HEALTH AND BIOSECURITY
www.csiro.au

Bioinformatics Student Exchange Program

CSIRO – Germany

1 BSEP 2021

Australia was featured in a Nature article stating that “Scientists from across the world are attracted to the country, which competes internationally by focusing on its strengths”, it was also named the “most productive of all G20 nations” with respect to papers published [nature Index].

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, we deliver great science and innovative solutions that benefit industry, society and the environment.

In order to give overseas students the opportunity to contribute to world-class research and gain experience in an international research environment, the eHealth program is running the Bioinformatics Student Exchange Program (BSEP) with foreign Universities. The program is aimed at Master and Honours students and invites them to join CSIRO to conduct original research. This is an exciting opportunity to forge new collaboration with CSIRO as the hub for bioinformatics research in Australia.

Master and Honours students in Bioinformatics will have the opportunity to join CSIRO for 23 weeks (5 months) and undertake a research project that contributes towards their Thesis. The project will be proposed by CSIRO researchers who also agree to co-supervise the student and assist in writing the thesis.

COVID-19: International travel has been opened back up in August 2020. However, with volatility in cases future restrictions might prevent the 2021-participants to conduct all or any part of their studies in Australia. Catering for this, we 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	A/Prof. Dr. Denis Bauer Transformational Bioinformatics, eHealth, CSIRO Phone: +61 2 9325 3174 Email: denis.bauer@csiro.au

1.1 Key dates

Date	
June	CSIRO calls for project proposals
31st July	Program Booklet sent to the Universities
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	Students choose proposals and CSIRO starts recruitment process (interview, 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*

1.2 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 Internationale Forschungsaufenthalte für Informatikerinnen und Informatiker (IFI).
https://www.daad.de/de/im-ausland-studieren-forschen-lehren/forschen-im-ausland/ueberblick-ifi-foerderschienen/

- 818.93 EUR a month after deductions
- Travel expenses covered 1,975 EUR.
- When receiving the IFI scholarship, the student is automatically insured with “Die Continentale” for the duration of the scholarship + two weeks before and after starting/end date.

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

1.3 How to apply and 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:

The visa subclass 402 doesn’t exist anymore, for us the Temporary Activity (subclass 408) visa – Research Activities applies now: https://immi.homeaffairs.gov.au/visas/getting-a-visa/visa-listing/temporary-activity-408/research-activities#Overview

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

1.4 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**.

1.5 Experience Reports from previous students

Figure 1 Janina in Australia.

We are Pascal and Janina, Bioinformatics Masters students at the FU Berlin. We are taking part in the exchange program with Denis Bauer's Transformational Bioinformatics group at CSIRO in Sydney. We'd like to share our experience about working and living here in Sydney, the application process and the challenges that the years 2019 and 2020 brought with them.

The Commonwealth Scientific and Industrial Research Organisation is Australia's science agency, similar to the RKI in Germany, and also collaborates with industry partners. As a consequence, doing research here will give you insights not only into the workings of a government agency but also something close to an industry experience.
The group consists of two teams with their own team leads, one of which you will work more closely with, but you will also have plenty of opportunities to meet with the whole group in weekly meetings. There are also CSIRO wide meetings, where you can get a glimpse at all the research going on.

The application process for the exchange program is rather informal. You'll get to know the team leads and work out a project that suits you with them. However, the procedure to acquire funding is a longer process. One of us got the PROMOS funding, the other the DAAD IFI scholarship. For both of them you'll need a letter of recommendation from a professor at your university. It makes sense to ask the person, who will be grading your thesis, for the recommendation as they can attest to how motivated you are to successfully complete the project! For the DAAD scholarship you will also need to do some reading on your own, as they expect a (rough) project plan. Make sure to get a good idea of your project, when discussing it with your supervisor. Furthermore, you will need a proof of your English language proficiency. So plan some time for doing a test, if you don't have one already (the university offers one a few times a year).

After acquiring the funding, you'll be able to get a contract from CSIRO with which you can apply for a visa. The processing times vary, so be sure to check out the webpage.

Looking for apartments in Sydney can be tricky, when you're not able to visit the apartments in person, so it may be a good idea to arrive a couple of weeks prior to starting your placement. Once you're all set, it's time to start experiencing all the fun things Sydney has to offer like the beach walks, Blue Mountains hikes, barbeques and much more. If you're looking to meet new people, you're not alone, many travellers and short-term residents pass through Sydney. Roommates are of course always a good starting point, but there are also plenty of other ways like apps such as We3, Meetup, Bumble BFF and groups on social media. There are also international student meetups organized by the universities that often allow others to join, for example the UTS outdoor adventure club.

