Principal Supervisor:	**Reiko Tanaka**
Further Information	
Department:	Other, Bioengineering
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/3	Healers vs. non-healers: A systems biology approach for predicting the molecular dynamics of wound healing processes *Co-spvr Jacques Behmoaras*
Wound healing is a physiological process under strong genetic control, which determines the kinetics of repair mechanisms.Â This project aims to identify biomarkers and genetic factors that can differentiate between fast and slow-healers in a skin wound healing model. The student will apply statistical and machine learning analysis to the comprehensive data obtained from about 1400 genetically heterogeneous mice that underwent wound healing following ear punching. Through analysis of both genetic and plasma biochemistry measurements, we aim to identify the biomarkers, develop mathematical models to predict and stratify fast & slow-healers, and propose experimentally-verifiable hypotheses to enhance ourÂ understanding of wound healing processes.	
Project 2/3	Dynamic mechanisms behind heterogeneous developmental profiles of atopic dermatitis (AD) *Co-spvr Adnan Custovic*
This project brings together data from the large population-based birth cohort studies, with a mathematical in silico model of AD pathogenesis that describes the complex and dynamic interplay between skin barrier function, immune responses, and environmental stressors. The project will capitalise on the complex, rich and expanding datasets from five UK birth cohort studies, which will allow the student to identify developmental trajectories of AD from birth to early adulthood. The student will develop a mathematical model of AD pathogenesis that explains the heterogeneity in developmental profiles of eczema, by linking our previously published mathematical model of AD pathogenesis that describes the complex and dynamic interplay between skin barrier function, immune responses, and environmental stressors, with population-based data on >14â€…000 children with many repeat measures over time. We will achieve a mechanistic understanding of the heterogeneity in developmental profiles of eczema and inform stratified treatment and prevention strategies. Belgrave et al. Developmental Profiles of Eczema, Wheeze, and Rhinitis: Two Population-Based Birth Cohort Studies. PLoS Med. 2014;11(10):e1001748. Dominguez-Huttinger et al. Mathematical modeling of atopic dermatitis reveals â€œdouble-switchâ€ mechanisms underlying 4 common disease phenotypes. J Allergy Clin Immunol 2017, 139 (6) 1861â€“1872.e7.	
Project 3/3	Development of mathematical models to predict the eczema occurrence and treatment outcome using machine learning methods *Co-spvr Adnan Custovic*
Atopic dermatitis (AD) is one of the complex diseases with a large variation in the disease severity and responses to treatments among patients. Designing personalized treatment strategies for AD	

based on patient stratification, rather than the â€œone-fit-allâ€ treatments, is of high clinical relevance. Better prognoses could help stratify patients who are likely to respond to treatments favourably and choose appropriate, effective and personalised treatment strategies. This project aims to develop mathematical models that predict treatment outcomes by applying machine learning analysis to longitudinal data from previously published clinical studies. The model is expected to identify the biomarkers that can stratify the patients depending on the predicted efficacy of the treatments and also to predict the likely occurrence of eczema flares.

Contact Details	
Email: r.tanaka@imperial.ac.uk	**Tel:** x46374

Principal Supervisor:	**Virginia Fairclough/Derek Huntley**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,
Dates of absence of more than two weeks? ,	**Provision for supervision during primary supervisor's absence (if applicable):** Other Derek Huntley/Pietro Spanu
Project 1/1	The molecular basis for evolution of obligate biotrophy in plant disease *Co-spvr Pietro Spanu*

Plant and microbes commonly interact in nature; these interactions may lead to mutualistic symbioses or disease. In some cases, the interactions become so close that an obligate dependency evolves. These are then called obligate biotrophic fungi. Although these interactions are common, have been known for a long time and lead to some of the most prominent and devastating diseases of important staple crops (such as wheat and barley), it is not known WHAT THE MOLECULAR BASIS for this obligate relation is. Why we cannot grow these fungi on a Petri dish?

As a result of the last decadesâ€™ efforts in sequencing genomes of these fungi, it is clear that the main primary metabolic pathways are active in obligate biotrophs. In this project you will test the hypothesis that what is different here, is that the REGULATORY elements controlling the expression of genes encoding enzymes on primary metabolic pathways have mutated to become dependent on growth in a plant environment. Preliminary evidence pointed out that this may be the case. Now, with the availability of much broader data-sets, this hypothesis needs to be tested, using a deeper and more rigorously statistical analysis of the genome data available in the data bases.

Contact Details	
Email: virginia.fairclough08@imperial.ac.uk/d.huntley@imperial.ac.uk	**Tel:**

Principal Supervisor:	**Robert Endres**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, 3rd Floor SEC (CISBIO)
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):** Primary supervisor via email/Skype
Project 1/3	Deep learning for classifying hematopoietic stem cells under various conditions *Co-spvr Cristina Lo Celso*
Deep learning is a powerful machine learning approach to detect subtle details for classifying images and imaging data. Depending on student preferences there are two potential ways forward in this project: (1) In one option we will investigate hematopoietic stem cells as imaged in Lo Celso's lab, recorded under varying conditions. (2) In another option we will explore different ways of visualising and interpreting deep learning results, such as plotting trained networks with edge thickness reflecting rates or lower dimensional embedded manifolds.	
Project 2/3	Implement spectral method for solving reaction-advection-diffusion model for embryonic polarity development *Co-spvr*
Reaction-advection-diffusion models are tricky to solve numerically. Here we will first develop a spectral method as explained in (de la Hoz & Vadillo, J Comp Applied Math, 2014), and then apply it to solve the polarity development model for C. elegans zygotes in (Goehring et al. Science, 2011). Student needs very good mathematical and computational skills.	
Project 3/3	How do bacteria find their middle for cell division? Exploration of a minimal model of the Min system *Co-spvr na*
In order to find the middle for cell division, the Min system - which consists of three proteins, MinC, MinD and MinE - forms spatio-temporal oscillations similar to Turing patterns, restricting the FtsZ-ring assembly for cytokinesis to the lateral mid-cell region [1]. Current models, despite being relatively simple, can only be solved numerically or be simulated. Here, we will implement and explore an analytical minimal model of the reaction-diffusion model, including both molecules in the cytosol and in the membrane. A key aspect will be linear-stability analysis to quantify the onset of Min oscillations upon reaching a critical cell length. Using this approach, we will address cell division in growing normal cells and filamentous cells after stress has been removed [2]. Besides some programming skills, the project will require some math and the understanding of biophysical concepts. References: [1] Rowlett VW , Margolin W, Curr Biol 23: R553 (2013) [2] Wehrens et al., Curr Biol 28: 1 (2018)	
Contact Details	
Email: r.endres@imperial.ac.uk	**Tel:** 4 9537

Principal Supervisor:	**Prof Mike Sternberg**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, in group
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/4	Machine learning and protein structure prediction *Co-spvr Dr Lawrence Kelley*
The protein folding problem is one of the major unsolved problems in molecular biology. Central to any approach to solving this problem is the energy function used to assess a 3D model. In this project you will investigate the usefulness of a machine learning technique, Gaussian Mixture Models (GMMs), for this task. You will write programs that break known structures into fragments and use internal distances within these fragments to train a GMM. This GMM will then be tested on models produced by the Phyre protein modelling tool to determine its viability in discriminating good models from bad.	
Project 2/4	Machine learning gene signatures for patients with abnormal serum cholesterol levels. *Co-spvr Mr Luis Leal*
Defects in the lipid metabolism induce abnormal levels of serum cholesterol, increasing the risk of diabetes, atherosclerosis and cardiovascular diseases. In the UK, inherited lipid disorders such as Familial Hypercholesterolimia are known to have a prevalence of 1 in 500 people and most of the people with these conditions are undiagnosed. Therefore, characterisation of patients with extreme lipid profiles will give them prioritized treatment to decrease the risk of related diseases. In this project we aim at using machine learning and network-based methods to improve high cholesterol prognosis. We will explore the formulation of prognostic gene signatures by integrating patient clinical with genetic data (e.g, serum lipid levels, Single Nucleotide Variants (SNVs)). Sets of prioritised variants will be obtained using regression models and matrix factorisation methods, followed by the implementation Random Forests algorithms to classify patients. We will also analyse the protein-protein interactions that could be affected in the lipid metabolism. Coding skills in R are desired. Reference: 1). Okser S., et al. 2014. Regularized Machine Learning in the Genetic Prediction of Complex Traits. PLOS Genetics. 2) Johnson L. et al. 2015. An Examination of the Relationship between Lipid Levels and Associated Genetic Markers across Racial/Ethnic Populations in the Multi-Ethnic Study of Atherosclerosis. PLOS One.	
Project 3/4	Application of Molecular Dynamics to protein/protein docking *Co-spvr Mr Tarun Khanna*
Protein undergo conformational change when they dock. The aim of the project is to assess molecular dynamics starting from the unbound can generate a structure that is sufficiently close t the bound that it assists in protein docking. We will use an established benchmark of protein complexes with their respective unbound molecules. Docking servers will then be used to test of the predicted bound conformation enhances protein docking.	
Project 4/4	Structure-based evaluation of the effect of missense variants at protein interfaces *Co-spvr Mr Tarun Khanna*

We recently have developed a program 3DVAR that assess the structural effect of a mis-sense mutation in a protein tertiary structure. A list of altered conformational features are identified. We have shown that many disease associated variants occur at protein interfaces (1). We will use known (and possibly predicted) protein complexes and evaluate which feature in 3DVar remains effective when assessing the structural effect at a protein/protein interface.

