HEALTH AND BIOSECURITY
www.csiro.au

# Bioinformatics Student Exchange Program

CSIRO – Germany – China

# 1 BSEP 2016

Australia was featured in a recent Nature article stating 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 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 students from Germany and China 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 German and Chinese 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.

| University | Contact Person |
|---|---|
| **FU Berlin** | Prof. Dr. Annalisa Marsico<br>RNA Bioinformatik<br>Phone: +49 30 8413 1843<br>Fax: +49 30 8413 1960<br>Email: Annalisa.Marsico@fu-berlin.de |
| **Tübingen** | Dr. Julian Heinrich<br>Applied Bioinformatics Group<br>Email: heinrich@informatik.uni-tuebingen.de |
| **CSIRO** | Dr. Denis Bauer<br>Transformational Bioinformatics, eHealth, CSIRO<br>Phone: +61 2 9325 3174<br>Email: denis.bauer@csiro.au |



## 1.1 Key dates

| Date | |
|---|---|
| **July** | CSIRO calls for project proposals |
| **31st August** | Program Booklet sent to the Universities |
| **Early October 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 |

## 1.2 Funding

Students are encouraged to apply for funding

## 1.3 Germany

### PROMOS

German funding through PROMOS (Deadline Early October to early November), which will cover

- from 300 to 500 EUR per month and / 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 FIT- Internationale Forschungsaufenthalte in der Informationstechnologie für Masterstudierende, which can be applied for at any stage, with notification of success within 3 Months (recommended application date no later than October)

- 875 EUR per months
- contributions to travel costs
- contributions to insurances

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

## 1.4 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.

VISA:
https://www.border.gov.au/Trav/Visa-1/402-

Information about the VISA subclass 402 Trainee and Research.

Address where to send the application:
https://www.border.gov.au/Lega/Lega/Help/Location/australia
Tasmania-Hobart office

Health insurance:
http://www.health.gov.au/internet/main/Publishing.nsf/Content/Overseas+Student+Health+Cover+FAQ-1#insurersofferoshc
e.g. the BUPA caters for VISA subclass 402.

# 1.5 Experience Report from 2015

I want to share with you some of the fantastic experiences I had during the Bioinformatics Student Exchange Program (BSEP) as well as give you some advice so you'll get the most out of your exchange.

I lived in Australia for months (mid May until end of November 2016) and chose a project that was based in Sydney. Sydney is the best city for an ambitious and nature loving student. The city has everything you need for a good night: pubs, restaurants and lot of clubs. Sydney can be best described like a combination of New York city and London. Really fascinating is the fact that you can take the bus from the city center and in an hour you are at the beaches, relaxing and sun bathing. Not enough? There are a lot of bush walk tours around Sydney. Maybe join a group via meetup.com, you will get some astonishing impressions. By taking the train you have also the opportunity to travel to the Blue mountains. The Blue mountains offers one of the most spectacular views you can imagine. Katoomba is well-known for the Three Sister and a good point to start. You see, there is plenty of stuff for you to discover.

But do not get lost in all these new things. You still have to work :). But I can tell you, you will even enjoy this part of your trip. The team at CSIRO is very friendly and open-minded. They find a solution for any problem you have and provide ideas and inspiration to evolve your work, still leaving enough room for your creative mind and expertise. The work environment at CSIRO is one of the best I have seen so far. Everybody is very friendly and talkative. When you need a little break, then you can use the pool table, table tennis and some darts in the basement. At lunch time there is even a little bush walk group to innervate your lazy legs.

But before enjoying all of this you need to plan your trip. I have two recommendations for any applicant. First apply for a funding as soon as possible. I took PROMOS which was less labour intensive than other fundings. For the PROMOS you have to find a supervisor (professor). Next you have to confirm your language skills by taking a language test. I recommend to do the DAAD or the language test from the FU because they do not involve any fees, in contrast to TOEFL and others.

