Trait prediction analysis using VariantSpark • Use HAIL/AWS/TREs with Jupyter notebooks for large scale data analysis • Use ML algorithms to build predictive models with genetic data • Compare currently available PRS models (e.g., LD-pred2, LDAK-SumHer, PRS-CS, PRSice) with ML algorithms
Relevant field/s of study	• Machine learning • Bioinformatics • Statistical genetics
Supervisor	Letitia Sng Anubhav Kaphle
Contact Details	Letitia.sng@csiro.au Anubhav.kaphle@csiro.au
Location	Remote or in person (Sydney/Melbourne)

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. Student will implement a pipeline with agent-based extractors, transformers and analytical script executors that will be executed using LLM based function calling.
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	Anuradha Wickramarachchi
Contact Details	Anuradha.Wickramarachchi@csiro.au
Location	Remote or in person in Adelaide

Project Title	**BSEP03 Modelling Digital Consent Framework for Patient-Controlled Genomic Data Sharing**
Brief description of the project highlighting expected outcomes	This project will aim to generate a conceptual framework for integrating Verifiable Credentials (VCs) with FHIR Consent to enable patient-controlled, portable consent management in clinical genomics via personal datastores, modelling VC issuance, lifecycle (updates/revocation), and OIDC4VP-based presentations, and also assessing ISO 27560 alignment. The aim is to cater to a model scenario involving three entities and 2 clinical contexts: 2 *healthcare* institutions and a patient: • **Genomic screening context (Institution A):** issues a signed FHIR Consent VC to be stored on the patient's personal datastore; stores genomic data locally on their server. • **Genomic diagnostic context (Institution B):** uses OIDC4VP to request VC presentation from the patient, verify the existing VC, then issues a new completed and signed consent VC for data access and testing. We will assume that both institutions run FHIR servers with OIDC4VP support and middleware to verify VCs, extract FHIR Consent, and enforce patient consent. **Key specs:** W3C VC, OIDC4VP, FHIR Consent, ISO 27560. The student will help develop a conceptual framework for consent interaction in clinical genomics application using latest development around datastores and VCs.
Duties/Tasks	The student will: • Model issuance of VCs embedding FHIR Consent to personal datastores, including OIDC4VP presentation and verification flows. • Outline consent lifecycle management—updates, amendments, revocation—and mechanisms for checking current validity. • Compare the FHIR Consent/VC framework to ISO IEC 27560:2023, identifying alignments, gaps, and integration recommendations. • Assess personal datastore platforms (e.g., Solid, Verida) for securely storing, managing, and presenting consent VCs, and specify necessary features (secure storage, management APIs, authorization).
Relevant field/s of study	• Software engineering • Health Informatics • Cybersecurity
Desired skillsets	- Understanding of verifiable credentials and decentralised identity concepts (DIDs, VCs), healthcare data interoperability (HL7 FHIR) - Exposure to system design and architecture principles - Ability to conduct rigorous conceptual analysis and design.
Supervisor	Anubhav Kaphle
Contact Details	Anubhav.kaphle@csiro.au
Location	Remote or in person in Melbourne

Project Title	**BSEP04 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, this project aims to automate the extraction and classification process of influenza host labels, thus improving efficiency and accuracy.
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	Nehleh Kargarfard Carol Lee
Contact Details	nehleh.kargarfard@csiro.au carol.lee@csiro.au
Location	Remote or in person in Sydney

Project Title	**BSEP05 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. The student will help develop anti-viral capabilities that are robust against viral evolution.
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	**BSEP06 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. 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.
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	**BSEP07 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. The student will help understand how climate influences viral evolution.
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	Nehleh Kargarfard Laurence Wilson
Contact Details	nehleh.kargarfard@csiro.au laurence.wilson@csiro.au
Location	Remote or in person in Sydney

