Program and Abstract Book

swedish bioinformatics workshop 2019

October 21-22, 2019
Conference Centre Wallenberg
Göteborg

Swedish Bioinformatics Workshop
21-22 October 2019

Conference Center Wallenberg,
Gothenburg, Sweden

Welcome to the 2019 Swedish Bioinformatics Workshop!

The organizing committee would like to welcome you to Göteborg for this yearly occasion that gathers researchers in bioinformatics, computational biology, systems biology, genomics, and other such fields from Sweden and abroad.

This year we anticipate over 100 participants to the workshop and we are thrilled to share this experience with you. In this two day event, we have an exciting program including presentations, posters, workshops, and a career session. We have a very diverse collection of research areas covered, including AI and machine learning, single-cell technologies, and genomics and gene regulation.

We hope that you will also join us for the SBW dinner on October 21st, which will be served at the Lyktan restaurant in the Wallenberg conference center.

Finally, if you have any inquiries regarding the workshop, any member of the organizing committee will be easily recognizable throughout SBW with a green name badge, and will be happy to assist you.

Thanks for joining SBW 2019!

The organizing committee

Schedule

Monday 21 Oct.		
9:00 - 10:00	**Registration and welcome + coffee** (Set up posters)	
10:00 - 10:05	**Opening address**	
10:05 - 11:00	**Christina Leslie** Decoding epigenomic programs in cancer and immunity	
11:00 - 11:30	**Nikolay Oskolkov** National Bioinformatics Infrastructure Sweden (NBIS)	
11:30 - 12:00	**Christoph Börlin** Modeling TF binding profiles through competitive binding	
12:00 - 12:30	**Fredrik Karlsson** Bioinformatics and machine learning for generation and characterization of advanced cell models for drug discovery	
12:30 - 13:30	**Lunch**	
13:30 - 14:30	**Lior Pachter** Some solved and unsolved bioinformatics problems in single-cell genomics	
14:30 - 15:00	**Aleksej Zelezniak** Uncovering DNA grammar using deep learning	
15:00 - 16:00	**Coffee break + Poster session 1**	
16:00 - 18:00	**Anna Reymer** *Workshop* Understanding the biological function of a protein using structural bioinformatics	**Johan Bengtsson-Palme** *Workshop* Wait a minute! – On annotation errors and how to spot them
18:00	**Dinner (at Wallenberg Conference Center)**	

Tuesday 22 Oct.	
9:00 - 10:00	**Mihaela Zavolan** Learning miRNA-dependent regulatory networks from high-throughput data
10:00 - 10:30	**Sara Younes** Identification of combinatorial markers in Systemic Lupus Erythematosus and disease subtypes
10:30 - 11:00	**Coffee break + Poster session 2**
11:00 - 11:30	**Olukayode Daramola** Differential Expression of Annotated Adenylate Cyclase Genes in various Life Stages of *Fasciola hepatica*
11:30 - 12:00	**Marcela Davila** University of Gothenburg Bioinformatics Core Facility
12:00 - 12:30	**Johanna Hörberg** Regulatory role of torsional stress on bZIP transcription factors Interactions with DNA
12:30 - 13:30	**Lunch**
13:30 - 14:00	**Francesco Gatto** Elypta
14:00 - 14:30	**Erik Larsson Lekholm** Understanding and modeling regional mutational heterogeneity in skin cancers
14:30 - 15:00	**Coffee break**
15:00 - 16:00	**Careers in Computational Biology** (hosted by RSG-Sweden) [http://www.rsg-sweden.iscbsc.org/]
16:00 - 16:05	**Closing remarks**

Legend

- Keynote speaker
- Sponsor presentation
- Invited speaker
- Oral presentation
- Workshop/Career session
- Poster session

Organizing Committee

The organization of SBW2019 is a joint collaboration between PhD students, postdocs, and other scientists from Chalmers & GU, as well as from the Regional Student Group in Sweden (RSG-Sweden).

Jonathan	**Robinson**	**jonrob@chalmers.se**
Johan	Bengtsson-Palme	johan@microbiology.se
Johan	Gustafsson	gustajo@chalmers.se
Hao	Wang	hao.wang@chalmers.se
Anna	Reymer	anna.reymer@gu.se
Ahmed	Waraky	ahmed.waraky@gu.se
Angelo	Limeta	angelol@chalmers.se
Jari	Martikainen	jari.martikainen@gu.se
Sanna	Abrahamsson	sanna.abrahamsson@gu.se
Marcela	Davila	marcela.davila@gu.se
Nazeefa	Fatima	nazeefa@iscb.org
Sri Harsha	Meghadri	harshameghadri@gu.se
Deborah	Oliveira	de0580ol-s@student.lu.se
Suze	Roostee	su2781ro-s@student.lu.se
Gustaw	Eriksson	gustaw.eriksson.649@student.lu.se

Contents

Keynote Speakers

Christina Leslie

Associate Member, Computational Biology Program, Sloan Kettering Institute, USA

Bio: Christina Leslie received a PhD in Mathematics from the University of California, Berkeley, and did her postdoctoral training in the Mathematics Department at Columbia University in 1999-2000. She then joined the faculty of the Computer Science Department and later the Center for Computational Learning Systems at Columbia University, where she became the principal investigator leading the Computational Biology Group. In 2007, she moved her lab to the Computational Biology program of Memorial Sloan Kettering Cancer Center, where she is currently an Associate Member.

Christina Leslie's research group uses computational methods to study the regulation of gene expression in mammalian cells and the dysregulation of expression programs in cancer. She is well known for developing machine learning approaches for analysis of high-throughput biological data, particularly from next-generation sequencing. Focus areas in the lab include dissecting transcriptional and epigenetic programs in differentiation, microRNA-mediated gene regulation, alternative cleavage and polyadenylation, and integrative analysis of tumor data sets.

Talk: **Decoding epigenomic programs in cancer and immunity**

Dysregulated epigenetic programs are a feature of many cancers, and the diverse differentiation states of immune cells as well as their dysfunctional states in tumors are in part epigenetically encoded. We will describe recent analysis work and computational methodologies from our lab to decode epigenetic programs from chromatin accessibility data (e.g. from ATAC-seq) and other genome-wide data in cancer and immunity.

We will present several vignettes to study how somatic alterations in epigenetic regulators lead to lineage plasticity in tumors. With the Sawyers lab, we show that mutations in the pioneer transcription factor FOXA1 in prostate cancer lead to altered differentiation programs and tumor phenotypes through analysis of ATAC-seq and ChIP-seq in mouse prostate organoid systems. With the Scaltriti and Baselga labs, we show that loss of function of ARID1A in ER+ breast cancer models leads to loss of luminal identity and resistance to endocrine therapy.

We will also describe a unified chromatin state and single cell expression analysis underlying T cell differentiation to "exhaustion". Numerous studies in chronic viral infection and tumor models have shown that most T cells progress to a terminally dysfunctional ("exhausted") state from which they cannot be rescued by current immunotherapies. We performed a systematic analysis of over 280 ATAC-seq and RNA-seq experiments from eight published studies of CD8 T cell dysfunction, using a statistical batch correction to define a common differentiation trajectory to terminal exhaustion in all settings of chronic antigen exposure. We further performed scRNA-seq analysis of CD8 T cells at multiple time points during acute and chronic viral infection to elucidate this trajectory at single cell resolution.

Finally, we will present a novel machine learning approach called BindSpace to leverage massive in vitro transcription factor (TF) binding data from SELEX-seq experiments through a joint embedding of DNA k-mers and TF labels, leading to improved prediction of TF binding.

Lior Pachter

Bren Professor of Computational Biology and Computing and Mathematical Sciences, California Institute of Technology, USA

Bio: Lior Pachter received a PhD in Applied Mathematics from MIT in 1999. He then moved to the University of California at Berkeley where he was a postdoctoral researcher (1999-2001), assistant professor (2001-2005), associate professor (2005-2009), and until 2018 the Raymond and Beverly Sackler professor of computational biology and professor of mathematics and molecular and cellular biology with a joint appointment in computer science. Since January 2017 he has been the Bren professor of computational biology at Caltech.

Lior Pachter's research interests span the mathematical and biological sciences, and he has authored over 100 research articles in the areas of algorithms, combinatorics, comparative genomics, algebraic statistics, molecular biology and evolution. His lab develops computational and experimental methods for genomics, and is currently focused on the development of single-cell genomics technologies and their application to RNA biology. The computational challenges addressed in his group involve the analysis of high-dimensional data.

Talk: **Some solved and unsolved bioinformatics problems in single-cell genomics**

Mihaela Zavolan

Professor in Computational Biology/Genomics, Biozentrum, University of Basel, Switzerland

Bio: Mihaela Zavolan received a PhD in Computer Science from the University of New Mexico in Albuquerque. Between 1993 and 2003 she conducted research at the Santa Fe Institute in Santa Fe, the Los Alamos National Laboratory in Los Alamos, as well as at the Rockefeller University in New York. In 2003, Mihaela Zavolan was appointed Professor of Computational and Systems Biology at the Biozentrum of the University of Basel. She is also a group leader in the Swiss Institute of Bioinformatics (SIB).