We both warmly recommend you consider this opportunity to participate in cutting edge research in a friendly atmosphere with helpful and knowledgeable coworkers and supervisors in this wonderful city. Although there are some bureaucratic hurdles to overcome during the application process, you will not regret this effort.

Cheers

COVID-19

As mentioned above, the years 2019/2020 have been out of the ordinary for Australia and the rest of the world. As we've been here at different times, we each would like to address the unique challenges we faced separately.

2019 summer - Pascal
When I arrived in Sydney, the summer was just about to start. I pretty fast got an idea of the Australian summer with one of my first days reaching temperatures of over 35 °C (the day of a little German Oktoberfest party at work). However, little did I know that this would become one of the most extreme summers of all time. With temperatures above 45 °C in Sydney, it was the warmest summer on record. So there were plenty of good opportunities to spend the weekends at one of the various beautiful beaches of Sydney.
Unfortunately in combination with little to no rain during the whole summer, Australia saw one of the worst bushfire seasons in history. Although Sydney was never at direct risk of the fire itself, it still had a huge impact on the daily life. Because of the smoke of the bushfires, there were days with hazardous air quality. So one was encouraged to reduce all outdoor activities to a limit during these days. However, already during this time CSIRO was quite flexible with working from home. So it was unproblematic to avoid going outside on the few really bad days. Due to the bushfires I had to cancel my plans of visiting the surrounding areas of Sydney and the Blue Mountains. Since Sydney itself has already so many things to do and see that I could just do a fraction of them and definitely have to come back at one day, this was not really dramatic.
Just when the bushfire season was about to end, the weather turned into the other extreme and there were a few days of heavy rainfall. But again, since CSIRO and the team were so flexible and supportive with the working conditions, it was easy to overcome the few extreme days, were it was hard to get to work.

2020 March-July: Janina
I arrived two weeks before the borders closed down for international travellers. Before the offices were closed, I had a few days to personally get to know my coworkers, but sadly also had to stay home for some time quarantining myself while waiting for the results of a Coronavirus test result. Unfortunately the work site will not fully reopen before my time here is up, but we have all adapted to working from home. While communication is more difficult, when you can't just walk up to someone's desk to ask them a question, weekly meetings and catch-ups with my supervisor have helped to keep me on track.
Seeing CSIRO and in particular our team quickly respond to the challenge of the pandemic outbreak and being able to observe so closely the critical research that is being conducted has truly been a fascinating experience.

1.6 Projects

BSEP01: sVEP motif and regulatory features annotation module

Using cloud-native technology to accelerate disease gene annotation for clinical pathology.

When a patient’s genome has been sequenced, a genetic pathologist must review any changes in the sequence (called variants or mutations) to identify any variants that are the likely cause of a disease or could help inform treatment decisions. As part of this clinical reporting, the pathologist categorises variants in the sequence as either ‘pathogenic’, ‘non-pathogenic’ or levels of uncertainty in between. An automated pre-annotation stage reduces the huge number of variants in each sequence to a set of potential variants for the pathologist to review. This pre-annotation stage is slow and does not take into account a range of information, such as previous decisions made by the pathologist, data in local or international knowledge bases, or data only reported in academic publications.

We have developed a new version of the variant effect predictor (VEP), a software tool originally developed by the European Bioinformatics Institute and in widespread use. Addressing both the flexibility and time constraints of VEP, our version uses cloud-native architectures to provide a highly modular and resource-efficient implementation of the variant annotation pipeline, serverless VEP (sVEP). This includes using machine learning (ML) methods to incorporate previous decisions from the curator.

Student will design a python or perl module to annotate motif and regulatory features from a given VCF input. The outcome of this project will be successful integration of this module in sVEP architecture. The student will apply the newly designed module on disease cohorts (Alzheimer’s, ALS) to benchmark the performance and gain new insights into the molecular bases of the diseases.

The student will perform:

- Automate data collection from ensembl
- Design annotation module for motif and regulatory features
- Integrate annotation module to sVEP architecture

Relevant fields of study:

- Statistical analysis
- Scripting programming (python/perl)
- Human genetic disease research
- Cloud architecture design

Supervisor
Dr. Natalie Twine and Yatish Jain (CSIRO, H&B)
Email: Yatish.jain@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers to top-up stipend for this project: $300/week contribution towards living expenses

BSEP02: Random Forest with population stratification for plant genomics

Implementing population stratification into CSIRO's genomics platform, VariantSpark.