Project Title	**BSEP08 Developing bioinformatic workflow pipeline**
Brief description of the project highlighting expected outcomes	Workflow pipelines are widely used in the bioinformatics world to streamline and automate the execution of various tools. In this project, you will work with a range of bioinformatics software, including traditional machine learning models and modern large language models (LLMs). Your task will involve assessing the resource usage, as well as the input and output requirements for each tool. Subsequently, you will develop an automated pipeline to integrate these tools using a domain-specific language (DSL) like Nextflow. You will deploy this pipeline on our in-house HPC with opportunities to practice deployment on AWS cloud. You will help develop standardized workflow generation capability for bioinformatics task on HPC and cloud.
Duties/Tasks	The student will perform: Review of existing bioinformatics tools and literature • Wrap and deploy software in containers • Develop automated pipelines using DSL • Deploy pipelines on HPC and AWS
Relevant field/s of study	• HPC, Cloud, docker, kubernet • Bioinformatics • Programming (bash, python, java, Nextflow)
Supervisor	Qinying Xu (Christina)
Contact Details	Qinying.xu@csiro.au
Location	Remote or in person in Brisbane

Project Title	**BSEP09 BacXGen: an integrated cloud-based pipeline for pathogen genome analysis**
Brief description of the project highlighting expected outcomes	This proposal builds on a previous SC grant combining two projects, "A web-based tool for target identification in bacterial and fungal genomes" and "BacXGen". These projects initially proposed to make a GUI and improve two separate existing lightly documented pipelines for bacterial and fungal genome assembly and target identification more efficient. The present team has focused their efforts to ensure that the pipeline is modular, and user-friendly enough so that non-bioinformaticians can easily use the pipeline, circumventing the need for the GUI. Moreover, the team has extensively documented the pipeline for easy user use, significantly improving the efficiency and speed of the workflow. The new modular structure of the workflow will enable for ease of integration of our collaborators' tools. The student will continue to develop other aspects of this pipeline, primarily summarisation and interactive visualisation of the analyses results. This is because the pipeline incorporates several command line tools and custom scripts which are difficult to integrate and visualise as a whole. Development of an interactive interface of results will facilitate a clearer understanding of the analysis results, significantly increasing the usability of the pipeline for both CSIRO and their collaborators. We currently have some R scripts for summary graphs that can be modified for our purpose. Another main goal would include converting this snakemake pipeline to a cloud-based workflow.
Duties/Tasks	The student will perform: 1. Summaries of the quality control data – which can be achieved with the package multiqc at this point, however having an R shiny app for the results would make the results more interactive and usable. 2. Interactive tree visualisation – build an R shiny app for easy navigation of graphs showing relatedness between individual samples. 3. Convert pipeline to a cloud-based workflow (nextflow)
Relevant field/s of study	Pathogen genome assembly and analysis
Supervisor	Priya Ramarao-Milne
Contact Details	Priya.Ramarao-Milne@csiro.au
Location	Remote or in person in Brisbane

Project Title	**BSEP10: 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. However, heterogeneous data types (continuous vs. count data), normalisation strategies, and missing values make this analysis challenging. Random Forest (RF) is a widely-used, flexible machine learning method, but it (Breiman's RF) exhibits bias when handling mixed data types (Strobl *et al*. 2007), can be sensitive to large missingness, and can lead to instability of importance measures with correlated features. Specifically, RF favours predictors with more split option - continuous variables or categorical features with many level - leading to skewed feature importance and wrong emphasis on certain omics features. In this project, we will use simulated data to better understand this behaviour in integrative omics analysis (*n*-integration) and use different bias reduction techniques to test for improvements in performance. The student will help adapt Random Forest Machine Learning approaches to multi-omics analysis, potentially adding a new functionality to our flagship software, VariantSpark.
Duties/Tasks	Tasks include: - Setting up the scMultiSim (or any other alternative simulation method, as applicable) environment and generate initial multi-omics datasets - Measuring influence of data properties on RF performances - Apply bias correction techniques (conditional inference forests, adjusted gini split gain, layer aware sampling, block RF) and check for improvements in RF performance - Benchmark against MOFA+, covariance optimisation methods in mixomics package based on true recovery rates, robustness to different data properties - Provide recommendations for best-practice RF variants when analysing heterogeneous omics data.
Relevant field/s of study	Bioinformatics, Statistical Data Science
Supervisor	Anubhav Kaphle
Contact Details	Anubhav.kaphle@csiro.au
Location	Remote or in-person in Melbourne