The primary research focus in Mihaela Zavolan's group is on microRNAs (miRNAs), which regulate the expression of protein coding genes to control cell differentiation, metabolism, and immune responses. Through the development of high-throughput experimental methods and computational analyses, she has contributed to the discovery of miRNAs in various organisms ranging from viruses to humans. Her group has developed algorithms to predict miRNA genes and miRNA targets, and has worked on the development of the CLIP method (cross-linking and immunoprecipitation) for mapping the binding sites of RNA-binding proteins in RNAs.

Talk: **Learning miRNA-dependent regulatory networks from high-throughput data**

Some two decades ago, the veil has been lifted on vast numbers of RNAs with regulatory function. Among these, miRNAs are small regulators of gene expression that guide Argonaute proteins to mRNA targets, inducing target degradation and translational repression. Comparative genomics studies showed that a miRNA typically has hundreds of targets, yet target prediction still remains challenging. MiRNA target characterization has been a focus of research in my group since the discovery of miRNAs and over the years we have contributed various models and methods that use large-scaled measurements to infer how miRNAs regulate their targets. With these methods we have uncovered key miRNAs and transcription factors that determine cell type-specific transcriptomes. Furthermore, taking advantage of the commonality of mechanisms between miRNAs and small interfering RNAs (siRNAs), which are frequently used to manipulate gene expression, we have integrated the prediction of siRNA on- and off-targets to improve the accuracy of identification of phenotype-conferring genes from genome-wide siRNA screens. In this presentation I would like to review this work to highlight the principles of miRNA-dependent regulation that have been derived primarily with computational approaches.

Invited speakers

We have invited two local (Gothenburg) researchers to give a talk about their research activities. Aleksej Zelezniak, an Assistant Professor at Chalmers University, and Erik Larsson Lekholm, a Professor at the University of Gothenburg will share some of their exciting work.

Aleksej Zelezniak

Assistant Professor, Division of Systems and Synthetic Biology, Department of Biology and Biological Engineering, Chalmers University, Gothenburg, Sweden

Talk: **Uncovering DNA grammar using deep learning**

Erik Larsson Lekholm

Professor, Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden

Talk: **Understanding and modeling regional mutational heterogeneity in skin cancers**

Speakers from Industry/Infrastructure

We have three speakers joining us from beyond the realm of academia, providing some insight into the interesting work in which they are currently involved. The speakers are Nikolay Oskolkov from the National Bioinformatic Infrastructure of Sweden (NBIS), Marcela Davila from the University of Gothenburg Bioinformatics Core Facility (BCF), and Francesco Gatto from local start-up, Elypta.

Nikolay Oskolkov

NBIS Expert, National Bioinformatics Infrastructure Sweden (NBIS)
nbis.se

Marcela Davila

Head of Bioinformatics Core Facility at Sahlgrenska Academy, Gothenburg University
cf.gu.se/bioinformatics

Francesco Gatto

Chief Scientific Officer, Elypta
elypta.com

Workshops

Workshop 1. Understanding the biological function of a protein using structural bioinformatics

Moderator: Anna Reymer (anna.reymer@gu.se)

NOTE: You need to bring your own laptop for this workshop, and have UCSF Chimera installed!

Abstract: Structural bioinformatics studies molecular structures and interactions to explain the function of a biological molecule in question, and when necessary find a way to block this function for medicinal purposes. In this workshop, we will combine homology modeling to create a model of a protein, and molecular docking to study interactions of the protein with a set of small ligands, potentially pharmacologically potent.

Workshop 2. Wait a minute! – On annotation errors and how to spot them

Moderator: Johan Bengtsson-Palme (johan@microbiology.se)

NOTE: You need to bring your own laptop for this workshop!

Abstract: Bioinformatic approaches for functional predictions rely upon correctly annotated database sequences. However, the assumption that the information in biological databases can be trusted is only sometimes true. The presence of inaccurately annotated or otherwise poorly described sequences introduces noise and bias to biological analyses and in the worst cases lead us do draw totally wrong conclusions. In this workshop, we will look at some things to be suspicious about in bioinformatic results and discuss strategies to deal with poorly annotated reference sequences.

Career Session

To give experienced researchers a chance to share their career experiences with young researchers, the Regional Student Group in Sweden (RSG-Sweden; www.rsg-sweden.iscbsc.org) has arranged a Panel discussion at 15:00 on Tuesday 22nd of October in the hall Europa. The session will offer possibilities for researchers to ask a panel of experienced scientists about the important skills needed to establish a career in industry and academia.

Oral Presentations

OP01. Differential Expression of Annotated Adenylate Cyclase Genes in various Life Stages of Fasciola hepatica
Presenting author: Olukayode Daramola (Olukayode.Daramola@liverpool.ac.uk)

OP02. Regulatory role of torsional stress on bZIP transcription factors Interactions with DNA
Presenting author: Johanna Hörberg (johanna.horberg@gu.se)

OP03. Identification of combinatorial markers in Systemic Lupus Erythematosus and disease subtypes from gene expression data using Rule Based Networks
Presenting author: Sara A. Younes (sara.younes@icm.uu.se)

OP04. Bioinformatics and machine learning for generation and characterization of advanced cell models for drug discovery
Presenting author: Fredrik Karlsson (fredrik.h.karlsson@astrazeneca.com)

OP05. Modeling TF binding profiles through competitive binding
Presenting author: Christoph S Börlin (borlinc@chalmers.se)

OP01. Differential Expression of Annotated Adenylate Cyclase Genes in various Life Stages of Fasciola hepatica

Olukayode Daramola[1], Jane Hodgkinson[2], Steve Paterson[1]

1. Centre for Genomic Research, Institute of Integrative Biology, University of Liverpool, United Kingdom
2. Institute of Infection and Global Health, University of Liverpool, United Kingdom

Presenting author: Olukayode Daramola (Olukayode.Daramola@liverpool.ac.uk)

Adenylate Cyclases catalyse the conversion of Adenosine Triphosphate (ATP) to 3',5'-cyclic AMP (cAMP) and pyrophosphate. These genes are regulated by G protein-coupled receptors and facilitate production of cAMP which serves as regulatory signals. These genes are part of the Ras, adenylyl cyclase, protein kinase A (PKA) nutrient-sensing pathway, regulates metabolism, cell division, entry into stationary phase, ion gates and the stress response. These gene families have been investigated extensively many organisms due to their roles in ATP-related biochemical activities. In Fasciola hepatica for example, previous studies indicate adenylate activity is reduced in liver flukes that are resistant to Triclabendazole (TCBZ), a drug of choice. TCBZ interestingly has been found to increase resistance to stress in yeast. The importance of AC genes makes them a candidate gene family to target as potential drug targets in ongoing research efforts to find new drug targets effective against Fasciola hepatica, this is needed to curb the rapidly spreading incidences of TCBZ resistance across Europe. To facilitate this, it is important to annotate these genes in the Fasciola genome. Using available Fasciola genomes; 2 Fasciola hepatica genomes (from UK and US), Fasciola gigantica genomes (from US and India), we explore this gene family. Using our published UK Fasciola hepatica genome and RNA Seq data from various stages of the Parasite, we assess the expression profile of these AC genes. Our results indicate that these AC genes are more expressed in newly encysted juvenile stages of the parasite with lower expression in adult flukes. We observed that only 2 of these AC genes are predominantly expressed. The orthologs of these genes in the other assemblies were identified and assessed for selective pressure. These findings provide insights into the biology of liver flukes, the annotation of these AC genes, as well as instigate more research questions.

Time and place: Tuesday, October 22, 11.00, Room: Europe

OP02. Regulatory role of torsional stress on bZIP transcription factors Interactions with DNA

Johanna Hörberg[1], Kevin Moreau[1], Anna Reymer[1]

1. Department of chemistry and molecular biology, University of Gothenburg, Sweden

Presenting author: Johanna Hörberg (johanna.horberg@gu.se)

Torsional stress on DNA, introduced by molecular motors when the molecule undergoes under- or overtwisting, constitutes an important regulatory mechanism of gene expression. Bringing local conformational alterations that facilitate opening of promoters and nucleosome remodelling, torsional stress might be the key factor in modulation of specific binding of transcription factors to DNA. Using all-atom microsecond scale molecular dynamics simulations together with a torsional restrain that controls the helical twist of DNA, we addressed the impact of torsional stress on MafB-DNA interactions. MafB (PDB ID: 4AUW) is a representative of the bZIP family of transcription factors, which recognises the palindromic DNA sequence (TGCTGACGTCAGCA). We over- and underwind free DNA and DNA in complex with MafB by ±5 per dinucleotide step, and monitor the evolution of the protein-DNA contacts at different degrees of torsional strain. Our computations show that MafB changes the DNA sequence-specific response to the torsional stress; the dinucleotide steps that are anticipated to absorb most of the stress become more torsionally rigid as these are involved in the specific protein-DNA contacts. Also, the protein undergoes substantial conformational changes to adjust to modified groove geometries to maintain, key contacts with DNA. This results in a significant asymmetric free energy profile with respect to free DNA, where overtwisting is energetically unfavourable. Our data suggest that MafB could act as a torsional stress insulator, to prevent the propagation of torsional stress along the chromatin fibre.