(1) David, A., Razali, R., Wass, M. N., & Sternberg, M. J. (2012). Proteinâ€“protein interaction sites are hot spots for diseaseâ€associated nonsynonymous SNPs. Human mutation, 33(2), 359-363.

Contact Details

Email: m.sternberg@imperial.ac.uk

Tel: 020 7594 5212

Principal Supervisor:	**Giorgio Gilestro**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, 4th floor SEC
Dates of absence of more than two weeks? Dates, 1/8 - 23/8	**Provision for supervision during primary supervisor's absence (if applicable):** Primary supervisor via email/Skype
Project 1/1	A web app to convert manuscripts to publication quality PDFs *Co-spvr G. Gilestro*

The publishing industry in the biomedical field has recently been taken by a revolutionary storm: the pre-print movement. The pre-print server bioRxiv (http://www.biorxiv.org) born on the model of the physic's big brother arXiv, has grown almost exponentially in the past two years (see https://3spxpi1radr22mzge33bla91-wpengine.netdna-ssl.com/wp-content/uploads/2017/04/Preprint-Growth-in-Life-Sciences.jpg) attracting funding from the Zuckerberg foundation.

One bottleneck of the system remains the actual output of manuscripts submitted. Most authors would simply upload a copy of their manuscript as saved directly from the word processor of choice, without any kind of formatting to increase aesthetics and readability. This, I believe, is hindering the growth and adoption of the system.

The goal of this project will be to build a web app able to transform a manuscript into a nicely formatted PDF. The user should be able to upload their work as .doc, .rtf or .odt and through a minimal UI being able to output a nicely formatted PDF with multiple columns layout and in-text figures. From the technical point of view, the project entails:

- principles of NPL and ML to automatically characterised which part of the text is which (title, abstract, main text, references, etc)
- understanding and implementing principles of UI, including nodeJS or similar to handle formatting in a WYSIWYG fashion

The project has the potential to develop beyond the MSc.

Contact Details	
Email: g.gilestro@imperial.ac.uk	**Tel:** 2075945443

Principal Supervisor:	**Vahid Shahrezaei**
Further Information	
Department:	Mathematics,
Location: South Kensington,	**Facilities:** Desk only. State location in other,Other, SEC level 3
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/1	Identifying models of stochastic gene expression using single cell data *Co-spvr Anthony Bowman*
Gene expression is a stochastic process due to randomness in the timing of biochemical reactions. Single cell methods can capture this phenotypic variability. In this project we use Approximate Bayesian Computation to explore how to infer model parameters and structure from such data. In particular, we explore how not using the correct model can be misleading. One example is fitting models that ignore cell growth and division to data that is from growing and dividing cells. Familiarity with Julia programming language would be useful for this project.	
Contact Details	
Email: v.shahrezaei@imperial.ac.uk	**Tel:**

Principal Supervisor:	**Kirsten McEwen**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, 417 SEC
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):** Co Supervisor, listed with project
Project 1/1	Stochastic modelling of gene transcription in healthy and diseased states *Co-spvr David Schnoerr*
The advent of single-cell biology has opened new avenues for exploring gene regulation. Molecular species can be quantified at the single-cell level and we now know that gene transcription occurs in bursts, switching between inactive and active states. However, the identification of the underlying mechanisms from such data poses significant mathematical and statistical challenges. This project will extend previous work to explore the impact of gene bursting using stochastic models of transcription. Using experimental data of RNA from single cells, Approximate Bayesian Computation will be employed to infer parameters and identify suitable models. This method will then be extended to the generalised method of moments, which will allow analysis of hundreds of data sets. The developed methods will be applied to compare transcriptional bursting under healthy versus cancerous states to identify disrupted mechanisms of gene regulation.	
Contact Details	
Email: kirsten.mcewen@imperial.ac.uk	**Tel:**

Principal Supervisor:	**Clive Hoggart**
Further Information	
Department:	Medicine,
Location: St Mary's,	**Facilities:** Desk only. State location in other,Other, 2nd floor, Medical School
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/3	Estimating the heritability of meningococcal infection *Co-spvr Myrsini Kaforou*

Meningococcal disease (MD) is a life threatening infectious disease which further caries the risk of other severe complications such as skin grafts, amputations and neurological impairment. Genetic factors are known to influence both occurrence and severity of the disease. Studies of affected families have identified rare variants (minor allele frequency (MAF) < 1%) with large effect sizes conferring risk to MD and a genome-wide association study (GWAS) identified common variants (MAF~20%) in the region of the complement factor H gene (CFH) but with a substantially smaller effect size. Subsequent GWASs for MD have confirmed the CFH region but have not identified any further susceptibility loci at genome-wide significance.

We now have three case control MD GWAS data sets totalling 1,000 cases and 8,000 controls. The aim of this project is to utilise this data to estimate genome-wide heritability of MD and hence the residual heritability not explained by the CFH region. There are a variety of published software tools that can be used. These tools allow the heritability attributable to variants within a given minor allele frequency range to be calculated, thus we can estimate the proportion of missing heritability attributable to common variants that can be detected by increasing GWAS sample sizes and rare variants that can be detected by studies of affected families.

These methods can be extended to study co-heritability of multiple phenotypes, co-heritability is a measure of the contribution of common genetic effects common across phenotypes. The project can be extended to estimate the co-heritability of MD and other infectious and inflammatory diseases.

GWAS data sets are typically very large. Therefore this project will require the student to use Imperial's high performance cluster (hpc) for some analyses. The majority of post processing can be done on a desktop in R or python.

Further reading:

The initial genome-wide association study for susceptibility to meningococcal infection upon which this work builds was published in Nature Genetics.

Davila et al. (2010) Nature Genetics 42, 772â€“776

Methods to estimate heritability:

LDAK

Speed et al (2017). 49, 986-992. Nature Genetics Reevaluation of SNP heritability in complex human traits.

http://dougspeed.com/ldak/

GCTA

Yang et al (2010). Method for estimating the variance explained by all SNPs (GREML method) with its application in human height. Nat Genet. 42(7): 565-9.

http://cnsgenomics.com/software/gcta/#Overview

LDScore

Brendan et al (2015). LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies. Nat Genet. 47(3): 291â€“295.

Project 2/3	Developing a proteomic signature to distinguish bacterial from viral infection *Co-spvr Dr Shea Hamilton*

Due to the difficulty in distinguishing bacterial from viral infection based on clinical features, many children are misdiagnosed and receive unnecessary antibiotic treatment, while bacterial infection is missed in others. To identify protein biomarkers to distinguish bacterial from viral infection, we have analysed 200 serum samples by SELDI-TOF mass spectrometry. The serum samples were analysed on three different surface chemistries including cationic, anionic and IMAC arrays to enrich for potential biomarkers. The advantage of mass-spectrometry is that it is hypothesis free capturing all proteins in a sample. The primary aim of this project will be to evaluate protein levels that are differentially expressed between bacterial, viral and healthy controls and determine which combination of proteins form the best signature.
A variety of machine learning prediction tools such as the LASSO will be used to estimate optimal prediction models. The project can also explore:
1. whether the signature can be further refined by incorporating clinical data
2. the performance of the signature by type of bacterial or viral infection
3. an additional signature to differentiate gram-positive and gram-negative bacterial infections which would aid subsequent clinical treatment
References
LASSO:
Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1â€“22.
See also software for its implementation
https://cran.r-project.org/web/packages/glmnet/index.html
Publication from our group of a RNA signature derived from the same group of patients.
Herberg et al (2016). Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children. JAMA. 23-30;316(8):835-45.