After you applied for the PROMOS, the next big thing you have to tackle is the application for the VISA. Do that as soon as possible. Keep in mind that the Australian government needs at least 3 months to process the VISA application and costs around $300. This was for me a big problem because I planned my trip a bit too early and everything got delayed for a few months. But this is not a disaster. CSIRO is very flexible. You have the chance to shift your time and even extend it. Yet prepare everything beforehand for the VISA application: have a valid passport, get a financial report of your bank, find a health insurance and write an English CV. If you have finished the VISA application the hard part is over. The rest is relatively easy to organize and I would recommend to do that after you get the VISA. A bank account can be arranged when you are in Australia. I found a cheap accommodation via homestay.com, no serious advertising but for some help in that sense you can contact me (you can even use airbnb.com). You do not need a car, except you want to be a bit more flexible. For a smooth start I would recommend to enter Australia one or two weeks before you begin your work at CSIRO. You definitely need some time to get used to the new culture and do a bit of sight-seeing.

I definitely had an amazing time and I encourage you to get in touch with me if you want to know more.
Cheers,

Florian

# 1.6 Projects



| ID | Details |
|---|---|
| **BSEP01** | **Improving CRISPR/Cas9 genome editing strategies in a context of infectious diseases**<br><br>Genome editing technologies such as CRISPR/Cas9 are powerful and advanced techniques to interrogate the function of the genes.<br><br>The Burgio laboratory at the Australian National University is developing the genome editing technology to study the interaction between a pathogen (malaria, hospital acquired infections) and its host by generating mouse models of disease (cancer, neurological, immunity).<br><br>In collaboration with Denis Bauer, CSIRO Sydney, we aim to develop a tool to assess and predict the efficiency of a CRISPR/Cas9 guide RNA (gRNA) within a mammalian system in silico and to correlate the findings with biological assays in mouse zygotes and human cell lines. We also aim to develop a gene drive approach to eradicate infectious diseases in wild mouse or rat population. The student will develop a novel algorithm to predict the efficiency of the Cas9 cleavage given a specific gRNA sequence and a genomic position. The student will then assess this predictor tool with biological results in a context of human cell lines and mouse zygote and will adjust the algorithm accordingly. The second part of this project will be to model the spread of the 'gene drive' approach in a wild population. The student will develop an algorithm, given a specific gRNA and a gene target to predict and optimize the spread of a mutation in a wild mouse population and to predict the likelihood of the "off-targets" effect of the gene-drive mutation.<br><br>The student will be part of an important project to optimize the design of guide RNAs to generate mouse models of diseases and to better control or eradicate deleterious infectious diseases.<br><br>**The student will perform**<br><br>• Genome sequence analysis to identify potential targets of the gRNA (using SeqAn, e.g. Triplexator)<br>• Optimize the binding motif model using optimization techniques<br>• Adjust the search algorithm to handle sequence uncertainty from wild populations<br>• Use functional annotation (e.g. ENCODE) to predict effects for off-targets.<br><br>**Relevant field/s of study** |