Time and place: Tuesday, October 22, 12.00, Room: Europe

OP03. Identification of combinatorial markers in Systemic Lupus Erythematosus and disease subtypes from gene expression data using Rule Based Networks

Sara A.Yones, Alva Annett, Patricia Stoll, Jennifer R. S. Meadows, Fredrick Barrenas, Jan Komorowski

1. Department of Cell and Molecular Biology, Uppsala University,
2. Institutionen för Biologisk Grundutbildning, Uppsala University

Presenting author: Sara A. Yones (sara.younes@icm.uu.se)

Systemic Lupus Erythematosus (SLE) is an autoimmune disease characterized by unpredictable periods of flares that is driven by stochastic execution of a complex inherited program. The flares are presented as different SLE disease activities (DA). Due to the unprecedented flares, disease complexity and heterogeneity among individuals it is very hard to tailor a personalized treatment for SLE patients. Several studies have been conducted to explore the genetic differences between healthy controls and SLE patients. However, efforts directed to study the combinatorial effects of genes towards the manifestation of different SLE disease activities in different patients, subgroups have been very limited. Patient subgroups potentially indicate different SLE subtypes.

In this study we mainly focus on understanding the manifestation of SLE for DA1 and DA3 from a multivariate perspective and in different patients, subgroups. In order to achieve the latter we developed a rule based model on gene expression data to model the complex combinatorial effects of genes. Rule based models are highly interpretable and transparent and can be visualized in the form of rule networks. The networks produced from the developed model not only showed a clear distinction between DA1 and DA3 but also the state of the genes and their interactions to discern between the two DAs. We fine-tuned the model by pruning certain objects from the training set according to certain criteria in order to discover the molecular heterogeneity in SLE disease subtypes.

Time and place: Tuesday, October 22, 10.00, Room: Europe

OP04. Bioinformatics and machine learning for generation and characterization of advanced cell models for drug discovery

Fredrik Karlsson[1]

1. Data Sciences and Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

Presenting author: Fredrik Karlsson (fredrik.h.karlsson@astrazeneca.com)

A key challenge for early drug discovery is the availability of translatable in vitro cellular assay systems that can be used to assess the potency and safety of a molecule in consistency with clinical outcome. The prospect of creating more translatable models for drug discovery has substantially expanded with the introductions of CRISPR and induced pluripotent stem cells (iPSCs). As part of assessing cellular model translatability, we use transcriptomics to obtain a detailed molecular characterization of in vitro cellular systems and compare to human tissue.

In this talk, I will present 3 examples of bioinformatics approaches that we use in AstraZeneca to support generation and assess translatability of advanced cellular models.

1) A machine learning model for predicting variants generated from CRISPR genomic cuts. This is being used to faster generate better cellular models.

2) Integration of diverse scRNASeq datasets is crucial for comparing data from model systems to human tissue. Therefore, we have developed an approach to use autoencoders for integration and batch correction of single cell RNASeq data. We have trained and applied an autoencoder to integrate scRNASeq data from 4 different studies. Integration with the autoencoder clusters cell types from individual studies together and outperforms established methods for scRNASeq integration.

3) scRNASeq characterization of kidney organoids derived from iPSCs. Sequencing of 50.000 cells from different timepoints in the differentiation process and comparison to human kidney.

Time and place: Monday, October 21, 12.00, Room: Europe

OP05. Modeling TF binding profiles through competitive binding

Christoph S Börlin, David Bergenholm, Verena Siewers and Jens Nielsen

1. Chalmers University of Technology, Department of Biology and Biological Engineering, Gothenburg, Sweden

Presenting author: Christoph S Börlin (borlinc@chalmers.se)

Regulation of gene expression is one of the major determinants of cellular phenotype and is mainly influenced by binding of transcription factors (TFs). To this date, the factors effecting the binding of TFs are not well understood and prediction of TF binding relays solely on experimentally or computational derived motifs. Known motifs for TFs however can only highlight were the TF can potentially bind, but it does not provide conclusive information whether the TF is actually going to bind there and if so with which strength. This lack of knowledge has hindered the knowledge driven design of promoters with a predictable TF binding pattern.

In our approach, we use high quality TF binding profile data obtained using chromatin immunoprecipitation followed by exonuclease treatment and sequencing (ChIP-exo) to discover the underlying motif binding strengths and model this as a competitive process between the TF of interest and all other TFs. This added competition between TFs can explain why the same motif is bound with divergent strengths in different genomic loci and is therefore essential to accurately predicting TF binding patterns over the full promoter sequence.
Using cross validation, we were also able to show that our model can learn the underlying motif strengths and not only memorizes it. This demonstrates that a rather simple mechanistic model using only motif strength for each kmer as parameter is able to accurately capture TF binding patterns if one accounts for the overall competition for the same stretch of DNA.

Time and place: Monday, October 21, 11.30, Room: Europe

Poster Presentations

P01. Dissecting Cell-to-Cell Variation in Single-Cell RNA-Seq Data
Presenting author: Johan Gustafsson (gustajo@chalmers.se)

P02. Clustering of the Baltic sea metagenome
Presenting author: Luis Fernando Delgado (luis.delgado.zambrano@scilifelab.se)

P03. Transcriptome Reconstruction in Picea abies
Presenting author: Karl Johan Westrin (westrin@kth.se)

P04. Reserve capacities of yeast metabolism and translation
Presenting author: Rosemary Yu (yuta@chalmers.se)

P05. RSG-Sweden: Sweden's Computational Biology Organisation
Presenting author: Martin Rydén (rsg-sweden@iscbsc.org)

P06. Integrative and Conjugative Elements Stability, Maintenance and Activity
Presenting author: Ibikunle Idowu (msxioi@nottingham.ac.uk)

P07. Deciphering glucocorticoid-mediated stress responses in the pancreatic beta cell using bioinformatic methods
Presenting author: Alexandros Karagiannopoulos (alexandros.karagiannopoulos@med.lu.se)

P08. Predicting crosstalk between phosphosites from large scale quantitative phosphoproteomics data
Presenting author: Augustine C. Amakiri (hlaamaki@liv.ac.uk)

P09. Inferring the effect of DEFA1A3 total genomic copy number on mRNA gene expression levels in human immune cells.
Presenting author: Reem Alhamidi (reem.alhamidi@nottingham.ac.uk)

P10. HumanGEM, an open resource for community curations of genome-scale metabolic model of Homo sapiens
Presenting author: Hao Wang (hao.wang@chalmers.se)

P11. Radiotherapy biomarkers discovery using machine learning approaches
Presenting author: Björn Andersson (bjorn.andersson@medic.gu.se)

P12. A Modeling Approach for BioinformaticsWorkflows: A Design Science Study
Presenting author: Marcela Davila (marcela.davila@gu.se)

P13. A website for exploration and visualization of metabolic networks at metabolicatlas.org
Presenting author: Mihail Anton (mihail.anton@chalmers.se)

Poster Presentations

P14. Digging into viral transcriptomes from publically available datasets, does your data make sense?
Presenting author: Sanna Abrahamsson (sanna.abrahamsson@gu.se)

P15. Ioniser: Searching for additional glycoproteins.
Presenting author: Dagmara Gotlib (dagmara.gotlib@gu.se)

P16. Next-Generation Sequencing and Genotyping for Swedish Research
Presenting author: Sara Sjunnebo (sara.sjunnebo@scilifelab.se)

P17. Benchmarking of four tools designed for pseudotime analysis in scRNA-seq data
Presenting author: Vanja Börjesson (vanja.borjesson@gu.se)

P18. Structural characterization of apomyoglobin unfolding intermediates
Presenting author: Leocadie Henry (leocadie.henry@gu.se)

P19. Large-scale identification of genes important for bacterial community invasion
Presenting author: Johan Bengtsson-Palme (johan.bengtsson-palme@microbiology.se)

P20. UMIerrorcorrect, a pipeline for analyzing targeted sequencing data with UMIs for ultrasensitive detection of low-frequency variants
Presenting author: Tobias Österlund (tobias.osterlund@gu.se)

P21. Identification of Skin Transcription Modules for Analysis of Human Skin Transcriptome
Presenting author: Malin Östensson (malin.ostensson@gu.se)

P22. Vaginal microbiota and HPV infection of adolescent young girls
Presenting author: Yue O. O. Hu (yue.hu@ki.se)

P23. Evaluation of Single-Molecule Long-Read Sequencing Technologies for Structural Variant Detection in Human Genomes
Presenting author: Nazeefa Fatima (nazeefa.fatima@medsci.uu.se)

P24. Multilocus Sequence Typing: Improving Canine Leptospirosis Molecular Epidemiology in Resource-Limited Countries
Presenting author: Adewole Adekola (aadekola18@rvc.ac.uk)

P01. Dissecting Cell-to-Cell Variation in Single-Cell RNA-Seq Data

Johan Gustafsson[1,2], Jonathan Robinson[1,2], Juan S. Inda-Diaz[3], Elias Björnson[1,4], Rebecka Jörnsten[3] and Jens Nielsen[1,2,5]

1 Department of Biology and Biological Engineering, Chalmers University of Technology, Sweden.
2 Wallenberg Center for Protein Research, Chalmers University of Technology, Sweden.
3 Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, Gothenburg, Sweden
4 Department of Molecular and Clinical Medicine/Wallenberg Laboratory for Cardiovascular and Metabolic Research, University of Gothenburg, Gothenburg, Sweden
5 BioInnovation Institute, Copenhagen, Denmark

Presenting author: Johan Gustafsson (gustajo@chalmers.se)

Single-cell RNA sequencing has become a valuable tool for investigating variation between individual cells. However, the information of interest is often obscured by non-biological artefacts contributing to this variation. The dominant contributor is often sampling noise, which is primarily a function of the number of counts per cell and thus provides little information on dataset quality. Quantifying the remaining variation would give more insight into the quality of datasets. We developed a method to partition the intercellular variation of single-cell RNA-Seq data into two components; sampling noise and BTM (biological, technical, and misclassification) variation. We show that it is possible to quantify these two variation types individually without the use of spike-in genes, and that the BTM variation can be compared across datasets. Our method supports an overall variation score for cell populations, cell-wise and gene-wise variation metrics, and the means to evaluate the pool size needed to obtain the same variation in pooled single-cell data as compared to bulk. We found that different public datasets exhibit large differences in BTM variation despite containing cells of the same cell type and from similar conditions, suggesting that the difference in such cases is primarily technical. We propose this method for use as a standard procedure for quality control of single-cell RNA-Seq datasets to ensure that the technical variation, apart from sampling, is comparable to that of other published work.