Project 3/3	Identifying mechanisms of tolerance or progression to severe infectious diseases from whole blood RNA-seq *Co-spvr Dr Aubrey Cunnington*

Tolerance is defined as the ability of a host to maintain health in the presence of a defined pathogen load, and specific tolerance mechanisms have been characterized in experimental model systems. This has attracted considerable interest because therapeutically tractable tolerance mechanisms could transform management of infectious diseases. However, identification of tolerance mechanisms in humans has been challenging because pathogen load is often unknown and patients present at different stages of infection. We have recently performed whole blood RNA-seq on large numbers of subjects with two different infectious diseases in which pathogen load has also been measured â€“ malaria and meningococcal disease. This project will aim to integrate pathogen load data with RNA-seq data to identify transciptomic signatures of tolerance which are common across two important human diseases. An extension of this project will be to use these datasets and additional data from around 1000 patients with a variety of infections to identify transcriptional trajectories which lead to severe illness, and using a fate-mapping approach, identify biological markers and determinants of branch points in these trajectories.
References:
1. Disease tolerance and immunity in host protection against infection. Soares MP, Teixeira L, Moita LF. Nat Rev Immunol. 2017 Feb;17(2):83-96. doi: 10.1038/nri.2016.136
2. Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. Hyun Jae Lee, Michael Walther, Athina Georgiadou, Davis

Nwakanma, Lindsay B. Stewart, Michael Levin, Thomas D. Otto, David J. Conway, Lachlan J. Coin, Aubrey J. Cunnington doi: https://doi.org/10.1101/193631

Contact Details

Email: c.hoggart@ic.ac.uk	**Tel:** 020 7594 3915

Principal Supervisor:	**Derek Huntley**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, 301 SEC
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/2	Identification of neuroblastoma differentiation associated biomarkers by RNA-seq analysis *Co-spvr Dr Ximena Montano UCL, Great Ormond Street Institute of Child Health*
Neuroblastomas are the most frequent childhood extracranial neoplasms and represent up to 10% of all childhood cancers worldwide. Close to 90% of cases are diagnosed before 5 years of age and 30% of these arise within the first 12 months. Prognosis partly depends upon patient age and tumour characteristics. Good-prognosis neuroblastomas (differentiated, favourable cytogenetics), express high levels of the nerve growth factor (NGF) receptor tyrosine kinase trkA, whereas poor-prognosis tumours (poorly differentiated, with chromosomal abnormalities and amplified MYCN), express the neurotrophin receptor tyrosine kinase trkB. On NGF-stimulation, trkA is activated by tyrosine-phosphorylation and outcome is cell/tissue-dependent, and can signal for cell differentiation. Importantly, incidence of mutation of the tumour suppressor p53 in neuroblastomas is 5% to10%. This project test the hypothesis that comparative RNA sequence analysis of neuroblastoma cells expressing this p53 dependent trkA-activation versus not activated trkA, will identify potential differentiation biomarkers. The analysis will identify differentially expressed isoforms and miRNAs, and reveal possible gene-fusion transcripts that will be either p53 targets unrelated to trkA and differentiation, or deregulated as a result of p53-dependent trkA-Y674/Y675 phosphorylation. Overall, the identification of key differentiation associated RNAs will facilitate the discovery of potential biomarkers and provide information for therapies that will promote targeted differentiation of neuroblastoma cells.	
Project 2/2	Identification of neuroblastoma differentiation associated biomarkers by proteomic analysis *Co-spvr Dr Ximena Montano UCL, Great Ormond Street Institute of Child Health*
This project will determine the differential expression between the proteomic content of the neuroblastoma cell lines used in the RNA-seq project. This analysis will identify differentially expressed proteins that will be either p53 targets unrelated to trkA and differentiation, or proteins deregulated as a result of p53-dependent trkA-Y674/Y675 phosphorylation via SHP-1 repression. We will identify proteins in which the concentration has increased/decreased on cells and compare to control cells (cells transfected with trkA, mutant-trkA, tsp53 and non-transfected cells). Overall, the identification of key differentiation associated proteins will facilitate the discovery of potential drugable biomarkers and provide information for therapies that will promote targeted differentiation of neuroblastoma cells.	
Contact Details	
Email: d.huntley@imperial.ac.uk	**Tel:** 47478

Principal Supervisor:	**Antonio J. Berlanga-Taylor**
Further Information	
Department:	Medicine,
Location: St Mary's,	**Facilities:** Other, Epidemiology and Biostatistics provides access to communal spaces in the computer, MSc rooms and desks available on a day to day basis. The library is also available. There are no permanent spaces unfortunately.
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):** Co Supervisor, listed with project
Project 1/3	Genetic basis of bacterial antibiotic resistance *Co-spvr Dr. Huw Williams and Dr. Jake Bundy*

Genetic basis of bacterial antibiotic resistance

Background

Antibiotic resistance poses a global threat to humans (Holmes et al., 2016; Laxminarayan et al., 2016). Understanding the genetic basis of bacterial adaptation to antibiotic treatment is a crucial area which can provide important biological insight and potential therapeutic and preventive targets (Blair et al., 2015; Bush et al., 2011). Genome-wide association studies (GWAS) have emerged as a powerful method to understand the genetic basis of disease in humans (WTCCC, 2007) and are gaining traction in bacterial research (Power et al., 2017). Bacterial GWAS is a novel field however and faces specific challenges. Previous studies have applied tools designed for human genetic association studies but these usually do not translate well because of homologous recombination, clonality and population structure in bacteria (Power et al., 2017). Bacteria specific methods are rapidly emerging however (Collins et al., 2018; Earle et al., 2016).

Pseudomonas aeruginosa is an important human pathogen considered a critical priority in terms of antibiotic development (Morita et al., 2014; Tacconelli, 2017). Cystic fibrosis is a devastating disease (Davis et al., 1996) with patients almost invariably developing lung infection from P. aeruginosa in early adulthood (Crull et al., 2018; Folkesson et al., 2012; Ramsay et al., 2017).

In this study we propose to understand the genetic basis of antibiotic resistance in P. aeruginosa derived from patients with cystic fibrosis using bacterial GWAS approaches. We have whole genome sequenced more than 90 strains from sputum of patients suffering cystic fibrosis and have access to several phenotypic assays from the same strains, including metabolomics readouts with hundreds to thousands of measurements.

Study question

- What is the genetic basis of antibiotic resistance in cystic fibrosis P. aeruginosa?

Study design

- Bacterial genome-wide association study

Aims

The overall aim of this study is to understand the underlying bacterial genetic adaptations that occur in long term infections in humans. Specific objectives include:

- To identify genetic variants which increase bacterial resistance to clinically relevant antibiotics (we currently have assays for five antibiotics plus other bacterial phenotypes of relevance to pathogenicity)
- To build a bioinformatics pipeline for bacterial GWAS. This is a multi-step process for which many tools already exist. The key element is to create an installable, re-usable, extensible pipeline with excellent code documentation (see below for further details on the specific tools and concepts that will be used)

Proposed methods and access to data

Drs. Jake Bundy and Huw Williams (Imperial College London) have generated a wealth of phenotypic information for more than 90 strains of P. aeruginosa sampled from sputum of patients suffering cystic fibrosis. The patient cohort spans more than a decade of investigation and serves as an excellent natural model of bacterial evolution. We will use whole genome sequences as well as bacterial assays for antibiotic resistance, amongst other data. The main analytical method is bacterial GWAS (Collins et al., 2018; Earle et al., 2016). We may also apply quantitative trait loci tools for genotype-phenotype associations in the context of thousands of measured features. To achieve the study objectives several analyses must be undertaken. These include de novo assembly, pan and core genome analysis, identification of recombination hotspots, mapping and variant calling, variant annotation, amongst several others.