| | |
|---|---|
| | • Proficiency with SeqAn for motif search<br>• C++ advanced<br>• Optimization techniques (Linear optimization Constrained programming, polynomial approximation)<br>• Basic understanding of the CRISPR/Cas9 system<br><br>**Supervisor**<br>Gaetan Burgio (ANU), Denis Bauer (CSIRO, Health&Biosecurity)<br>phone 0261259428 or email Gaetan.burgio@anu.edu.au<br>Location: Sydney NSW |
| **BSEP02** | ## Predicting exon splicing changes triggered by epigenetic profiles<br><br>Splicing transcripts is an essential and tightly regulated process whose interruption can cause disease like Familial dysautonomia (FD). While it is well established how DNA sequence variations can influence splicing events with recent publications in Science and Cell, less focus has been given to the role of DNA methylation despite evidence of its involvement. In preliminary work we have used RNA sequencing and Whole genome bisulphite sequencing to predict exon splicing using a deep neural network.<br><br>This project aims to use convolution neural network (Angermueller et al. Molecular Systems Biology, 2016) to provide a better insight into the complex interacting features of methylation and histone marks. We will be able to test the hypothesis that DNA methyaltion at histone marks or at the protein binding sites of CTCF affects chromatin folding and in turn leads to changed splicing events by clustering observed DNA methylation profiles with an unsupervised deep learning approach (Zheng et al. Nucleic Acids Research, 2015).<br><br>**The student will perform**<br><br>• Assemble the data sets in a form suitable for a learning algorithm<br>• apply various algorithms, modifying them as required for the problem<br>• interpret the results and write up the project<br><br>**Relevant field/s of study**<br><br>• R, Bash, Python<br>• Machine learning and statistical concepts.<br>• Understanding of metabolic, gene-gene interaction and regulatory networks<br><br>**Supervisor**<br>Robert Dunne (CSIRO, Digital61)<br>Rob.Dunne@csiro.au<br>Location: Sydney NSW |
| **BSEP03** | ## Linkage-directed prioritisation for disease variant detection<br><br>Even healthy individuals have hundreds of variants that disrupt the normal function of the resulting protein. Hence identifying causative variants by filtering for deleterious variants may sometimes not be sufficiently powerful. Inherited high impact deleterious variants amongst family members is observed clinically as a familial syndrome. Recently developed tools such as pVAAST (Nat Biotech, 2014) make use of shared chromosomal segments between related |

| | |
|---|---|
| | individuals to identify genetic variants that directly influence disease risk. These approaches extend the variant prioritization and case-control association features with linkage analysis methods specifically designed for sequence data. These models are broadly similar to traditional linkage analysis but capable of modeling de novo mutations and handling incomplete penetrance or locus heterogeneity.<br><br>This project proposes to apply pVAAST to existing whole genome and whole exome data of two disease cohorts, one from familial colorectal cancer and one from familial motor neuron disease patients. We also wish to compare with other approaches, including non-parametric linkage in Merlin, the Generalized Family-Based Association Test and methods for identifying shared chromosomal segments in distantly related individuals. This will require parsing VCF files into standard linkage/association file formats used by PLINK and Merlin. We propose to extend SeqAn to output these formats.<br><br>**The student will perform**<br>• Perform a literature review to identify all linkage-aware association predictors<br>• Apply pVAAST and related tools to two disease cohorts<br>• Develop a SeqAn-based parser for linkage file formats<br>• Compare pVAAST results to other linkage and family-based association test outputs<br>• Develop approaches for identifying distantly related individuals<br><br>**Relevant field/s of study**<br>• Knowledge of genetics<br>• Bioinformatics<br><br>**Supervisor**<br>Jason Ross (CSIRO, Health&Biosecurity)<br>phone +61 2 9490 5015 or email Jason.Ross@CSIRO.au<br>Location: Sydney NSW |
| **BSEP04** | ## Distribution of transponsable elements distributed in the Tiwi genomes<br><br>A large proportion of the human genome (approximately 45%) comprises of transposon and transposon-like repetitive DNA elements, although only a very small number remain active. They comprise a large number of families based on sequence comparison. Depending on their genomic location and activity, they can influence gene expression, individual phenotypic diversity and disease.<br><br>The McMorran group has generated 120 complete human genome sequences from the Tiwi people, generated as part of a study investigating causes of renal disease. This represents the first resource of its kind.<br><br>This project aims to conduct a bioinformatics characterisation of transposable elements in whole genome sequences of the Tiwi people. Elements will be identified, characterised and mapped for each individual. An assessment of common and unique elements that may influence gene expression and protein function will also be conducted.<br><br>**The student will perform**<br>• Whole genome sequence analysis identifying transposable elements and distinguish them against other repetitive elements using third party tools<br>• Utilize annotation information (ENCODE, UCSC) to characterise functional classes |

- Compare findings between 120 individuals to identify structural groups

**Relevant field/s of study**

- Proficiency with SeqAn for motif search
- C++ advanced
- statistical concepts

**Supervisor**

Brendan McMorran (ANU), Denis Bauer (CSIRO, Health&Biosecurity)

brendan.mcmorran@anu.edu.au

Location: Sydney NSW

**BSEP05**

## Medical image analysis of the retina with machine learning

The eye is the only part of the body where central nervous system tissue and microvasculature can be imaged directly and non-invasively. Modern imaging technology allows high-resolution, 3-dimensional imaging of the retina at the back of the eye.