P02. Clustering of the Baltic sea metagenome

Luis Fernando Delgado[1], Anders F. Andersson[2]

1Department of Biology, Lund University, SE-221 00 Lund, Sweden
2 KTH Royal Institute of Technology, School of Engineering Sciences in Chemistry, Biotechnology and Health, Department of Gene Technology, Science for Life Laboratory, Stockholm, Sweden.

Presenting author: Luis Fernando Delgado (luis.delgado.zambrano@scilifelab.se)

For many environments, biome-specific gene catalogues have been recovered using shotgun metagenomics followed by assembly and gene-calling on the assembled contigs. The assembly can be carried out either by individually assembling reads from each sample or by co-assembling reads from all the samples. The co-assembly approach has the advantage that certain genes displaying too low abundance to be assembled from individual samples may reach enough coverage to be recovered in the co-assembly. However, combining data from many samples often means mixing data from a diversity of closely related bacterial strains, which can lead to difficulties in the assembly. The individually assembled samples approach will minimize the mixing of data from different strains and therefore potentially result in more completely assembled genes. However, the reconstruction of (relatively) identical genes from multiple samples will occur. Therefore, to serve as a reference dataset, sequence redundancy removal is necessary. The aims of this project are: 1) to perform gene calling on assemblies generated using the individual assembly approach, and clustering of the resulting genes, 2) to evaluate the individual assembly gene set and compare it to the co-assembly approach gene set, and 3) to propose a new approach of assembly, mix-assembly, aiming to combine advantages of both individual and co-assembly approaches. In this project, assemblies were obtained from metagenomic reads from 124 samples of the Baltic Sea. Genes were identified on the resulting contigs and sequence redundancy removal was achieved by clustering the gene sequences at protein level. The resulting gene set was evaluated and compared using both statistics on genes (number of non-redundant genes found, the fraction of genes that are predicted to be complete) and on mapping efficiency. The mix-assembly approach resulted in a completer and more extensive non-redundant gene data set than the other approaches.

P03. Transcriptome Reconstruction in Picea abies

Karl Johan Westrin[1], supervised by Warren W. Kretzschmar & Olof Emanuelsson

1. KTH

Presenting author: Karl Johan Westrin (westrin@kth.se)

While the Norway spruce (Picea abies) is important for the Swedish economy, it lacks a complete reference genome -- the published draft is highly fragmented -- and its long juvenile period makes breeding difficult. Early cone setting has been proven associated with expression of a MADS-box gene: DAL19, but no existing transcriptome assembler have managed to reproduce all transcript isoforms of DAL19. A novel de novo-assembly pipeline Abeona is proposed for this issue, outperforming traditional assemblers in reproducing isoforms of DAL19. However, Abeona performs less good when reconstructing the entire transcriptome. Further studies on optimizing Abeona for full transcriptome assembly will be made.

P04. Reserve capacities of yeast metabolism and translation

Rosemary Yu, Kate Campbell, Rui Pereira, Johan Björkeroth, Qi Qi, Egor Vorontsov, Carina Sihlbom, Jens Nielsen

1. Department of Biology and Biological Engineering, Chalmers University of Technology, Sweden
2. Proteomics Core Facility, Sahlgrenska Academy, Gothenburg University, Sweden

Presenting author: Rosemary Yu (yuta@chalmers.se)

Cells maintain reserves in their metabolic and translational capacities as a strategy to quickly respond to changing environments. Here we quantify these reserves by stepwise reducing nitrogen availability in yeast steady-state chemostat cultures, imposing severe restrictions on total cellular protein and transcript content. Combining multi-omics analysis with metabolic modeling, we found that seven metabolic superpathways maintained >50% metabolic capacity in reserve, with glucose metabolism maintaining >80% reserve capacity. Cells maintain >50% reserve in translational capacity for 2,490 out of 3,361 expressed genes (74%), with a disproportionately large reserve dedicated to translating metabolic proteins. Finally, ribosome reserves contained up to 30% sub-stoichiometric ribosomal proteins, with activation of reserve translational capacity associated with selective upregulation of 17 ribosomal proteins. Together, our dataset provides a quantitative link between yeast physiology and cellular economics, which could be leveraged in future cell engineering through targeted proteome streamlining.

P05. RSG-Sweden: Sweden's Computational Biology Organisation

Martin Rydén, Nazeefa Fatima

Lund University, Uppsala University

Presenting author: Martin Rydén (rsg-sweden@iscbsc.org)

The Regional Student Group for Sweden (RSG-Sweden) is a non-profit organization affiliated with the leading International Society of Computational Biology. Our aim is to bring together people from different academic backgrounds and institutions, through networking and knowledge exchange activities. As a team, we work towards promoting Swedish research, getting more people interested in computational biology, bioinformatics, and related areas, and developing interactions between people in industry and academia.

RSG-Sweden is excited to be a part of the Swedish Bioinformatics Workshop 2019. This year, we are organising a session where industrial and/or academic researchers will share experiences that helped shape their careers. Speakers will give a talk about their work, how did they get their current position, what skills they use from their degree, and so on. Each talk will be followed by questions from the audience. We welcome and encourage everyone to join us and contribute to the discussion!

RSG-Sweden currently has five branches across Sweden including Gothenburg. Membership is free and easy; simply visit rsg-sweden.slack.com and sign up with your university e-mail address. We highly encourage students and researchers, at all academic levels, to get involved with their local branch and help achieve our goals of developing the computational biology communities across Sweden.

P06. Integrative and Conjugative Elements Stability, Maintenance and Activity

Ibikunle Idowu, Rob Delahay

1. Nottingham Digestive Diseases Centre (NDDC), Nottingham.
2. University of Nottingham, UK, Nottingham.

Presenting author: Ibikunle Idowu (msxioi@nottingham.ac.uk)

Helicobacter pylori, a gram-negative bacterial pathogen colonises the mucosal layer and the superficial epithelium of the human stomach. The bacteria are found in ~50% of the world's population, with variable prevalence between countries due to ethnicity, geography, age and socioeconomic factors. In developing countries, H. pylori infection is more prevalent across younger ages than in developed countries and increases progressively with age reflecting a cohort phenomenon. Across the world, the disease spectrum of H. pylori infection include gastritis, peptic ulcer disease (PUD) and gastric adenocarcinomas associated with bacterial, host and environmental factors. Several key H. pylori virulence factors have been reported including the cag pathogenicity island (cag-PAI), vacuolating cytotoxin (vac), and more recently products encoded within two integrating and conjugating (ICE) elements, tfs3 and tfs4 which are known to vary in prevalence in different populations. Both tfs3 and tfs4 encodes cluster of gene homologues to VirB/D T4SS in Agrobacterium tumefaciens, of which some have been implicated in disease outcomes in some populations. Both tfs ICEs can co-exist within a single strain of H. pylori as a consequence of inter-strain transfer, although often, one or both ICEs are fragmented, potentially resulting in loss of function. We employ a bioinformatics work-flow and the yeast two hybrid system here to investigate the protein-protein interaction(s) that may mediate Tfs T4SS assembly and investigate the potential for functional cross-complementation of tfs3 and tfs4 ICE T4SS assembly proteins. The results reveal a variety of interactions which indicate functional flexibility of some Tfs T4SS proteins.