VariantSpark is a powerful machine learning platform built for the high-impact domain of human genomics. It deploys as a self-serve privacy and data-ownership preserving cloud-platform, which enables academic consortia, big pharma and hospitals to analyse large-cohort genomic data to identify disease genes and develop new diagnostics or treatment avenues. With VariantSpark being application agnostic, the student can lead the first-time application to plant genomics, specifically *A. Thaliana*. For this, population stratification for plant genomics need to be addressed.

Student will learn how to run VariantSpark and run it with the *A. Thaliana* dataset. By tweaking the quality control variables address the population stratification issue with plant genome can be investigated and compared with existing literature.

The student will perform:

- VariantSpark analysis in *A.Thaliana*
- Understand the quality control measures in GWAS study
- Do a literature review for existing literature to compare results

Relevant fields of study:

- Machine learning
- Population genetics
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Dr. Natalie Twine and Yatish Jain (CSIRO, H&B)
Email: Yatish.jain@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP03: Building complex disease risk model using genomics and ML

Finding polygenic epistatic disease genes in the world's largest genomic dataset.

Polygenic risk models (PRS) are used to predict genetic disease risk from multiple genes in the genome. PRS are experiencing an uptake in the clinical space. However, current ways of generating PRS are not generalized and are unable to identify the specific genes contributing to the overall risk. Our Random Forest approach, VariantSpark, allows polygenic risk prediction but is able to attribute effects to individual genes, which helps for disease understanding and therefor treatment design.

The student will learn how to run VariantSpark with large sample cohort (e.g. MinE/ADNI) and optimise prediction/classification using a set of genetic drivers for phenotype of interest. This will be demonstrated in complex genetic disease cohort e.g. Alzheimer's/ALS. The student will help build standardized VariantSpark models for robust PRS-like predictions. The approach will be demonstrated on the world's largest genomic dataset, UK Biobank.

The resulting insights can be served out in an automated fashion akin to https://panelapp.genomicsengland.co.uk/.

The student will perform:

- VariantSpark analysis
- Use HAIL/AWS with python notebooks for large scale data analysis
- use ML to build predictive models with genetic data

Relevant fields of study:

- Machine learning
- Population genetics
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor

Dr. Natalie Twine (CSIRO, H&B)
Email: Natalie.Twine@csiro.au
Location: Sydney (or remote)

Funding:

CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP04: Ancestry prediction in the world’s largest genomic dataset.

Find out who may be related to who amongst 200,000 individuals and advance disease gene analytics.

TRIBES is a platform for relatedness detection, we are improving the performance of TRIBES to analyse >10,000 genomes at once, in a pairwise manner. The student would showcase the capability and computational performance of TRIBES 2.0 on international genomics cohort (e.g. UKBiobank). Finding the hidden ancestry in BioBank.

The student will perform:

- Perform analysis with TRIBES/Python notebooks/HPC
- Interpret and summarise results
- Highlight areas for further development of the platform

Relevant fields of study:

- Machine learning
- Population genetics
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Dr. Natalie Twine (CSIRO, H&B)
Email: Natalie.Twine@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP05: Develop a platform for CRISPR-diagnostics.

Help develop lab-free diagnostics approaches to detect genetic and infections diseases.

Effective response to an emerging pandemic hinges on the rapid and accurate diagnosis of the causative pathogen. The recent outbreaks of Ebola, Zika virus, and seasonal influenza pandemics have highlighted the significant diagnostic gaps present, meaning there is a desperate need for new approaches. While new developments such as the CRISPR-Cas based SHERLOCK system of diagnosis provide cheap and highly sensitive diagnosis, the constantly evolving nature of pathogens means the platforms must be continuously optimized for new strains. This is a time consuming and complex task.

In this project, we intend to combine powerful machine learning and modelling approaches to develop a platform for the analysis and application of CRISPR based detection methods. This will include developing methods for the analysis of populations to identify the optimal target site, as well as improving gRNA effectiveness.

The student will perform:

- Develop predictive models of CRISPR efficiency using machine learning approaches
- Develop a platform for high-throughput design of CRISPR-diagnostics.
- Interpret and summarise results
- Highlight areas for further development of the platform

Relevant fields of study:

- Machine learning
- Population genetics
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Dr. Laurence Wilson (CSIRO, H&B)
Email: Laurence.Wilson@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP06: Tools for CRISPR-based gene therapy.