Development of tools for computational biology

We will follow principles in reproducibility and best practice for computational biology (see for example (Noble, 2009; Sandve et al., 2013)). Specifically, the student will work with the following:

Reproducibility concepts and best practice implementation (Wilson et al., 2014; Wilson et al., 2017)

Use of Ruffus (Goodstadt, 2010) as a pipeline tool and CGAT tools (Sims et al., 2014) for support

Python programming and packaging restructuredText and Sphinx for reporting

Travis and tox for testing, Conda and Docker for management and development, GitHub for version control

Skills required

In order to successfully complete the project, we expect that the student already has:

Excellent skills in basic bioinformatics, statistics and general scientific method

Experience with at least one statistical and/or programming language (R and/or Python preferably)

Strong interest in genomics, epidemiology and human disease

General problem-solving and self-directed learning within a supportive environment

Outcomes

Students will gain knowledge and experience in:

Biological and statistical analysis and experience of bacterial genomics and human disease related to antibiotic resistance

Understanding of statistical methods in high-throughput analyses

Deep understanding and experience in â€œbest-practiceâ€ approaches in research software engineering for reproducibility

This project is suitable for students with backgrounds in either biomedical or numerical areas with experience in computer programming and general bioinformatics. We expect students to have a desire to learn topic specific issues (biomedical) and develop research software that is usable by others. Deriving biological insights as well as re-usable computing tools are key outcomes of this project.

There are further questions that will follow from this study, which may be suitable for someone interested in developing knowledge and skills in this area.

Supervisor:

Dr. Antonio J Berlanga-Taylor MBBS, MSc, DPhil

Co-supervisors:

Dr. Jake Bundy, Dr. Huw Williams

Key references

Blair, J. M., Webber, M. A., Baylay, A. J., Ogbolu, D. O. and Piddock, L. J. (2015). "Molecular mechanisms of antibiotic resistance." Nat Rev Microbiol 13(1): 42-51.

Bush, K., Courvalin, P., Dantas, G., Davies, J., Eisenstein, B., Huovinen, P., Jacoby, G. A., et al. (2011). "Tackling antibiotic resistance." Nat Rev Microbiol 9(12): 894-896.

Collins, C., Didelot, X., Earle, S. G., Wu, C. H., Charlesworth, J., Stoesser, N., Gordon, N. C., et al. (2018). "A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination
Identifying lineage effects when controlling for population structure improves power in bacterial association studies." PLoS Comput Biol 14(2): e1005958.

Crull, M. R., Somayaji, R., Ramos, K. J., Caldwell, E., Mayer-Hamblett, N., Aitken, M. L., Nichols, D. P., et al. (2018). "Changing rates of chronic Pseudomonas aeruginosa infections in cystic fibrosis: a population-based cohort study." Clin Infect Dis.

Davis, P. B., Drumm, M. and Konstan, M. W. (1996). "Cystic fibrosis." Am J Respir Crit Care Med 154(5): 1229-1256.

Earle, S. G., Wu, C. H., Charlesworth, J., Stoesser, N., Gordon, N. C., Walker, T. M., Spencer, C. C. A., et al. (2016). "Identifying lineage effects when controlling for population structure improves power in bacterial association studies." Nat Microbiol 1: 16041.

Folkesson, A., Jelsbak, L., Yang, L., Johansen, H. K., Ciofu, O., Hoiby, N. and Molin, S. (2012). "Adaptation of Pseudomonas aeruginosa to the cystic fibrosis airway: an evolutionary perspective." Nat Rev Microbiol 10(12): 841-851.

Goodstadt, L. (2010). "Ruffus: a lightweight Python library for computational pipelines." Bioinformatics 26(21): 2778-2779.

Holmes, A. H., Moore, L. S., Sundsfjord, A., Steinbakk, M., Regmi, S., Karkey, A., Guerin, P. J., et al. (2016). "Understanding the mechanisms and drivers of antimicrobial resistance." Lancet 387(10014): 176-187.

Laxminarayan, R., Matsoso, P., Pant, S., Brower, C., Rottingen, J. A., Klugman, K. and Davies, S. (2016). "Access to effective antimicrobials: a worldwide challenge." Lancet 387(10014): 168-175.

Morita, Y., Tomida, J. and Kawamura, Y. (2014). "Responses of Pseudomonas aeruginosa to antimicrobials." Front Microbiol 4: 422.

Noble, W. S. (2009). "A quick guide to organizing computational biology projects." PLoS Comput Biol 5(7): e1000424.

Power, R. A., Parkhill, J. and de Oliveira, T. (2017). "Microbial genome-wide association studies: lessons from human GWAS." Nat Rev Genet 18(1): 41-50.

Ramsay, K. A., Sandhu, H., Geake, J. B., Ballard, E., O'Rourke, P., Wainwright, C. E., Reid, D. W., et al. (2017). "The changing prevalence of pulmonary infection in adults with cystic fibrosis: A longitudinal analysis." J Cyst Fibros 16(1): 70-77.

Sandve, G. K., Nekrutenko, A., Taylor, J. and Hovig, E. (2013). "Ten simple rules for reproducible computational research." PLoS Comput Biol 9(10): e1003285.

Sims, D., Ilott, N. E., Sansom, S. N., Sudbery, I. M., Johnson, J. S., Fawcett, K. A., Berlanga-Taylor, A. J., et al. (2014). "CGAT: computational genomics analysis toolkit." Bioinformatics.

Tacconelli, E. (2017). "Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics." Essential medicines and health products, WHO.

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H., et al. (2014). "Best practices for scientific computing." PLoS Biol 12(1): e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T. K. (2017). "Good enough practices in scientific computing." PLoS Comput Biol 13(6): e1005510.

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447(7145): 661-678.

Project 2/3	Genome-wide association analysis of multi-morbidity in UK Biobank participants *Co-spvr Dr. Ioanna Tzoulaki and Dr. Deborah Schneider-Luftman*

Genome-wide association analysis of multi-morbidity in UK Biobank participants

Background

Multi-morbidity is generally defined as the co-occurrence of disease in a single individual (Academy of Medical Sciences, 2018; Valderas et al., 2009). Multiple chronic diseases are the main problem for most healthcare systems in the world (WHO, 2014). However, most healthcare, medical education and research systems follow reductionist approaches and are geared towards care of individuals with individual diseases (Barnett et al., 2012). In the US Medicare healthcare system for example, 65% of beneficiaries aged 65 and older had multiple chronic conditions (Wolff et al., 2002). In Scotland, 23% of patients were multi-morbid (Barnett et al., 2012). Multi-morbidity is highly heterogeneous and inherently difficult to study however (Nunes et al., 2016). It currently has no consensus definition but can be thought of as the presence of more than one disease in an individual where none of the conditions can be considered causal of others (Lefevre et al., 2014) and where the conditions are not due to frailty (i.e. vulnerability due to exhaustion (Villacampa-Fernandez et al., 2017)). Multi-morbidity generally increases with deprivation and age with non-communicable disease risk factors such as smoking and obesity likely being substantial causal contributors (Barnett et al., 2012). Genetic determinants have not been studied systematically in large cohorts however and may play an important role. Recent studies indicate that multi-morbidity of certain conditions may be due to shared familial risk and/or genetic factors (van Hecke et al., 2017). Thus, a subset of cases of multi-morbidity may be due to shared disease pathways, genetic or other, which may have common underlying pathogenesis.

Here we hypothesise that multi-morbidity can be treated as a complex trait with a common biological basis in certain individuals. Complex traits do not follow Mendelian inheritance patterns, are very likely caused or influenced by a multitude of genetic variants and express themselves in a variety of ways (phenotypes) (Boyle et al., 2017). The genetic basis to complex disease has been robustly studied in recent years using genome-wide association studies (GWAS) (Lander et al., 1994; WTCCC, 2007). GWAS approaches determine the association between genotypes and phenotypes in order to understand the genomic basis of complex traits (Anderson et al., 2010; Wang et al., 2018). In this study we propose to carry out a GWAS of multi-morbidity in UK Biobank participants, a cohort of 0.5 million individuals (Sudlow et al., 2015).

Study questions

- Can we identify subsets of individuals who share common environmental and/or biological pathways?
- Can we identify a genetic basis of multi-morbidity?

â€¢ Study design

- Cluster analysis and genome-wide association study of UK Biobank participants

Aims

The overall aim of the study is to identify genetic determinants associated with multi-morbidity.