P07. Deciphering glucocorticoid-mediated stress responses in the pancreatic beta cell using bioinformatic methods

Alexandros Karagiannopoulos[1,3], Jones K. Ofori[1,3], Lena Eliasson[1,3], Jonathan LS Esguerra[1,2,3]

1. Islet Cell Exocytosis, Department of Clinical Sciences-Malmö, Lund University
2. Bioinformatics and Computational Infrastructure Unit, Lund University Diabetes Centre, Lund and Malmö, Sweden
3. Lund University Diabetes Centre, Lund and Malmö, Sweden

Presenting author: Alexandros Karagiannopoulos (alexandros.karagiannopoulos@med.lu.se)

Glucocorticoids (GCs) are a class of steroid hormones that regulate various metabolic and endocrine processes and are widely used as anti-inflammatory and immunosuppressant drugs. GCs act within the cell mainly via the soluble glucocorticoid receptor (GR), a transcription factor which may activate or repress target genes. Despite the effectiveness of GCs, they have been associated with steroid-induced diabetes mellitus, which is characterized by increased blood glucose levels. Meanwhile, impaired insulin secretion in the pancreatic beta cells (β-cells) plays a crucial role in the progression of the disease. As the impact of GCs on B-cells at the molecular level has not been studied adequately, this project aims at deciphering some of the aspects of GC effect that can potentially lead to B-cell dysfunction. In order to identify potential direct gene targets of the GCs in the B-cells, a customized bioinformatics pipeline was developed. The pipeline is based on the integration of data derived from the differential gene expression analysis of B-cells and human islets treated with the glucocorticoid dexamethasone, and publicly-available data of chromatin immunoprecipitation followed by sequencing (ChIP-seq) analyses. The result revealed both established and novel GR-responsive genes to be among the most potent GR targets in B-cells. Additionally, other properties of GR DNA binding, as demonstrated in other human tissues, were also found to apply to B-cells. Finally, the discovery of transcription factor (TF) binding motifs other than the GRs within the GR binding regions, as well as the existence of conserved regions in the vicinity of these regions, infers the involvement of additional co-regulating TFs in the GC-regulated gene transcriptional mechanism. This study provides a better understanding of the way GCs could potentially disrupt the normal B-cell function that triggers the development of diabetes in patients undergoing GC treatment.

P08. Predicting crosstalk between phosphosites from large scale quantitative phosphoproteomics data

Augustine C. Amakiri[1], Anton Kalyuzhunyy[1], Andrew Jones[1]

1. Institute of Integrative Biology, University of Liverpool, United Kingdom

Presenting author: Augustine C. Amakiri (hlaamaki@liv.ac.uk)

Proteins undergo different kinds of post translational modifications (PTMs) which enable them to perform more specific functions in the body. There are over 480 types of PTM recorded in www.uniprot.com including phosphorylation, acetylation, methylation and many others. Phosphorylation and dephosphorylation of proteins is the hallmark of cell signalling. Several studies have shown that while an interplay between some of these PTMs lead to important biological processes, others have been implicated in disease pathways including cancer and neurodegeneration. In this research, we aim to predict crosstalk events between phosphorylation sites on one hand and phosphorylation and acetylation on the other hand from large scale quantitative phosphoproteomics datasets and sequence motif analysis, indicative that different sites within the same protein are under coordinated control. Our method normalises quantitative signals across multiple conditions (on a per protein basis), enabling us to detect significant correlations in quantitative profiles between sites. Where significant correlations are observed, we can then compare the predicted responsible kinases to infer that the same kinase is phosphorylating multiple sites on the same protein, or potentially that different kinases are working together. In addition to the quantitative analysis, we aim to investigate evolutionary conserved domains, short protein segments (motifs) from protein sequences across different species and kinase recognition motifs. Using these information, we hope to predict more accurately active PTMs sites, crosstalk and kinases that might be responsible for the PTM interplay observed. This will further enhance our knowledge of the signalling events occurring within cells which might be crucial to understanding pathophysiology of diseases. Initial results from quantitative analysis suggest that particular pairs of kinases e.g. CDK1 and CDK2a2 appear to be working together based on enrichments for pairs of phosphosites putatively phosphorylated by these kinases having correlated quantitative patterns.

P09. Inferring the effect of DEFA1A3 total genomic copy number on mRNA gene expression levels in human immune cells.

Reem Alhamidi, John Armour

1. University of Nottingham, UK

Presenting author: Reem Alhamidi (reem.alhamidi@nottingham.ac.uk)

Background: Alpha-defensins 1 and 3 (DEFA1A3) are important antimicrobial peptides that are highly expressed in neutrophils and have important roles in innate immunity. DEFA1A3 is a multi-allelic copy number variable region located on chromosome 8p23.1 that consists of a 19kb full repeat in tandem and a 10 kb partial repeat at the centromeric region. Measuring DEFA1A3 copy number (CN) becomes increasingly challenging as the gene CN increases, therefore finding an application to produce accurate estimates of CN is essential. It was reported that individuals exhibit 3 to 16 copies of DEFA1A3 for the European population; associations of nearby SNPs with IgA nephropathy and periodontitis have been demonstrated by GWAS. It is expected that gene copy number should correlate with the levels of mRNA expression. Investigations of DEFA1A3 copy number relationship with mRNA levels have been inconclusive, with studies conducted so far having small sample size.

Methods and Results: DEFA1A3 CN measurements from the 1000 Genomes Project (1kGP) with 2504 samples across different populations and the BLUEPRINT consortium (BP) data with 197 samples were determined. The DEFA1A3 CN ranged from 3 to 16 copies in 1kGP and 4 to 12 copies in BP. No relationship was found between DEFA1A3 total CN and levels of (RNAseq) expression in neutrophils and T-cells.
Five long amplicons (tiling path) for two samples were loaded to a flongle for sequencing from Oxford Nanopore Technologies (ONT). The amplicons span the 19 kb repeat and flanking regions in the telomeric and centromeric regions of the gene. Each amplicon is ~6kb in length. The 24-hr run produced 10.8k reads yielding 26.76 Mb with only 20.52% of reads passing the quality threshold. The amplicons were separated to determine the haplotypic structures.
Future Work Nanopore amplicon sequencing will be used to reconstruct the full haplotypic structures of NA12878 and NA12760. We are expecting to obtain 5 copies of DEFA1A3 from both samples. The expected ratio of DEFA1:DEFA3 in NA12878 is 4:1 while in NA12760 it is 3:2.

P10. HumanGEM, an open resource for community curations of genome-scale metabolic model of Homo sapiens

Hao Wang, Jonathan L. Robinson, Pinar Kocabas, Pierre-Etienne Cholley, Mihail Anton, Jens Nielsen

1. Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden

Presenting author: Hao Wang (hao.wang@chalmers.se)

Genome-scale metabolic models (GEMs) are valuable tools in studying metabolism by providing scaffolds for integrative analysis of omics data. Significant efforts have been dedicated to GEM development for analyzing metabolic disorders and human diseases. However, the previous human GEMs are disconnected among different sources which impedes further progress. It is therefore necessary to integrate previous resources through an extensive and transparent curation process that involves concerted community efforts.

By integrating the information from HMR2, iHsa, and Recon3D models, we constructed a curated consensus model of human metabolism, Human1, which consists of 13,520 reactions, 10,103 metabolites, and 3,627 genes. The stoichiometric matrix was provided with comprehensive external associations through mapping 89.5% of reactions and 95.2% of metabolites to external standard identifiers. To establish a systematic framework allowing community-wide collaborations, we deployed the model as HumanGEM repository on GitHub.

The insufficient use of standard identifiers in previous models was tackled by extensive association with external databases in a semi-automated process through using the MNXref namespace. This led to the identification and removal of 7,995 duplicated reactions and 3,155 duplicated metabolites; revision of 1,869 metabolite formulas; re-balancing of 2,061 reaction equations. The subsequent curations include correction of reversibility for 59 reactions; and the inactivation or removal of 603 mass- and/or energy-imbalanced reactions. Enzyme complexes from previous models were combined and integrated with the CORUM mammalian protein complexes database for the refinement of gene-reaction associations. The entire curation process has been implemented with a total of 23 releases, through a task-driven and issue-guided workflow.

This workflow converted the GEM reconstruction and curation pipelines into a software development-like process and transformed the complex curation tasks into a well-organized and transparent information flow, which ensures HumanGEM to serve the community as open and inclusive biomedical research resources with continuous upgrading with external biological databases.

P11. Radiotherapy biomarkers discovery using machine learning approaches

Björn Andersson[1], Peidi Liu[1], Britta Langen[2], Eva Forsell[2], Marcela Davila[1]

1. Bioinformatics Core Facility, The Sahlgrenska Academy at the University of Gothenburg, Göteborg, Sweden
2. Dept. of Radiation Physics, Inst. of Clinical Sciences, Sahlgrenska Cancer Center, The Sahlgrenska Academy at the University of Gothenburg, Göteborg, Sweden

Presenting author: Björn Andersson (bjorn.andersson@medic.gu.se)

Background: The radiation research and radiotherapy treatments heavily depend on the accuracy of the relation among the absorbed dose, radiation duration and the biological effects on tissues after the irradiation. Misconducting of the dose-response will cause severe problems of under-treatment or damage to healthy tissues. Thereafter, certain biomarkers that are relevant to dose-response are needed in reducing the risks.
Aim: The purpose of the project is to develop an unbiased machine learning tool/pipeline using different omics data (including transcriptomic and proteomic) to determine a panel of biomarkers that can reflect the dose-dependency for radiation treatments.
Methods: Transcriptomic data (microarray) and proteomic data (mass spectrometry) from experiments of animals (mouse & rat) that were radiated with different radiation (iodine-131, astatine-211, 177Lu-octreotate and 177Lu-chloride) and absorbed dose, radiation duration were evaluated in the following tissues; thyroid, kidneys, liver, lungs, spleen and blood.
Using Genetic Algorithm and K-nearest Neighborhood (GA/KNN, Li L et al 2001) machine learning approach to select top ranked genes/proteins that could classify radiated samples from untreated controls based on tissue types, dose and radiation duration.
Results: 1. both differential expression analysis of the omics data and the GA/KNN algorithm, identified potential biomarkers. 2. Regression with identified biomarkers showed that the predicted doses are well correlated to the true doses.
Conclusion: We found significantly regulated transcripts and shared transcripts between tissues and/or conditions. GA/KNN is a powerful iterative machine learning algorithm that could select a small panel of features (in this case genes and proteins) that are potential biomarkers with radiation doses. This biomarkers can then be evaluated in lab experiments to confirm their functions.