Making Gene Therapy personalized by using a patients unique genomic mark-up to guide editing.

Gene therapies are a powerful way to treat the root cause of genetic diseases, which are often incurable with standard treatments. In particular, CRISPR-Cas therapies have the potential to permanently cure patients with otherwise intractable genetic disorders. However, these therapies need to be carefully designed in order to prevent off-target effects.

In this project, the student will develop bioinformatic tools for optimal CRISPR-Cas guide design. Using our published SNP-aware CRISPR-guidance tool VARSCOT, the student will develop a pipeline to enable personalized gene-therapy.

The student will perform:

- Contribute to tools for CRISPR-Cas guide design. There are several options, including:
- Build a machine learning model of Cas function, and use this model to improve guide design
- Build a tool to optimise guide design for variability in patient genomes
- Build a machine learning model to optimize gene therapy vectors for specificity and expression

Relevant fields of study:

- Machine learning
- Population genetics
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Dr. Suzanne Scott (CSIRO, H&B)
Email: Suzanne.Scott@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP07: Tools for monitoring the spread of gene drives across populations.

Save Australia's unique biodiversity by making Gene Drive applications safer.

Gene drives are powerful systems that can be used to propagate genetic changes through a population. However, controlling the spread of gene drives to the target population is crucial for managing the risks associated with gene drive.

This project will focus on developing computational tools to monitor the spread of gene drive systems across populations. The student will develop synthetic dataset from literature derived problem spaces to demonstrate the applicability of the technology

The student will perform:

- Develop alignment-free tools using machine learning approaches
- Interpret and summarise results
- Highlight areas for further development

Relevant fields of study:

- Machine learning
- Biosecurity
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor

Dr. Aidan Tay (CSIRO, H&B)
Email: Aidan.Tay@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP08: A universal convolutional deep learning tool for identifying optimal CRISPR-Cas targets.

Make genome editing safer with a continuously learning model.

Current CRISPR prediction tools require guide sequences to predict the efficiency of. This is usually a list of guides, as specified by the user. To simplify usage, some tools will use simple pattern matching tools to identify guides based on their PAM within a longer DNA sequence. However, this adds human-specified rigidity to an otherwise machine learning (ML) model. Because, despite targets with the correct PAM (I.e. -NGG for CRISPR-Cas9 being the most efficient), in some cases alternative PAMs (such as -NAG) can be optimal. When the choice of targets may be limited, having a computational tool miss the optimal target can lead to a poor less-than-optimal experiment.

Here the idea is to integrate the target identification step into the ML model. This is theoretically possible with convolutional neural network (CNNs). CNNs can identify objects in images. Here, we will instead use a CNN to identify efficient guides in DNA sequences. As well as being PAM-agnostic for a given CRISPR system (I.e. Cas9), it may be possible to make the model applicable to other systems (I.e. Cas12a). By training different layers on these two systems, the one model will be able to identify targets for both systems. Another benefit will be the ability to add and optimize the model as new CRISPR systems are discovered, or additional data for current CRISPR systems becomes available.

Currently there are many datasets available with CRISPR target efficiency. These can be combined with off-target efficiency data (GUIDE-Seq) to improve the sample size. This should be adequate to train a model able to represent genome wide CRISPR activity.

The student will perform:

- Identify available datasets (Cas9 and Cas12a for now)
- Acquire data
- Merge data into a consistent format
- Model the data

Relevant fields of study:

- Deep learning
- CRISPR modelling
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor

Dr. Laurence Wilson (CSIRO, H&B)
Email: Laurence.Wilson@csiro.au
Location: Sydney (or remote)

Funding:

CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP09: Population Genomic Simulation for Complex Phenotype.

Web-page for benchmarking genomic data.

The discovery of genomic drivers of a genetic disease is a health critical application. Scientists process population genomic data using a variety of programs to identify genes that are statistically relevant to the disease of interest. Algorithms used for this analysis are constantly evolving to deal with more complex genomic drivers. One barrier in the development of such programs is the inaccessibility of a genomic dataset for verification. An alternative approach is to use synthetic datasets.

The goal of this project is to create a platform to generate synthetic genomic data with a controllable level of complexity. The resulting software can be used to evaluate existing methods. But the more important application is to help scientists to develop more efficient methods for the discovery of complex genomic drivers.