Specific objectives include:

- To carry out cluster analysis of individuals suffering more than one condition
- To carry out a GWAS of multi-morbidity in the UK Biobank cohort
- To build a GWAS bioinformatics pipeline. This is a multi-step process for which many tools already exist. A key element is to create an installable, re-usable, extensible pipeline with excellent code documentation (see below for further details on the specific tools and concepts that will be used)

Proposed methods and access to data

This study will make part of a broader research project (AJ Berlanga-Taylor) investigating chronic, low-grade inflammation in richly phenotyped cohorts. We will mainly use data from the UK Biobank study (www.ukbiobank.ac.uk) (Sudlow et al., 2015). There is also the possibility of using other cohorts such as the Airwave Health Monitoring Study (www.police-health.org.uk) (Elliott et al., 2014).

Development of tools for computational biology

We will follow principles in reproducibility and best practice for computational biology (see for example (Noble, 2009; Sandve et al., 2013)). Specifically, the student will work with the following:
Reproducibility concepts and best practice implementation (Wilson et al., 2014; Wilson et al., 2017)
Use of Ruffus (Goodstadt, 2010) as a pipeline tool and CGAT tools (Sims et al., 2014) for support
Python programming and packagingrestructuredText and Sphinx for reporting
Travis and tox for testing, Conda and Docker for management and development, GitHub for version control
Skills required
In order to successfully complete the project, we expect that the student already has:
Excellent skills in basic bioinformatics, statistics and general scientific method
Experience with at least one statistical and/or programming language (R and/or Python preferably)
Strong interest in genetics and genetics methodology, chronic disease and inflammation
General problem-solving and self-directed learning within a supportive environment
Outcomes
Students will gain knowledge and experience in:
Genomic and biomedical analysis and knowledge of human genomics and disease
Epidemiology of chronic inflammation and chronic disease
Statistical analysis of large cohort data and GWAS methods
Deep understanding and experience in â€œbest-practiceâ€ approaches in research software engineering for reproducibility
This project is suitable for students with backgrounds in either biomedical or numerical areas with experience in computer programming and general bioinformatics. We expect students to have a desire to learn topic specific issues (biomedical) and develop research software that is usable by others. Deriving biological insights as well as re-usable computing tools are key outcomes of this project.
There are further questions that will follow from this study, which may be suitable for someone interested in developing deeper knowledge and skills in this area, such as using statistical and bioinformatics methods to fine-map associated variants (see for example (van de Bunt et al., 2015)).
Supervisor:
Dr. Antonio J Berlanga-Taylor MBBS, MSc, DPhil
Co-supervisors:
Dr. Ioanna Tzoulaki
Dr. Deborah Schneider-Luftman
Web references
CGAT tools: https://github.com/cgat-developers
Python: https://www.python.org/
Python Packaging User Guide: https://packaging.python.org/
Conda: https://conda.io/docs/
Docker: https://www.docker.com/
project_quickstart: https://github.com/AntonioJBT/project_quickstart
reStructuredText: http://docutils.sourceforge.net/rst.html
Sphinx: http://www.sphinx-doc.org/en/stable/
Travis CI: https://travis-ci.org/
tox automation project: https://tox.readthedocs.io/en/latest/
Key references
Academy of Medical Sciences (2018). Multimorbidity: a priority for global health research.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P. and Zondervan, K. T. (2010). "Data quality control in genetic case-control association studies." Nat Protoc 5(9): 1564-1573.

Barnett, K., Mercer, S. W., Norbury, M., Watt, G., Wyke, S. and Guthrie, B. (2012). "Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study." Lancet 380(9836): 37-43.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017). "An Expanded View of Complex Traits: From Polygenic to Omnigenic." Cell 169(7): 1177-1186.

Elliott, P., Vergnaud, A. C., Singh, D., Neasham, D., Spear, J. and Heard, A. (2014). "The Airwave Health Monitoring Study of police officers and staff in Great Britain: rationale, design and methods." Environ Res 134: 280-285.

Goodstadt, L. (2010). "Ruffus: a lightweight Python library for computational pipelines." Bioinformatics 26(21): 2778-2779.

Lander, E. S. and Schork, N. J. (1994). "Genetic dissection of complex traits." Science 265(5181): 2037-2048.

Lefevre, T., d'Ivernois, J. F., De Andrade, V., Crozet, C., Lombrail, P. and Gagnayre, R. (2014). "What do we mean by multimorbidity? An analysis of the literature on multimorbidity measures, associated factors, and impact on health services organization." Rev Epidemiol Sante Publique 62(5): 305-314.

Noble, W. S. (2009). "A quick guide to organizing computational biology projects." PLoS Comput Biol 5(7): e1000424.

Nunes, B. P., Flores, T. R., Mielke, G. I., Thume, E. and Facchini, L. A. (2016). "Multimorbidity and mortality in older adults: A systematic review and meta-analysis." Arch Gerontol Geriatr 67: 130-138.

Sandve, G. K., Nekrutenko, A., Taylor, J. and Hovig, E. (2013). "Ten simple rules for reproducible computational research." PLoS Comput Biol 9(10): e1003285.

Sims, D., Ilott, N. E., Sansom, S. N., Sudbery, I. M., Johnson, J. S., Fawcett, K. A., Berlanga-Taylor, A. J., et al. (2014). "CGAT: computational genomics analysis toolkit." Bioinformatics.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., et al. (2015). "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS Med 12(3): e1001779.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C. and Roland, M. (2009). "Defining comorbidity: implications for understanding health and health services." Ann Fam Med 7(4): 357-363.

van de Bunt, M., Cortes, A., Brown, M. A., Morris, A. P. and McCarthy, M. I. (2015). "Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci." PLoS Genet 11(9): e1005535.

van Hecke, O., Hocking, L. J., Torrance, N., Campbell, A., Padmanabhan, S., Porteous, D. J., McIntosh, A. M., et al. (2017). "Chronic pain, depression and cardiovascular disease linked through a shared genetic predisposition: Analysis of a family-based cohort and twin study." PLoS One 12(2): e0170653.

Villacampa-Fernandez, P., Navarro-Pardo, E., Tarin, J. J. and Cano, A. (2017). "Frailty and multimorbidity: Two related yet different concepts." Maturitas 95: 31-35.

Wang, M. H., Cordell, H. J., Van Steen, K., Lappenschaar, M., Hommersom, A. and Lucas, P. J. (2018). "Statistical methods for genome-wide association studies
Probabilistic causal models of multimorbidity concepts." Semin Cancer Biol 2012: 475-484.

WHO (2014). GLOBAL STATUS REPORT on noncommunicable diseases 2014, World Health Organization.

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H., et al. (2014). "Best practices for scientific computing." PLoS Biol 12(1): e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T. K. (2017). "Good enough practices in scientific computing." PLoS Comput Biol 13(6): e1005510.

Wolff, J. L., Starfield, B. and Anderson, G. (2002). "Prevalence, expenditures, and complications of multiple chronic conditions in the elderly." Arch Intern Med 162(20): 2269-2276.

WTCCC (2007). "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447(7145): 661-678.

Project 3/3	GWAS and QTL functional genomics annotation *Co-spvr Dr Paul David Blakeley*

GWAS and QTL functional genomics annotation

Background

Complex traits do not follow Mendelian inheritance patterns, are very likely caused or influenced by a multitude of genetic variants and express themselves in a variety of ways (phenotypes) (Boyle et al., 2017). Genome-wide association (GWA, GWAS) and quantitative trait loci (QTL) studies have become popular methods to analyse complex traits from a genomic perspective. Both approaches are statistical methods which determine the association between genotypes and phenotypes in order to understand the genomic basis of complex traits (Kearsey, 1998; Lander et al., 1994).

Downstream annotation and biomedical interpretation of summary data obtained from GWAS and QTL analyses remains a key limitation in complex trait genomics however (Albert et al., 2015). Large scale projects such as ENCODE (Dunham et al., 2012), BLUEPRINT/IHEC (Kundaje et al., 2015; Stunnenberg et al., 2016) FANTOM5 (Andersson et al., 2014), Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010), International Cancer Genome Consortium (ICGC) (Hudson et al., 2010) and GTEx (Battle et al., 2017) have greatly enhanced the catalogue of annotations related to the effects of DNA sequence variation.