P12. A Modeling Approach for BioinformaticsWorkflows: A Design Science Study

Laiz Heckmann Barbalho de Figueroa, Rema Salman, Jennifer Horkoff, Soni Chauhan, Marcela Davila, Francisco Gomes de Oliveira Neto and Alexander Schliep

1. University of Gothenburg, Gothenburg, Sweden
2. Chalmers University of Technology, Gothenburg, Sweden

Presenting author: Marcela Davila (marcela.davila@gu.se)

Bioinformaticians execute daily, complex, manual and scriptedworkflows to process data. There are many tools to manage and conduct these workflows, but there is no domain-specific way to textually and diagrammatically document them. Consequently, we create methods for modeling bioinformatics workflows. Specifically, we extend the UnifiedModeling Language (UML) Activity Diagram to the bioinformatics do-main by including domain-specific concepts and notations. Additionally,a template was created to document the same concepts in a text for-mat. A design science methodology was followed, where four iterationswith seven domain experts tailored the artefacts, extending concepts andimproving usability, terminology, and notations. The UML extension re-ceived a positive evaluation from bioinformaticians. However, the writtentemplate was rejected due to the amount of text and complexity.

P13. A website for exploration and visualization of metabolic networks at metabolicatlas.org

Mihail Anton[*1] , Pierre-Etienne Cholley[*1] , Jonathan L. Robinson[2,3], Lena Hansson[1,4], L. Thomas Svensson[1], Jens Nielsen[2,3,5,6]

1. Department of Biology and Biological Engineering, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden
2. Department of Biology and Biological Engineering, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.
3. Wallenberg Center for Protein Research, Chalmers University of Technology, Kemivägen 10, Gothenburg, Sweden.
4. Novo Nordisk Research Centre Oxford, Oxford, UK
5. Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
6. BioInnovation Institute, Ole Maales Vej 3, DK2200 Copenhagen, Denmark
** These authors contributed equally to this work*

Presenting author: Mihail Anton (mihail.anton@chalmers.se)

Metabolic Atlas integrates open source genome-scale metabolic models (GEMs) of human and yeast for easy browsing and visualization. A wide range of attributes, including reaction equations, metabolite formulas, gene rules and subsystem contents, are presented in tabular and map views. For each of the integrated models, Metabolic Atlas integrates manually created 2D maps. Also, it creates 3D graphs for both compartments and subsystems. Via Interaction Partners, a dynamically generated and customizable graph of associations of genes and metabolites, one can interact with a restricted part of the metabolic network, or further expand the interaction partners of any element already on the graph. Moreover, RNA expression data from the Human Protein Atlas can be overlaid onto the graph. The GEM repository contains over 350 genome scale metabolic models. With the Global search feature one can search for any model component with the use of advanced filtering, being able to export the results.

P14. Digging into viral transcriptomes from publically available datasets, does your data make sense?

Sanna Abrahamsson[1,2], Yarong Tian[1], Guojiang Xie[1], Harsha Meghadri[1], Jonas Carlsten[3], Ka-Wei Tang[1,4]

1. Sahlgrenska Cancer Center, Department of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Sweden
2. Bioinformatics Core Facility, Sahlgrenska Academy, Gothenburg, Sweden
3. Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm, Sweden
4. Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Sweden

Presenting author: Sanna Abrahamsson (sanna.abrahamsson@gu.se)

EBV is a double stranded DNA virus that have infected 90 % of the human population. In most cases the virus does not cause any symptoms, but sometimes EBV infections causes different cancer types such as gastric adenocarcinoma, Burkitts' lymphoma and nasopharyngeal carcinoma. How EBV is causing cancer is not known. EBV RPMS1 is a viral long noncoding RNA postulated to be the major transcript expressed in some tumors. However, this claim has been disputed and RNA from the antisense strand may be the origin for the virus gene expression.

The main goal in this analysis was to screen multiple publically available cancer datasets to map the EBV transcriptome and to measure the true expression signal within the RPMS1 region.

It is easy to run through a common RNA-seq pipeline, glance through the results and make the conclusions, but does the result make sense? The EBV genome is very complex. There are overlapping genes within the RPMS1 coordinates on the antisense strand and miRNA within the RPMS1 introns on the sense strand. The main proportion of the downloaded datasets were unstranded which means that if only the mapping position of the read is used it is impossible to determine the origin of a read within overlapping genes.

To prove that the expression signal mainly arise from RPMS1 a detailed analysis was performed using poly A signal, 5' mapping and splice read analysis with linkage to the exon coordinates of RPMS1. Here we show the importance of understanding your input data and what to look for when investigating complex genomes.

P15. Ioniser: Searching for additional glycoproteins.

Dagmara Gotlib, Ekaterina Mirgorodskaya, Luciano Fernandez-Ricaud

1. University of Gothenburg's Core Facilities

Presenting author: Dagmara Gotlib (dagmara.gotlib@gu.se)

The Identification of glycosylated proteins is a challenging task in the proteomics field. Glycopeptides are commonly presented by multiple glycoforms. Identification of all the forms is often difficult, as not all of them are available in databases.

To facilitate the process of finding additional glycoforms from already known ones, we developed the Ioniser. This tool uses the information on the fragment ion spectra from identified glycoforms to aid at the identification of non-identified glycoforms (either not present in the database or novel structures). The program facilitates recognition of differently charged fragment ions ([M+H]+1, [M+H]+2, etc) for their correct assignment

Ioniser is a small and useful tool that processes and filters large amounts of mass-to-charge ratio and abundance data allowing the user to identify additional glycosylated proteins based on user-specified parameters. The tool then outputs the relevant masses found within the given ppm tolerance.

P16. Next-Generation Sequencing and Genotyping for Swedish Research

NGI Sweden

1. NGI Sweden

Presenting author: Sara Sjunnebo (sara.sjunnebo@scilifelab.se)

The National Genomics Infrastructure (NGI) is hosted by Science for Life Laboratory. NGI is one of SciLifeLabs largest technical platform both in terms of number of projects and number of users. It provides access to technology for massively parallel/next generation DNA sequencing, genotyping at all scales and associated bioinformatics support to researchers based in Sweden.

P17. Benchmarking of four tools designed for pseudotime analysis in scRNA-seq data

Vanja Börjesson

1. University of Gothenburg, Sahlgrenska Academy, Bioinformatics Core Facility

Presenting author: Vanja Börjesson (vanja.borjesson@gu.se)

Cells differentiate and specialize to do essential things in our body. Some cells produce proteins as important building blocks for e.g. tissues and organs, and others might specialize to perform a function such as communicating through chemicals. In research, studying the function of cells are very important in order to understand different phenotypes such as diseases.
Single cell RNA sequencing (scRNA-seq) allows researchers to measure the expression levels of RNA in individual cells. Studying how cells change in their specialization, i.e. the change in gene expression, over time is also very important to understand the whole cell process. Several bioinformatics tools exist for analyzing cell differentiation, all based on different algorithms for dimensional reduction, clustering and trajectory creation.

Four of the most used tools for pseudotime analysis; Monocle, SLICER, destiny and scanpy, were in this project compared. We defined the most efficient tool based on resulting trajectory, running time, required computational resources, and ability to select parameters for pre-work such as filtering and normalization.

The results shows that the running time and ability to reduce noise through pre-work differ a lot between the tools and also the possibility to set origin cells or number of branches expected. Monocle and Scanpy were the tools that was easiest to use and produced the most clear output figures of the trajectories.

P18. Structural characterization of apomyoglobin unfolding intermediates

Leocadie Henry, Matthjis Panman, Linnea Isaksson, Elin Claesson, Irena Kosheleva, Robert Henning, Sebastian Westenhoff, Oskar Berntsson

1. Department of Chemistry and Molecular Biology, University of Gothenburg, 40530 Gothenburg, Sweden
2. Center for Advanced Radiation Sources, The University of Chicago, Chicago IL 60637, USA

Presenting author: Leocadie Henry (leocadie.henry@gu.se)

An increased interest in the understanding of protein misfolding and unfolding mechanism has particularly been inspired in the last twenty years with the increasing prevalence of neurodegenerative disease. Unfortunately, the study of protein unfolding is often difficult to carry out with the currently available methods due to the time scale of the folding events and the transience of their intermediates. Apomyoglobin has been a well-studied test sample for small, single domain and globular proteins. Various studies have been done by NMR, SAXS, fluorescence and MD simulations to try to underpin the unfolding process of apomyoglobin. So far, three states consisting of a native, a molten globule and an unfolded state have been directly observed. However, very little is known about how and when the transition between those states occur. Here, we show that it is possible to use nano- to millisecond time resolved X-ray solution scattering in combination with molecular dynamics (MD) simulations to follow and characterize the early events of apomyoglobin unfolding. We report here that the formation of the unfolded state is actually a succession of discrete intermediate steps. The data shows an increase in radius of gyration, flexibility and solvent accessible surface in those intermediates over time as well as a loss of helicity and hydrogen bonds. We observe that water enters the structure from 3 μs and that conserved non-functional residues are involved in the maintenance of the 3D structure. The characterization of those states led us to the postulate that there is a rapid formation of a dry molten globule followed by a wet molten globule and finally an unfolded state. This study shows that we are able to directly observe secondary and tertiary structural changes in the early events of unfolding, bringing new insight into small globular protein unfolding mechanism.