Project Title	**BSEP11: New mathematical algorithms for epistasis**
Brief description of the project <u>highlighting expected outcomes</u>	Finding complex interactions in genome data is difficult, particularly the concept of epistasis around some kind of non-linear (non-additive) relation between specific genes and the probability of an effect - such as disease occurrence. Thus, one approach to finding epistasis involves searching for genes where the observed number of disease case&control cases can't make sense (ie. low maximum likelihood) under some linearity assumptions. Historic best practices in finding epistasis include the BOOST algorithm, which finds the maximum likelihood of case&control cases under linearity assumptions on the log-odds of disease. We have also coded a method under linearity assumption of disease probability itself (not the log-odds of it). However there is wider space between these – and we want to make an epistasis detector which is not (or is less) sensitive to specific form of the linearity assumption employed (whether characterised directly or indirectly). This may add resolving power to best practices and uncover interesting and new interactions between genes and disease. <u>Expected outcome: a coded and optimised alternative concept and method of epistasis detection.</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. (eg. code that maximises empirical likelihood: $\Sigma_i\, n_i \log(p_i)$ might be broadened to $\Sigma_i\, n_i f(p_i)$ for some monotonic (and convex?) function f , or investigate other datascience/statistical measures would indicate strongly non-linear (or even XOR-type) disease presentation observations, or alternative information-theoretic measures) 3. Code and optimise (for practical speed) an alternative method of epistasis detection for running on real data. Bonus: extend to continuous phenotype (rather than just binary case/control)
Relevant field/s of study	Coding, and some knowledge of optimisation / datascience / statistics.
Supervisor	Mark Burgess
Contact Details	Mark.a.burgess@csiro.au
Location	Remote or In-person

Project Title	**BSEP12: Automated input method for StrEpiFun**
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. The expected outcome is to develop a ML or statistical method to automate mutation information input into the existing pathogen tracking pipeline.
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 (viruses)
Supervisor	Carol Lee
Contact Details	Carol.lee@csiro.au
Location	Remote or in-person (Sydney)

Project Title	**BSEP13: Advancing VariantSpark to Unlock Complex Genetic Insights using Machine Learning**
Brief description of the project highlighting expected outcomes	VariantSpark is a powerful, scalable machine learning platform designed for genome-wide analysis. Unlike traditional approaches, VariantSpark can capture complex, non-linear genetic interactions (epistasis), making it especially effective for uncovering hidden associations in genomic information. While random forests are known for their robustness, research has shown that default hyperparameter settings often underperform when applied to genomic data. This project aims to systematically evaluate how key variables, such as sample size, number of genetic variants, and disease type, affect random forest outcomes in genetic association studies. The goal is to understand how to best tune these models and to find the most reliably way to measure which hyperparameter settings work best. The student will be helping to improve the accuracy and reliability of genetic research using machine learning.
Duties/Tasks	The student will perform: • Genetic association analysis using VariantSpark • Use HAIL/AWS/TREs with Jupyter notebooks for large scale data analysis • Systematically evaluate and determine best 'default' hyperparameters for random forests in a genetic context. • Identify the best metric to tune hyperparameters with • Report findings
Relevant field/s of study	Bioinformatics, machine learning, genetics
Supervisor	Letitia Sng
Contact Details	Letitia.Sng@csiro.au
Location	Remote or in-person (Sydney)