The number of statistical methods and tools available to rapidly annotate a given set of results from GWAS and QTL studies has increased dramatically (see for example (Fang et al., 2016; Heger et al., 2013; Pers et al., 2015; Trynka et al., 2015)). Different tools address overlapping but distinct questions and efficient analysis remains a problem however. Hence, a mechanistic understanding of how genetic loci influence complex traits continues to be a challenge in any given project. As a motivating example we are currently analysing QTL loci from genome-wide genotypes and multiple metabolomic datasets from several thousand individuals in the Airwave cohort (Elliott et al., 2014).

Here we propose to deeply annotate these results in order to enhance our understanding and aid the generation of mechanistic hypotheses for variant function in relation to physiology and disease.

â€¢ Study questions

o What is the biomedical context of genomic variants associated with metabolites? Are they enriched in enhancers, regulatory loci and disease associated loci? What variants should be prioritised for further analysis?

â€¢ Study design

o Bioinformatics analysis and annotation of GWAS and QTL summary data from publicly available and unpublished results.

Aims

The overall aim of this study is to annotate genomic variants associated with complex traits arising from GWAS and QTL studies. Specific objectives include:

o To determine the biological and medical context of complex trait associated variants in the Airwave cohort. This will include providing context for genomic location, architecture, predicted effect and enrichment analysis corrected for important confounders and background noise, amongst others.

o To build a bioinformatics pipeline for downstream annotation of genetic variants and other â€˜omics high-throughput data. This is a multi-step process for which many tools already

exist. A key element is to create an installable, re-usable, extensible pipeline with excellent code documentation (see below for further details on the specific tools and concepts that will be used).

Proposed methods and access to data

We will use consortia generated and publicly available data from GTEx (Battle et al., 2017) and GWAS (www.ebi.ac.uk/gwas) (MacArthur et al., 2017) as well as unpublished results from ongoing analyses of genotype-metabolomic associations (Berlanga-Taylor AJ, Elliott PE, et al.).

Development of tools for computational biology

We will follow principles in reproducibility and best practice for computational biology (see for example (Noble, 2009; Sandve et al., 2013)). Specifically, the student will work with the following:

â€¢ Reproducibility concepts and best practice implementation (Wilson et al., 2014; Wilson et al., 2017)
â€¢ Use of Ruffus (Goodstadt, 2010) as a pipeline tool and CGAT tools (Sims et al., 2014) for support
â€¢ Python programming and packaging
â€¢ restructuredText and Sphinx for reporting
â€¢ Travis and tox for testing
â€¢ Conda and Docker for management and development
â€¢ GitHub for version control

Skills required

In order to successfully complete the project, we expect that the student already has:

â€¢ Excellent skills in basic bioinformatics, statistics and general scientific method
â€¢ Experience with at least one statistical and/or programming language (R and/or Python preferably)
â€¢ Strong interest in functional genomics, epidemiology and human disease
â€¢ General problem-solving and self-directed learning within a supportive environment

Outcomes

Students will gain knowledge and experience in:

â€¢ Genomic and biomedical analysis and experience of human genomics and disease
â€¢ Understanding of statistical methods in high-throughput â€˜omics data
â€¢ Deep understanding and experience in â€œbest-practiceâ€ approaches in research software engineering for reproducibility

This project is suitable for students with backgrounds in either biomedical or numerical areas with experience in computer programming and general bioinformatics. We expect students to have a desire to learn topic specific issues (biomedical) and develop research software that is usable by others. Deriving biological insights as well as re-usable computing tools are key outcomes of this project.

There are further questions that will follow from this study, which may be suitable for someone interested in developing deeper knowledge and skills in this area.

Supervisor:
Dr. Antonio J Berlanga-Taylor MBBS, MSc, DPhil
Co-supervisor:
Dr Paul David Blakeley
p.blakeley@imperial .ac.uk

Web references

CGAT tools: https://github.com/cgat-developers
Python: https://www.python.org/
Python Packaging User Guide: https://packaging.python.org/
Conda: https://conda.io/docs/
Docker: https://www.docker.com/
project_quickstart: https://github.com/AntonioJBT/project_quickstart
reStructuredText: http://docutils.sourceforge.net/rst.html

Sphinx: http://www.sphinx-doc.org/en/stable/
Travis CI: https://travis-ci.org/
tox automation project: https://tox.readthedocs.io/en/latest/
Key references

Albert, F. W. and Kruglyak, L. (2015). "The role of regulatory variation in complex traits and disease." Nat Rev Genet 16(4): 197-212.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., et al. (2014). "An atlas of active enhancers across human cell types and tissues." Nature 507(7493): 455-461.

Battle, A., Brown, C. D., Engelhardt, B. E. and Montgomery, S. B. (2017). "Genetic effects on gene expression across human tissues." Nature 550(7675): 204-213.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., et al. (2010). "The NIH Roadmap Epigenomics Mapping Consortium." Nat Biotechnol 28(10): 1045-1048.

Boyle, E. A., Li, Y. I. and Pritchard, J. K. (2017). "An Expanded View of Complex Traits: From Polygenic to Omnigenic." Cell 169(7): 1177-1186.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., et al. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature 489(7414): 57-74.

Elliott, P., Vergnaud, A. C., Singh, D., Neasham, D., Spear, J. and Heard, A. (2014). "The Airwave Health Monitoring Study of police officers and staff in Great Britain: rationale, design and methods." Environ Res 134: 280-285.

Fang, H., Knezevic, B., Burnham, K. L. and Knight, J. C. (2016). "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits." Genome Med 8(1): 129.

Goodstadt, L. (2010). "Ruffus: a lightweight Python library for computational pipelines." Bioinformatics 26(21): 2778-2779.

Heger, A., Webber, C., Goodson, M., Ponting, C. P. and Lunter, G. (2013). "GAT: a simulation framework for testing the association of genomic intervals." Bioinformatics 29(16): 2046-2048.

Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., et al. (2010). "International network of cancer genome projects." Nature 464(7291): 993-998.

Kearsey, M. J. (1998). "The principles of QTL analysis (a minimal mathematics approach)." Journal of Experimental Botany 49(327): 1619â€“1623.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., et al. (2015). "Integrative analysis of 111 reference human epigenomes." Nature 518(7539): 317-330.

Lander, E. S. and Schork, N. J. (1994). "Genetic dissection of complex traits." Science 265(5181): 2037-2048.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., et al. (2017). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." Nucleic Acids Res 45(D1): D896-d901.

Noble, W. S. (2009). "A quick guide to organizing computational biology projects." PLoS Comput Biol 5(7): e1000424.

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H. J., Wood, A. R., Yang, J., Lui, J. C., et al. (2015). "Biological interpretation of genome-wide association studies using predicted gene functions." Nat Commun 6: 5890.

Sandve, G. K., Nekrutenko, A., Taylor, J. and Hovig, E. (2013). "Ten simple rules for reproducible computational research." PLoS Comput Biol 9(10): e1003285.

Sims, D., Ilott, N. E., Sansom, S. N., Sudbery, I. M., Johnson, J. S., Fawcett, K. A., Berlanga-Taylor, A. J., et al. (2014). "CGAT: computational genomics analysis toolkit." Bioinformatics.

Stunnenberg, H. G. and Hirst, M. (2016). "The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery." Cell 167(5): 1145-1149.

Trynka, G., Westra, H. J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., Klein, R. J., et al. (2015). "Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci." Am J Hum Genet 97(1): 139-152.

Wilson, G., Aruliah, D. A., Brown, C. T., Chue Hong, N. P., Davis, M., Guy, R. T., Haddock, S. H., et al. (2014). "Best practices for scientific computing." PLoS Biol 12(1): e1001745.

Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L. and Teal, T. K. (2017). "Good enough practices in scientific computing." PLoS Comput Biol 13(6): e1005510.