P19. Large-scale identification of genes important for bacterial community invasion

Johan Bengtsson-Palme[1,2,3], Adriana Osinska[2,4], Manuel F Garavito[3], Amanda Hurley[3], Jeyaprakash Rajendhran[3,5], Emil Burman[1,2], Erik Kristiansson[2,6], Gabriel L Lozano[3,7], Jo Handelsman[3]

1. Dept. of Infectious Diseases, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg
2. Centre for Antibiotic Resistance Research at University of Gothenburg
3. Wisconsin Institute for Discovery, University of Wisconsin-Madison, USA
4. University of Warmia and Mazury in Olsztyn, Poland
5. Dept. of Genetics, School of Biological Sciences, Madurai Kamaraj University, India
6. Dept. of Mathematical Sciences, Chalmers University of Technology
7. Dept. of Molecular, Cellular and Developmental Biology, Yale University, USA

Presenting author: Johan Bengtsson-Palme (johan.bengtsson-palme@microbiology.se)

Rapid antibiotic resistance development threatens our ability to perform modern healthcare. Alarmingly, there are indications that antibiotics not only drive increased resistance, but also enhance bacterial invasiveness and virulence. Thus, future pathogens could become multi-resistant ”superbugs” for which we not only lack treatment options, but also cause more severe infections and spread more easily. The ability of bacteria to establish themselves in new environments (colonization) and successfully become a member of already established communities (invasion) largely determines the outcome of pathogen interactions in the human body and thus partially dictates the infectiousness of pathogenic bacteria. Unfortunately, very little is known about how antibiotic exposure affects the ability to colonize niches and invade established bacterial communities. In this study, we have investigated how sub-lethal concentrations of antibiotics affect bacterial colonization and invasion ability and which specific genes that are important for these processes. To address this, we have used a technique for high-throughput sequencing of tags from transposon mutants called InSeq. In this way, we have been able to identify thousands of genes that are involved in increased or reduced fitness in community invasion compared to in colonization processes. Furthermore, we have been able to pinpoint a subset of 17 genes that not only are important for invasion ability, but also are further enriched in the presence of antibiotics. Some of these genes seem to be involved in virulence, while others seem related to antibiotic resistance traits. The latter category of genes are candidates to be mobilized and spread in bacterial populations, and could therefore become future resistance factors in pathogens, directly connecting environmental antibiotic exposure to human health risks.

P20. UMIerrorcorrect, a pipeline for analyzing targeted sequencing data with UMIs for ultrasensitive detection of low-frequency variants

Tobias Österlund[1,2,3], Stefan Filges[1,3], Gustav Johansson[1,3], Anders Ståhlberg[1,2,3]

1. Sahlgrenska Cancer Center, Department of Laboratory medicine, Institute of Biomedicine, University of Gothenburg, Sweden.
2. Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, 413 45 Gothenburg, Sweden.
3. Wallenberg Centre for Molecular and Translational Medicine, University of Gothenburg, Gothenburg, Sweden.

Presenting author: Tobias Österlund (tobias.osterlund@gu.se)

Targeted sequencing with Unique molecular index (UMIs) offers a way to perform "error-free" sequencing of genomic targets of interest. Each original molecule is tagged with a unique barcode which then can be traced in the sequencing data to remove PCR-duplicates, remove sequencing errors and to get an estimate of the number of molecules in the original sample by counting the number of unique barcodes.
This technique can offer ultra-sensitive variant detection, for example when targeting cell-free circulating tumor DNA (ctDNA) from liquid biopsies, where the mutations of interest often are at frequencies and 1% or below.

Here we present UMIerrorcorrect, a bioinformatics pipeline for analyzing targeted sequencing data with UMIs. The pipeline is easy to use and includes preprocessing of UMI-reads, UMI clustering, error correction, creation of consensus reads and low-frequency variant calling.

We used the UMIerrorcorrect pipeline to analyze targeted amplicon sequencing data using SiM-Sen-Seq (1). The sensitivity for the identification of single nucleotide variants and small indels with 0.25% variant allele frequency was 60.0% and the specificity was 100%. For detecting variants with 1% variant allele frequency the sensitivity was 100% and the specificity was 100%. The pipeline is open source and will be available at www.github.com/tobbeost/umierrorcorrect/

1. Ståhlberg A, Krzyzanowski PM, Egyud M, Filges S, Stein L, Godfrey TE. Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing. nature protocols. 2017;12(4):664.

P21. Identification of Skin Transcription Modules for Analysis of Human Skin Transcriptome

Emili A.A. Verge[1], Ali M. Harandi[1], Malin Östensson[1,2]

1. Department of Microbiology and Immunology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
2. Bioinformatics Core Facilities, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden

Presenting author: Malin Östensson (malin.ostensson@gu.se)

Introduction: Cutaneous delivery of vaccines and immunomodulators has proven to be efficient to mount immunity in humans. Understanding the molecular signatures of skin-derived immunity in humans can inform rational development of vaccines and immunomodulators/adjuvants for delivery through the skin.
Integrative network modeling is a powerful tool to identify the transcriptional signatures induced following infection and vaccination. This systems biology approach has recently been successfully used to identify blood transcription modules (BTM) in humans. Herein, we report the development of a novel framework consisting of a set of Skin Transcription Modules (STM) for human through large-scale network integration of publicly available data.

Materials and Methods: The experimentally inferred network from skin transcriptomic studies has been combined with bibliography and interactome derived networks. The resultant master network has been analyzed with the MCODE topological algorithm in Cytoscape to identify densely connected gene clusters. Modules have been annotated according to the top matched gene ontology and UniProt terms, KEGG, Reactome and Biocarta databases.

Results: Data integration has resulted in the creation of a final master network consisting of 20,532 genes and 432,472 interactions. From the master network, 349 modules have been pulled out by the MCODE algorithm, from which 67 have a size equal to or higher than 10 genes. 20 modules have been matched to a KEGG, Reactome or Biocarta pathway with a gene overlap of 60% or less.

Conclusions: We herein report, for the first time, the development of STM that can provide a context- specificity in the analysis of transcriptomics of vaccines, adjuvants and immunomodulators in human skin. Further improvements of STM is underway to remove noise and increase the range of biological processes and molecular functions included.

P22. Vaginal microbiota and HPV infection of adolescent young girls

Yue O. O. Hu[1*], Liqin Cheng[1*], Johanna Norenhag[2*], Nele Brusselaers[1], Emma Fransson[1], Andreas Öhrlund-Richter[3], Unnur Gudnadottir[1], Pia Angelidou[1], Yinghua Zha[1], Marica Hamsten[1], Ina Schuppe Koistinen[1], Matts Olovsson[2], Lars Engstrand[1], Juan Du[1]

1. Karolinska Institutet,
2. Uppsala University

Presenting author: Yue O. O. Hu (yue.hu@ki.se)

Changes in vaginal microbiota with the absence of Lactobacilli and increased microbial diversity facilitate sexually transmitted infections. Human papillomavirus (HPV) is the most common sexually transmitted virus. To define the HPV infection associated microbial community in Sweden, we analyzed the microbial community composition of Swedish young women respecting to HPV infection status and 27 HPV subtypes. 16S rRNA gene sequencing was used to characterize the vaginal samples from a youth clinic which covers age 14 to 22 (n=156), and health care centers from age 23-29 (n=126). Microbiota alpha diversity analysis revealed a significantly increased diversity in the HPV+ group compared to HPV- group, especially from young girls infected by oncogenic HPV sub-types. The vaginal microbiome in HPV+ women was characterized by higher levels of signature bacteria, such as Gardnerella compared to HPV- women. Our results suggest HPV infection is associated with increased vaginal microbiota diversity and potential microbial markers can be used for HPV infection.

P23. Evaluation of Single-Molecule Long-Read Sequencing Technologies for Structural Variant Detection in Human Genomes

Nazeefa Fatima, Adam Ameur

1. National Genomics Infrastructure, Science for Life Laboratory, Uppsala

Presenting author: Nazeefa Fatima (nazeefa.fatima@medsci.uu.se)

Chromosomes can undergo various changes such as large deletions and/or insertions, resulting in structural variation differences between individuals. Structural variants (SVs) are a common source of variability in the human genome and are known to be associated with several diseases. SVs often involve complex genomic rearrangements that are difficult to resolve using short read sequencing technologies. New approaches enabled by the latest generation of long-read single-molecule sequencing instruments, provided by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), can produce a sufficient amount of data to enable SV detection across entire human genomes to a reasonable cost.