The outcome will be a web-page enabling global challenges akin to CASP for protein folding.

The student will perform:

- Study and understand existing disease models
- Propose a configurable simulation model
- Implement the proposed model
- Explore the usage of the developled program in evaluation of existing GWAS methods.

Relevant fields of study:

- Machine learning
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics
- Genomics

Supervisor
Dr. Arash Bayat (CSIRO, H&B)
Email: Arash.Bayat@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP10: Algorithm development for the characterization of viral evolution.

Study BigData to gain insights into how viruses may acquire new traits.

We now have access to truly enormous datasets of strain information for several virus families, with over 100 thousand samples consisting of both closely related strains (Sars-COV-2) and more distantly related strains (Influenza). These datasets consist of not only the raw genetic sequences but also detailed metadata about each given strain.

This project will aim to develop improved models to use these datasets to better characterize viral evolution or epidemiology. Due to the large amount of data available this project will ideally utilize a machine learning based approach, potentially hybridized with more traditional bioinformatic approaches.

The resulting model will be able to give insights into plausible epidemiological traits (e.g. infectivity) and may shed light on how likely a virus trait has naturally occurred.

The student will perform:

- Apply machine learning to analyse viral strain datasets
- Interpret and summarise results
- Highlight areas for further development

Relevant fields of study:

- Machine learning
- Biosecurity
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Cameron Hosking (CSIRO, H&B)
Email: Cameron.Hosking@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP11: Algorithm development for the characterization of viral evolution.

Enabling genomic data sharing the cloud-native way.

Serverless Beacon is pioneering an approach to sharing genomic information between laboratories quickly and securely. The core of the project has been completed and is able to query thousands of genomes in a fraction of a second, however there is much scope for further development. This includes: Allowing a beacon to serve genomes from multiple different organisms simultaneously.

Adding sample-specific metadata to a beacon, including phenotypic data, collection information and clinical outcomes; and allowing searching and filtering by these pieces of metadata. Allowing searching and filtering by multiple pieces of variant-specific metadata.

The student will be able to extend this technology to standing up a COVID-19 Beacon, e.g. add clinical outcomes of the patient as meta-data. The student may also be able to expend the protocol to polyploid organisms (e.g.plants) updating the frequency calculation.

The student will perform:

- Adding one or more pieces of additional functionality to the serverless beacon architecture
- Compare beacon functionality with other genomic sharing implementations
- Benchmark process to establish performance impact (if any) of additional functionality
- Highlight areas for further development

Relevant fields of study:

- Machine learning
- Biosecurity
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Brendan Hosking (CSIRO, H&B)
Email: Brendan.Hosking@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

BSEP12: Streamlining Genome-scale CRISPR Knock-Out.

Enable researchers to study genomic function more systematically.

Knocking out genes systematically helps in determining the function of individual genes. CRISPR-Cas9 enables to do this systematically.

Student will combine our various tools (TUSCAN, VARSCOT) into a streamlined pipeline with pre and post processing using snakemake / nextflow. The student will be able to demonstrate the ability of the pipeline on a CSIRO dataset.

The student will perform:

- Add functionality for batch filtering and sorting
- Extract valid genes from a reference organism, then trim to CDS
- Find targets using TUSCAN (Cas9) or add functionality for other targets
- Filter off targets using VARSCOT
- Combine command line tools using snakemake
- Improve performance and throughput

Relevant fields of study:

- Machine learning
- Biosecurity
- Computer science and scripting programming (bash, R, perl/python etc.)
- Statistics

Supervisor
Daniel Reti (CSIRO, H&B)
Email: Daniel.Reti@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers to top-up stipend for this project: $300/week contribution towards living expenses

BSEP13: Developing a Visualization platform for VariantSpark

VariantSpark is a machine learning tool that can deal with high dimensional data and can identify top X features that individually or jointly influence the outcome with high sensitivity. Because of its availability on all major cloud computing platforms, its ease of use and its extensible and flexible algorithm, VariantSpark can be applied to diverse fields to generate insights. For example, In Genetics, VariantSpark can help in identifying the diseased gene, In Internet-of-things, VariantSpark can provide fresh insights into predictive markers about customers and behaviours, In supply chain management, VariantSpark can identify the specific component that contributes to an optimal outcome.

Currently, there exists no streamlined workflows that can generate reproducible visualizations from the results of VariantSpark which makes it difficult for beginners to understand the results of VariantSpark. In this project, we intend to create a serverless platform which accepts a VariantSpark output file and generates multiple different visualizations which can be customized by end-users.