Contact Details

Email: a.berlanga@imperial.ac.uk

Tel:

Principal Supervisor:	**Professor Sophia Yaliraki**
Further Information	
Department:	Other, Chemistry
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,
Dates of absence of more than two weeks? Dates, 20 July-August 5th	**Provision for supervision during primary supervisor's absence (if applicable):** Other email/skype as well as disussions with collaborator Prof. Barahona
Project 1/	*Co-spvr*
Project 2/	*Co-spvr*
Project 3/	*Co-spvr*
Project 4/	*Co-spvr*
Contact Details	
Email: s.yaliraki@imperial.ac.uk	**Tel:** 0207 5945899

Principal Supervisor:	**Peter DiMaggio**
Further Information	
Department:	Other, Chemical Engineering
Location: South Kensington,	**Facilities:** Desk only. State location in other,Other, Bone Building
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/1	Empirical Bayesian Analysis of Proteomic Modifications *Co-spvr Dorian Haskard; NHLI*
ADP-ribosylation, a protein post-translational modification, is integral to a diverse range of cellular processes. However, proteome-wide investigation of its cellular functions has not been possible due to numerous technical challenges. The DiMaggio lab has pioneered the development of a quantitative mass spectrometry proteomics workflow that has enabled the first comprehensive profiling of the "ADP-ribosylome" and elucidated the intracellular response of thousands of proteins to clinical PARP (the enzyme that "writes" ADP-ribosylation) inhibitors. Bioinformatic analysis of the large-scale proteomics data is performed using LIMMA (an empirical Bayes' approach) analysis of linear models designed to capture the expected protein responses. While this approach is certainly more robust than traditional methods (e.g. two-sample t tests are often used), we discovered an inherent bias in the data which arises from the sampling method utilised by the mass spectrometer. This project will involve extending these linear models, which are implemented in Jupyter notebooks, to investigate the methods for compensating for this replicate bias (e.g. performing quantitation at the peptide rather than protein level, correcting for the number of identified peptides across replicates, etc) and extending our current pathway analysis (e.g. Reactome) to relate these findings to differences in PARP enzyme structure.	
Contact Details	
Email: p.dimaggio@imperial.ac.uk	**Tel:** 020 7594 5589

Principal Supervisor:	**Matteo Fumagalli**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, Sir Ernst Chain Building
Dates of absence of more than two weeks? Dates, 23/07-13/08	**Provision for supervision during primary supervisor's absence (if applicable):** Primary supervisor via email/Skype
Project 1/1	A deep learning approach to estimate population parameters from large-scale genomic data *Co-spvr*
High-throughput sequencing data generate a large amount of DNA/RNA data. In parallel, novel computational methods have been developed to process such large amount of information to extract meaningful insights. One of the main challenges in population genetics is the quantification of natural selection, which is the process whereby a heritable trait becomes more common in a population if it confers an advantage to the carrier. Mutations targeted by natural selection are candidate for harboring important functionalities, so their detection is vital at the biomedical level. In our group we are introducing the use of convolutional neural network on images created from genomic data to identify signals of natural selection. The student will be in charge of benchmarking and optimising (and improving if necessary) this methodology by analysing a large collection of synthetic and natural data. Student should be familiar with python and unix and basic concepts of classification and statistics. Additional training will be provided. The main supervisor is based at Silwood Park. There is the possibility for the student to work at Silwood Park whenever she/he wants. Alternatively, the student will be mostly remotely (but constantly) supervised. The main supervisor aims at being at South Kensington, where the student will have a desktop, once a week.	
Contact Details	
Email: m.fumagalli@imperial.ac.uk	**Tel:** +44 (0)20 7594 3793

Principal Supervisor:	**Virginia Fairclough**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, Room 317
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):** Co Supervisor, listed with project
Project 1/1	What is the molecular basis for obligate biotrophy in plant pathogenic fungi? *Co-spvr Derek Huntley*
Plant and microbes commonly interact in nature; these interactions may lead to mutualistic symbioses or disease. In some cases, the interactions become so close that an obligate dependency evolves. These are then called obligate biotrophic fungi. Although these interactions are common, have been known for a long time and lead to some of the most prominent and devastating diseases of important staple crops (such as wheat and barley), it is not known what the molecular basis for this obligate relation is. Why can't we grow these fungi on a Petri dish? As a result of the last decadesâ€™ efforts in sequencing genomes of these fungi, it is clear that the main primary metabolic pathways are active in obligate biotrophs. In this project you will test the hypothesis that what is different here, is that the regulatory elements controlling the expression of genes encoding enzymes on primary metabolic pathways have mutated to become dependent on growth in a plant environment. Preliminary evidence pointed out that this may be the case. Now, with the availability of much broader data-sets, this hypothesis needs to be tested, using a deeper and more rigorously statistical analysis of the genome data available in the databases. Data for this project will be supplied by Professor Pietro Spanu.	
Contact Details	
Email: virginia.fairclough08@imperial.ac.uk	**Tel:**

Principal Supervisor:	**Myrsini Kaforou**
Further Information	
Department:	Medicine,
Location: St Mary's,	**Facilities:** Desk with Desktop . State location in other,
Dates of absence of more than two weeks? ,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/2	Unravelling signalling networks in malaria host-pathogen interaction for drug-development *Co-spvr Dr Moritz Treeck - the Francis Crick Insitute*

Malaria remains one of the most devastating tropical diseases in the tropics killing ~500.000 people every year, mainly children under the age of 5. An additional 200.000.000 people get infected, with devastating consequences for the economies and the healthcare systems of endemic countries. Resistance to all current drugs has been reported and there is no efficient vaccine. Therefore, identifying novel drug targets and a better understanding how the parasite evades the human immune system is paramount in the fight against the disease.

One of the key elements for successful host-immune evasion is the parasiteâ€™s ability to remodel the host cell it lives in1, the human red blood cell (RBC). It remodels the cell extensively and does so by exporting ~10% of its proteome into the RBC2. Among these ~450 proteins is an expanded kinase family3,4 that is unusual in several ways. We predict that these kinases are fundamental in host cell remodelling and immune evasion. We also predict that they are formidable drug targets because they are atypical kinases and as such different from human counterparts, which is important for drug specificity.

To investigate the functional role of these kinases in RBC remodelling we have first collected several large-scale phosphoproteome datasets that a) show that there is indeed phosphorylation of the RBC upon infection with malaria parasites, b) that these phosphorylation events only occur in virulent malaria species5. We then generated gene knock-outs for each of the ~20 kinases and measured the phosphoproteome to identify which kinase is mediating which phosphorylation events in the infected cell. Finally, in collaboration with GlaxoSmithKline, we identified highly specific drugs that specifically interfere with phosphorylation in the host cell upon infection.

The aim of this project is to analyse these datasets and draw a comprehensive map of kinase-mediated phosphorylation in the Malaria infected cell. Identifying which protein phosphorylation sites are regulated by kinases and in which manner will not only help to unravel the mechanisms in the host-cell that are important for immune evasion, it will also allow to relate proteins of unknown function and thus uncover novel biology. This knowledge is paramount for a better understanding of host-cell remodelling and virulence and will have direct impact on our further efforts in drug development.

The data analysis pipeline will be implemented in R. Statistical modelling, AI and machine learning tools will be used to build the kinase mediated phosphorylation map in Malaria infected cells.

1 Maier, A. G., Cooke, B. M., Cowman, A. F. & Tilley, L. Malaria parasite proteins that remodel the host erythrocyte. Nat Rev Microbiol 7, 341-354, doi:10.1038/nrmicro2110 (2009).

2 Marti, M., Good, R. T., Rug, M., Knuepfer, E. & Cowman, A. F. Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306, 1930-1933, doi:10.1126/science.1102452 (2004).

3 Nunes, M. C., Goldring, J. P., Doerig, C. & Scherf, A. A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell. Molecular microbiology 63, 391-403, doi:10.1111/j.1365-2958.2006.05521.x (2007).

4 Nunes, M. C., Okada, M., Scheidig-Benatar, C., Cooke, B. M. & Scherf, A. Plasmodium falciparum FIKK kinase members target distinct components of the erythrocyte membrane. PloS one 5, e11747, doi:10.1371/journal.pone.0011747 (2010).
5 Otto, T. D. et al. Genomes of all known members of a Plasmodium subgenus reveal paths to virulent human malaria. Nat Microbiol, doi:10.1038/s41564-018-0162-2 (2018).