Project Title	**BSEP14: Refining and documenting sBeacon data architecture as a white paper**
Brief description of the project highlighting expected outcomes	sBeacon implements GA4GH (Global Alliance for Genomics and Health) Beacon protocol. Beacon protocol is an API specification which outlines the schema of data resources and the REST API, however, provides the liberty of implementation (https://docs.genomebeacons.org). sBeacon takes a cloud first approach in implementing Beacon and leverages serverless technologies to create an implementation that is highly scalable (https://doi.org/10.1038/s41587-023-01972-9). The project documents sBeacon data architecture and the indexing strategy to compile a white paper, so that the architecture can be published for re-use in similar data exchange solutions.
Duties/Tasks	The student will perform • Study of Beacon protocol and sBeacon implementation. • Document potential areas of improvement for efficient metadata handling, after reviewing existing implementations for data exchange and storage. • Compile a white paper outlining the implementation for the Bioinformatics data exchange community.
Relevant field/s of study	• Computer science • Bioinformatics • Software Engineering
Supervisor	Anuradha Wickramarachchi
Contact Details	Anuradha.Wickramarachchi@csiro.au
Location	Remote or In-person in Adelaide

Project Title	**BSEP15: 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. 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.
Duties/Tasks	The student will: - Perofrm 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 appled 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	**BSEP16: 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. Expected outcomes include a fully deployed, scalable platform that makes advanced guide design tools accessible to the scientific community.
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	**BSEP17: Galaxy Workflow for Metagenomic Taxonomic Profiling of RNA-Seq Data**
Brief description of the project highlighting expected outcomes	Environmental RNA-Seq datasets contain a wealth of information not only about host gene expression but also about microbial and eukaryotic communities present in the sample. In this project, you will be integrating a range of tools, for streamlining taxonomic classification and analysis of RNA-Seq data within the Galaxy platform.
Duties/Tasks	**Expected outputs include:** • Integrating quality control, de novo assembly, and k-mer-based taxonomic classification. • Support comprehensive metatranscriptomic profiling. • Deliver a reproducible workflow available for use in the Galaxy environment **Duties** • Wrapping and validating KMA and CCMetagen as Galaxy tools for taxonomic classification. • Building and indexing custom reference databases for KMA usage. • Developing and testing a Galaxy-compatible workflow for metagenomics.
Relevant field/s of study	• Environmental metagenomics • Bioinformatics • Galaxy and HPC
Supervisor	Berenice Talamantes Becerra
Contact Details	Berenice.TalamantesBecerra@csiro.au
Location	Remote or in Person in Canberra

Project Title	**BSEP18: Make personalizing prescription drug design a breeze**
Brief description of the project highlighting expected outcomes	Avoiding adverse reactions to prescription drugs are one of the priority areas for healthcare. While a person's genome can inform drug and dosage choices (Pharmacogenomics), there currently does not exist a robust and scalable solution to decode this information. Building on PharmCAT as well as an in-house dbSNP rsID-based lookup workflow, the student will build a cloud-based solution that tackles Lambda's execution time limit, lack of parallel processing, and limited fault tolerance in reference data updates. This project will explore scalable architectural patterns—such as SNS-driven parallel orchestration or re-architecting the pipelines with AWS Batch or Fargate—for enhanced performance and resilience. The student will design and prototype a more robust execution model for pharmacogenomic applications, with measurable improvements in scalability, fault tolerance, and maintainability.
Duties/Tasks	The student will perform: • Evaluate limitations of the current Lambda-based pharmacogenomics workflows • Investigate AWS-native options for parallelism and long-running tasks (e.g. Batch, Fargate, Step Functions) • Design and prototype a scalable pipeline architecture • Benchmark execution time, cost, and fault tolerance of the proposed solution
Relevant field/s of study	Software engineering (Python), cloud computing (Terraform & AWS), bioinformatics
Supervisor	Nick Edwards
Contact Details	Nick.Edwards@csiro.au
Location	Remote or in person in Brisbane

Project Title	**BSEP19: 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. You will contribute a new feature to our flagship software, VariantSpark, with option for a peer-reviewed publication.
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	Nick Edwards
Contact Details	Nick.Edwards@csiro.au
Location	Remote or in person in Brisbane

Project Title	**BSEP20: 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 gaining insights of the data round wool quality and genetic determinants of traits to guide future breeding programs.
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

Image copyright Unsplash.Midjourney, GPT4

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

Health & Biosecurity
Denis Bauer
+61 2 9325 3174
Denis.Bauer@csiro.au
csiro.au/H&B