The goal of this project will be to:

- Create a platform to generate a generic visualization from a variantSpark output file.
- Generate visualizations to map VariantSpark output to raw input data.
- Give end-users the ability to customize their visualizations

The student will:

- Study existing visualization techniques for tree based data and interactions data
- Use existing visualization software to best visualize the interactions identified by VariantSpark.
- Research and identify different formats and visualizations that can be generated from VariantSpark output.
- Assist in building a full-stack serverless application to generate the generic visualizations.

Relevant fields of expertise:

- Data visualization and analytics.
- Computer science and knowledge of at least one scripting language (Python, R, Bash, Perl, etc).
- Basic understanding of any full-stack development.

Supervisor:

- Yatish Jain (CSIRO, H&B, Yatish.Jain@csiro.au)
- Natalie Twine (CSIRO, H&B, Natalie.Twine@csiro.au)
- Denis Bauer (CSIRO, H&B, Denis.Bauer@csiro.au)

Location: North Ryde, Sydney (or remote)

Funding:
CSIRO offers to top-up stipend for this project: $300/week contribution towards living expenses

BSEP14: Investigating the role of viral integrations in cancer

Viruses are estimated to cause 10-15% of all cancers. Links between a number of viruses and cancer types have been discovered, with well-known examples being hepatitis B with liver cancer, and human papilloma virus with cervical cancer. We are currently investigating the role of viral integrations in cancer, and are interested in where and how these viruses might integrate in the first place. Can we identify viral integration sites in publicly available data from patients with tumours? Are there particular signals for viral integration in the human genome? Are the mechanisms different for tumour and healthy tissue? In this project you will work towards answering some of these questions, gaining insight into the mechanisms whereby viruses might cause certain types of cancer.

The student will:

- Develop methods to identify viral integration sites in sequence data
- Develop methods to identify the mechanisms of viral integration
- Correlate viral integration data with phenotypic data

Relevant fields of study:

- Bioinformatics
- Scripting programming (Python, R, Bash, Perl, etc)
- Genetics

Supervisor:

- Suzanne Scott (CSIRO, H&B, suzanne.scott@csiro.au)
- Laurence Wilson (CSIRO, H&B, laurence.wilson@csiro.au)
- Denis Bauer (CSIRO, H&B, denis.bauer@csiro.au)

Location: North Ryde, Sydney (or remote)

Funding:
CSIRO offers to top-up stipend for this project: $300/week contribution towards living expenses

BSEP15: Apply genome-scale machine-learning to “big” datasets in other disciplines

The automatic collection of information such as from internet of things (IoT) devices causes datesets to grow rapidly in all disciplines. “Wide” data, that is millions of datapoints per sample, was originally encountered in the genomic discipline. Here, millions of genomic variants describe a patient, and with cohort sizes ranging in the 10 thousands, analysis tasks would easily reach several trillion datapoints. The analysis tool, VariantSpark, was developed to perform machine learning on such large, high-dimensional datasets. Be part of the team that applies the smarts of genome research to other disciplines.

The student will perform:

- Identify non-life-science dataset to apply VariantSpark to
- Compete in Kaggle competitions and make a name in the Machine Learning community
- Contribute to changes in VariantSpark to make it application agnostic.

Relevant fields of study:

- Computer science
- Machine learning
- Statistical analysis
- Programming (Python/Scala)

Supervisor
Denis Bauer
Email: Denis.Bauer@csiro.au
Location: Sydney (or remote)

Funding:
CSIRO offers a top-up stipend for this project: $300/week contribution towards living expenses

CONTACT US

t 1300 363 400
+61 3 9545 2176
e enquiries@csiro.au
w www.csiro.au

AT CSIRO WE SHAPE THE FUTURE

We do this by using science to solve real issues. Our research makes a difference to industry, people and the planet.

As Australia’s national science agency we’ve been pushing the edge of what’s possible for over 85 years. Today we have more than 5,000 talented people working out of 50-plus centres in Australia and internationally. Our people work closely with industry and communities to leave a lasting legacy. Collectively, our innovation and excellence places us in the top ten applied research agencies in the world.

WE ASK, WE SEEK AND WE SOLVE

FOR FURTHER INFORMATION

Health and Biosecurity
Denis Bauer
t +61 2 9325 3174
e Denis.Bauer@csiro.au
w https://aehrc.com