Project 2/2	Integrating different layers of human -omics data to improve clinical diagnostics for infectious diseases *Co-spvr Jethro Herberg*

Infectious diseases are amongst the most complex multifactorial diseases, as they depend on dynamic interactions between the host and the pathogen. The development of high-throughput, low-cost technologies for studying the human and pathogen genomes, along with the epigenome, transcriptome and proteome have resulted in an exponential growth in available data. However, much of the potential benefit remains untapped as the current standard analytical approach relies on single-level analyses of these datasets. As different levels of â€“omics data capture different levels of the immunological processes; information needs to be combined across DNA, RNA, protein and clinical data into predictive models of diagnosis and treatment decision.
Complex systems such as the host-pathogen interaction within an individual patient, can be understood more thoroughly if considered as a whole. This project aims to integrate and then explore the complex -omics and clinical data of unique cohorts of infectious disease patients using advanced bioinformatics techniques. This project also aims at generating novel disease specific signatures using the minimal number of multi-layer combinations of biomarkers that show optimal sensitivity and specificity. The integrative analysis will lead to a better understanding of the key biological processes involved in the host response to infection, to the discovery of robust biomarkers.
The student will have access to unique, well-phenotyped patient dataset of over ~1,000 patients with a range of severe infectious diseases, including N. meningititis, tuberculosis, and viral infections such as RSV and influenza.
The project can offer the student:
-Understanding host response in infectious diseases
-Experience in R
-Exploring and developing algorithms for data integration

1. Kaforou, M., V.J. Wright, T. Oni, N. French, S.T. Anderson, N. Bangani, C.M. Banwell, A.J. Brent, A.C. Crampin, H.M. Dockrell, B. Eley, R.S. Heyderman, M.L. Hibberd, F. Kern, P.R. Langford, L. Ling, M. Mendelson, T.H. Ottenhoff, F. Zgambo, R.J. Wilkinson, L.J. Coin, and M. Levin, Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. PLoS Med, 2013. 10(10): p. e1001538.
2. Herberg, J.A., M. Kaforou, V.J. Wright, H. Shailes, H. Eleftherohorinou, C.J. Hoggart, M. Cebey-Lopez, M.J. Carter, V.A. Janes, S. Gormley, C. Shimizu, A.H. Tremoulet, A.M. Barendregt, A. Salas, J. Kanegaye, A.J. Pollard, S.N. Faust, S. Patel, T. Kuijpers, F. Martinon-Torres, J.C. Burns, L.J. Coin, M. Levin, and I. Consortium, Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children. JAMA, 2016. 316(8): p. 835-45.
3. Chaussabel, D. and N. Baldwin, Democratizing systems immunology with modular transcriptional repertoire analyses. Nat Rev Immunol, 2014. 14(4): p. 271-80.
4. Wang, B., A.M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, Similarity network fusion for aggregating data types on a genomic scale. Nat Methods, 2014. 11(3): p. 333-7.

Contact Details

Email: m.kaforou@imperial.ac.uk	**Tel:** 4.47531828e+011

Principal Supervisor:	**Giovanni Sena**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk only. State location in other,Other, SAF 333 and Library
Dates of absence of more than two weeks? Dates, 23/7-13/8	**Provision for supervision during primary supervisor's absence (if applicable):** Co Supervisor, listed with project
Project 1/1	A new platform for automated quantification of tropism responses in plant roots. *Co-spvr Nick Oliver (Lab Technician)*
In this project, you will develop a stand-alone software for image processing and quantitative measurements of plant roots. In our lab, we are developing new automated assays for live imaging of Arabidopsis roots exposed to external electric fields, using a combination of tools ranging from 3D printing, Raspberry Pi programming and standard plant molecular genetics. One of our projects is currently collecting high-throughput time-lapse images on the electrotropic effect, or the progressive alignment of the root tip with the external electric field. The challenge for you is to develop a robust and intuitive pipeline to automate the image processing protocol, going from raw images to a measurement of root tip alignment with respect to the field direction. Preferred languages are MATLAB or Python. A user-friendly GUI would be appreciated, but not essential at this stage. Time permitting, you will also be involved in the actual analysis of the collected data, through the script you just developed. This will add statistical analysis to your project, and a strong link to quantification of population heterogeneity in electrotropic response.	
Contact Details	
Email: g.sena@imperial.ac.uk	**Tel:** x47448

Principal Supervisor:	**John Pinney**
Further Information	
Department:	Life Sciences,
Location: South Kensington,	**Facilities:** Desk with Desktop . State location in other,Other, Theosysbio lab, DoLS
Dates of absence of more than two weeks? No,	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/3	Exploring the landscape of glycan structure. *Co-spvr*
Glycans are tree-like polysaccharides that are frequently found attached to specific sites on eukaryotic cell surface and cellular matrix proteins, with diverse functions including protein folding, cell-cell signaling and immune response. Large databases describing observed glycan structures are now available, which provide an excellent resource for describing and exploring the space of naturally occurring glycans. In this project you will apply tree kernel methods to calculate distances between these structures and thereby develop clustering and data reduction methods to capture the most important differences between subsets of glycans, such as those observed in cancerous vs non-cancerous tissue.	
Project 2/3	Identification of cancer-specific protein isoforms *Co-spvr Oliver Pearce (QMUL)*
Oliver Pearce's lab at QMUL has gathered transcriptomic and proteomic data sets on 30 high grade serous ovarian cancer patient tissues and integrated these two data sets using a technique known as "proteomics informed by transcriptomics" (PIT). PIT identifies changes in splicing by pairing the transcript and peptide fragments together, significantly increasing the power of protein isoform detection. Preliminary analysis has revealed notable variation in protein sequences within the extracellular matrix (ECM), including candidate novel proteins. The aim of this project is to take this existing data integration, and extract from it a list of disease specific isoforms, and potentially candidate novel isoforms. Such targets are potentially useful for cancer diagnostics or potentially therapeutics. Having identified interesting isoforms, further sequence and/or structural analysis will be applied to explore the functional implications of the observed variations.	
Project 3/3	Predicting microsporidia-host interactions from genome sequence *Co-spvr*
Microsporidian parasites are an interesting model for host-pathogen interactions, as they are transmitted vertically from parent to child. There is therefore a very strong phylogenetic link between the host and parasite genomes, and we expect that coevolving protein pairs should be more easily identifiable in this type of system. In this project, you will apply a variety of methods to genome sequences to examine the evidence for host-parasite protein-protein interactions, including working with multiple alignments, phylogenetic trees and diversifying selection.	
Contact Details	
Email: j.pinney@imperial.ac.uk	**Tel:** 2075948629

Principal Supervisor:	**Becca Asquith**
Further Information	
Department:	Medicine
Location: St Marys	**Facilities:** Desk but no desktop. Communal area adjacent to supervisors office.
Dates of absence of more than two weeks? no	**Provision for supervision during primary supervisor's absence (if applicable):**
Project 1/2	Investigating the Maintenance of Immune Memory
We have recently identified a population of immune stem cells that we believe is responsible for the maintenance of immune memory [Costa del Amo, PLOS Biology 2018 in press]. The aim of this project is to study the transcriptional profile of these immune cells.	
Project 2/2	Model selection
We are investing classical (frequentist) tools for model selection. We are interested in how incorporating nominally "uninformative" data impacts on model selection.	
Contact Details	
Email: b.asquith@imperial.ac.uk	**Tel:** 0207 594 3731

Principal Supervisor:	**Tim Ebbels**
Further Information	
Department:	Medicine
Location: South Ken	**Facilities:** Desk only Hot desk in SAF
Dates of absence of more than two weeks? **30 July-29 August**	**Provision for supervision during primary supervisor's absence (if applicable): Co supervisor Rui CLIMACO PINTO**
Project 1/1	Matching features across untargeted MS metabolomics datasets
Metabolomics is the study of small molecules (metabolites) in biological tissues. The levels of metabolites constitute a â€˜health fingerprintâ€™ of a biological system and are perturbed in disease and physiological stress. Untargeted metabolomics using liquid chromatography - mass spectrometry (LC-MS) is widely used to capture variation in the largest possible number of these molecules. Although this approach can lead to the discovery of new biomarkers of disease, the detected features are not easily annotated to a specific metabolite but are identified simply by the mass-to-charge ratio of their ions and their chromatographic retention time. This makes it difficult to compare results between experiments, as there is no simple correspondence between metabolomic features in both. The development of a robust method to find correspondence between features in two datasets is of enormous importance in the field, as it would allow concatenation of studies to increase discovery power, or to use one study for discovery and another for validation. In this project, we will explore, validate and extend an existing strategy addressing the correspondence problem. New aspects include the use of adducts (multiple â€œflavoursâ€ of the same metabolite) as anchors for retention time shift or to calculate a matching score; development of a probability based matching method; the automation of the method; and the creation of a package in R/Python for wider use within the scientific community, with the possibility of publication in a relevant journal.	
Contact Details	
Email: t.ebbels@imperial.ac.uk	**Tel:** 02075943160