Previously, we performed PacBio sequencing of two Swedish human genomes, as part of the SweGen 1000 Genomes project (https://swefreq.nbis.se) and uncovered over 17,000 SVs per individual (Ameur et al, 2018). A majority of these SVs were not detectable in short reads. As a follow-up, we have now generated data for the same individuals on ONTs PromethION system, a new nanopore-based platform known for its higher throughput as compared to PacBio.

We present a pilot study that evaluates nanopore data derived from whole-genome sequencing on PromethION in comparison to the Single-Molecule Real-Time (SMRT) reads obtained from the PacBio RS II platform. We performed comparative analyses of single- molecule technologies in a context of mappability, and SV detection that resulted in an average of 17k and 24k variants across nanopore and SMRT datasets, respectively. The results will be useful for the large-scale SweGen project, while the study serves as a bioinformatics pipeline for future long-read data analyses and sets a basis for what to consider when designing future PromethION experiments.

P24. Multilocus Sequence Typing: Improving Canine Leptospirosis Molecular Epidemiology in Resource-Limited Countries

Adewole Adekola[1], Vicki Chalker[2], David Brodbelt[1], Brian Catchpole[1]

1. Royal Veterinary College, London, UK
2. Public Health England, UK

Presenting author: Adewole Adekola (aadekola18@rvc.ac.uk)

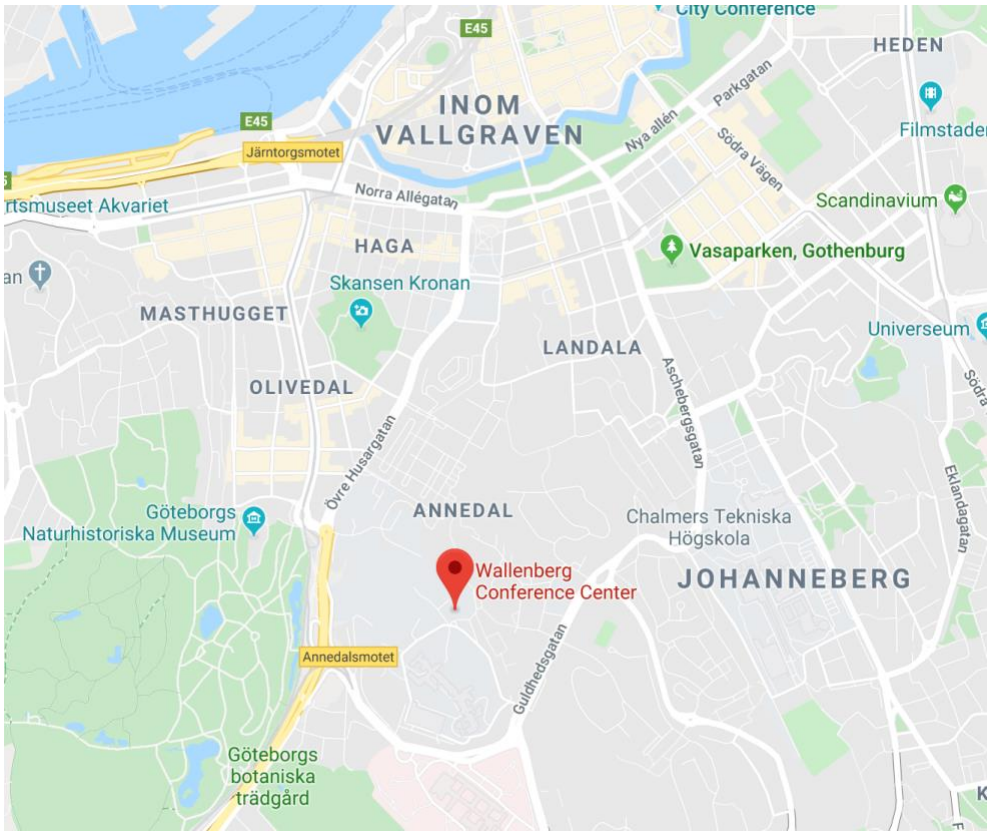
Canine leptospirosis is reported to be growing in prevalence with the emergence of new serovars and serogroups, which might represent a significant risk to public health and livestock production. The Leptospira taxonomic complexity, fastidious growth of the microorganism, poor serogroup cross-reactivity and the difficulty in discriminating between infected and vaccinated animals have impacted on the effectiveness of existing diagnostic tests. In addition, there is insufficient information on leptospirosis epidemiology (clinical and molecular), awareness and understanding of endemic pathogenic leptospire strains, particularly in resource-limited countries such as Nigeria. Multilocus sequencing typing (MLST) has been established as an important genotyping technique, designed to allow identification of allelic variations in selected bacterial genes. The choice of MLST for the genetic speciation of Leptospira is also predicated on the reproducibility, robustness, consistency, portability and how amenable the technique is to low and high-throughput scale. Molecular techniques for determining Leptospira genomospecies also avoids the ambiguity associated with the serovar-based classification. In addition, the technique also lends good applicability for the epidemiological investigation of host-pathogen spill-over and understanding of leptospirosis outbreaks. Use of the published MLST scheme 1 for interrogation of diagnostic samples from affected dogs is likely to be applicable for the identification of the most common pathogenic leptospira species in the canine population and which are of potential zoonotic risk. The incorporation of a nested amplification approach will potentially enhance the sensitivity of the assay when applied to clinical samples of varying quality/quantity.

Travel Instructions

The workshop will take place in the Conference Centre Wallenberg, Gothenburg.

Find the way to the conference

The address of the venue is Medicinaregatan 20 A, Konferenscentrum Wallenberg, Göteborg, Sweden.



Nearest tram stop

The closest public transport stop is **Medicinaregatan**. Trams 6,7, 8, and 13 and bus 753 stop by. The public transport system is managed by **www.vasttrafik.se**. From the stop, there is a 5-minute walk to the conference center.

How to reach central station

From Medicinaregatan, Trams 6, 7, and 13 go to Central Station.

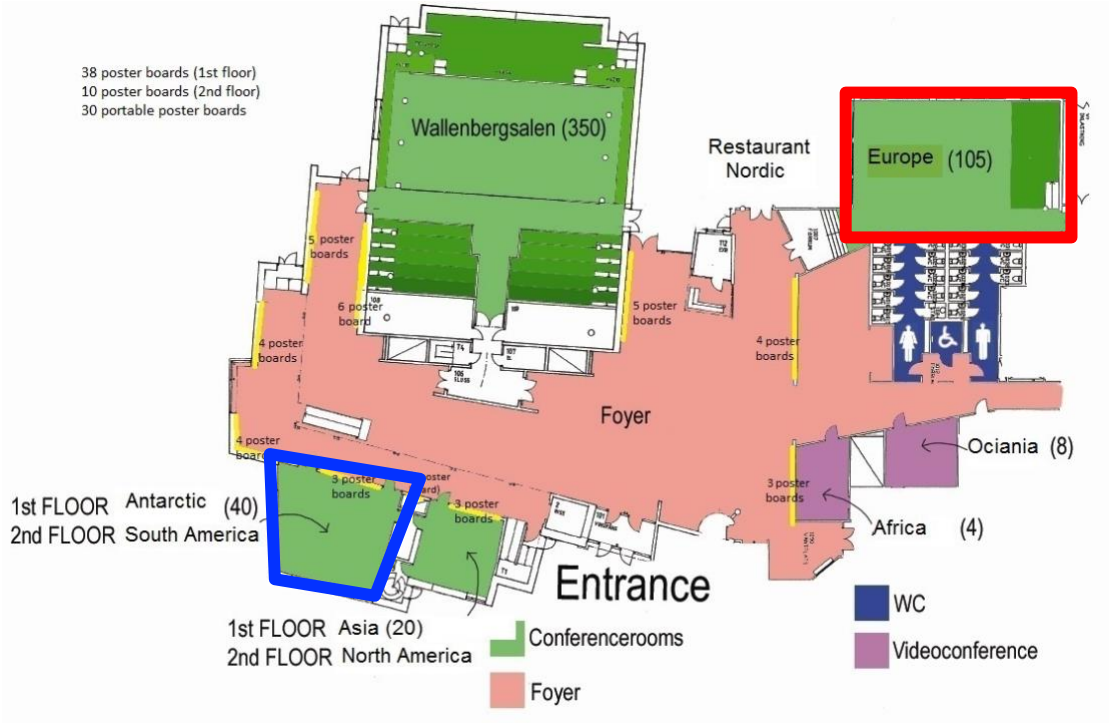
Where to get Västtrafik tickets

Buy your tickets at one of Västtrafik shops or at ticket sales points, Pressbyrån. A convenient alternative is also to use the app **Västtrafik To Go**.
NOTE: You can **not** pay with cash on the trams or buses.
See http://www.vasttrafik.se/ for more information.

Find dinner venue

The dinner takes place in the Lyktan Restaurant, at the same location as the conference.

Map of the Wallenberg Conference Center



The presentations will take place in Europe (outlined in red). One of the workshop sessions will take place in Antarctica (outlined in blue).