Changes to the eye occur in numerous diseases and might enable better management and treatment for these conditions. CSIRO has collected sophisticated imaging data from eyes of participants with some of these conditions and is conducting image analysis and machine learning / statistical analysis to develop clinical tests.

Eye imaging in Alzheimer's disease is showing promise for early detection of this form of dementia, making early intervention a real possibility and contributing toward the development of a successful treatment.

Retinal imaging is also showing signs that it could be beneficial in monitoring vascular health in conditions like resistant hypertension, stroke and also in managing general health and wellbeing.

The aim of the project is to analyse existing ocular imaging data from patients and controls, including retinal photographs, optical coherence tomography (OCT) scans and pupil response. Machine learning approaches will be utilised to combine markers from different imaging modalities and also other data, for clinical classification purposes.

**The student will perform**

- Literature review (will be guided)
- Extract medical imaging parameters from fundus, OCT and pupil data
- Machine learning / statistical analysis
- Write report

**Relevant field/s of study**

- Medical/Algorithmic Bioinformatics
- Biocomputing
- Applied Mathematics
- Statistics
- Image analysis

**Supervisor**

| | |
|---|---|
| | Dr Shaun Frost<br>phone on (+61) 893336137 or email shaun.frost@csiro.au<br>Location: Perth, WA (Floreat) |
| **BSEP06** | ## Targeted eQTL Analysis of SNPs and expression markers for Alzheimer's disease.<br><br>Using a targeted approach to molecular pathways and eQTL analyses, the current project will investigate approximately 2000 SNPs and genes to ascertain their relationship with Alzheimer's disease. Implementing a novel genomic design, the candidate will use complex statistical methods to define the interrelationships between SNPs, gene expression and disease phenotype. The outcome would be to build a Shiny App that would take data from a set of SNPs, data from a set of gene expression, and data regarding disease phenotype, and perform analyses to define optimal sets of markers associated with outcome. The app will be adjustable to vary numbers of markers identified, and present graphics to represent network based associations. It will be applicable to anyone who wants to analyse their favourite set of SNPs, and customise the parameters for network analyses.<br><br>**The student will perform**<br><br>• The student will perform statistical analyses, Shiny App design, and preparation of a manuscript to showcase the app.<br><br>**Relevant field/s of study**<br><br>• Bioinformatics<br>• Biostatistics<br><br>**Supervisor**<br><br>James Doecke (CSIRO, Health&Biosecurity)<br>+617 3253 3697 or james.doecke@csiro.au<br>Location: Herston Brisbane |
| **BSEP07** | ## Aligning sequencing reads with Spark<br><br>The alignment of reads generated from high throughput sequencing can be done very efficiently using high-performance-compute clusters due to advanced algorithms, such as BWA, Yara and RazerS. However, the subsequent tasks in sequence data analysis are less driven by compute intensive tasks but rather require the ingestion of large quantities of data at the same time. The Hadoop/Spark platform caters for this scenario better than traditional high-performance-commute, hence prompting efforts to re-implement GATK, the widely accepted gold standard for variant calling. As alignment task can also be reformulated, this new software suite, will also contain a BWA re-implementation in spark.<br><br>This project will re-implement Yara or RazerS in Spark by designing the underlying parallelization strategy using the functional programing concepts in scala. The student will compare the performance and accuracy of the developed tool against other Spark-based read aligners such as SNAP and the by then newly released Spark-based BWA implementation.<br><br>**The student will perform**<br><br>• Re-implementing Yara/RazerS in Spark<br>• Testing the tool against other Spark-based aligners |

| | |
|---|---|
| | • Publish and document the software code<br><br>**Relevant field/s of study**<br><br>• Experience with high-throughput sequencing data analysis, mapping and variant calling<br>• Knowledge of parallelization strategies<br>• Ideally knowledge of scala and Spark<br><br>**Supervisor**<br>Denis Bauer (CSIRO, Health&Biosecurity)<br>+61 2 9325 3174 or Denis.Bauer@CSIRO.au<br>Location: Sydney, NSW |
| **BSEP08** | ## Improving target accuracy for genome engineering applications<br><br>Both the structure of the sgRNA and the chromatin environment influence the on-target efficacy of CRISPR-Cas9. We have developed a chromatin-aware CRISPR predictor, however this program relies on publicly available models for predicting the effectiveness of a given sgRNA. The aim of this project is to develop a new method for analysing sgRNAs and predicting their activity and incorporate this into our full model.<br><br>**The student will perform**<br><br>• Identification of the best method for building the model (SVM, linear model, etc)<br>• Construction of model using publically available datasets<br>• Comparison of model with other publically available models<br>• Incorporation of the model into our CRISPR-predictor<br><br>**Relevant field/s of study**<br><br>• bioinformatics<br>• machine learning<br><br>**Supervisor**<br>Laurence Wilson (CSIRO, Health&Biosecurity)<br>+61 2 9325 3039 or Laurence.Wilson@CSIRO.au<br>Location: Sydney, NSW |
| **BSEP09** | ## Comprehensive characterization of non-conventional transcripts from human cells<br><br>Identification of unconventional RNA molecules such as fusion and read-through transcripts is a challenging task in the effort to comprehensively characterize the functional readout of human genome. RNA sequencing by Paired-End diTag (RNA-PET, previously known as GIS-PET) analysis possesses a unique capability to accurately and efficiently characterize the 5' and 3' ends of DNA fragments, which may have either normal or unusual compositions. This unique nature of RNA-PET analysis makes it an ideal tool for uncovering unconventional transcripts produced from the human genome. Previous studies have reported the identification of fusion transcripts by RNA-PET from human cancer cell lines. These discovered fusion transcripts directly corresponded to the gene fusion resulted by |

| | |
|---|---|
| | genome structure variation, which directly leads to disease. Similar read-through transcripts were reported in mammalian genomes, however, a comprehensive characterization of read-through transcripts and their function is still far from complete. High quality RNA-PET data from various health and disease cell types are available from the ENCODE project. This project aims to utilize these RNA-PET data to characterize and compare read-through transcripts from health and disease cells and the possible functional relationship to diseases.<br>**The student will perform**<br>• Design and implement novel algorithms to analyse RNA-PET<br>• Characterize transcripts from health and disease cells<br>• Write report summarizing findings<br>**Relevant field/s of study**<br>• Experience with high-throughput sequencing data analysis<br>• Statistical methodologies<br>**Supervisor**<br>Oscar Luo (CSIRO, Health&Biosecurity)<br>+61 2 9490 8989 or Oscar.Luo@CSIRO.au |
| **BSEP10** | SNP-aware graph-based Off-target finder<br>Genome engineering relies on target sites with minimal number of off-targets to avoid damage. Off-targets are typically identified by using traditional read mappers like bowtie to map the sequence of the target site to the genome allowing for a specific number of mismatches. This assumes perfect knowledge of the edited genome. However, the human as well as any other non-inbred population have a number of sequence variations (SNPs) occurring in the population. These SNPs can render some off-targets inactive or create new sites. It is hence important to take a variable genome into account when designing targets. We propose to construct a graph from a sequence library containing the locations of target-sites as well as variations thereof modulating known SNPs. This graph can hence be traversed for finding the optimal target-site with the fewest SNP modulated off-targets.<br>**The student will perform**<br>• Design and implement a novel graph algorithm for off-target search<br>• Characterize performance and power on publicly available data<br>**Relevant field/s of study**<br>• Experience in sequence analysis (SeqAn)<br>**Supervisor**<br>Denis Bauer (CSIRO, Health&Biosecurity)<br>+61 2 9325 3174 or Denis.Bauer@CSIRO.au<br>Location: Sydney, NSW |

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