swedish bioinformatics workshop 2014

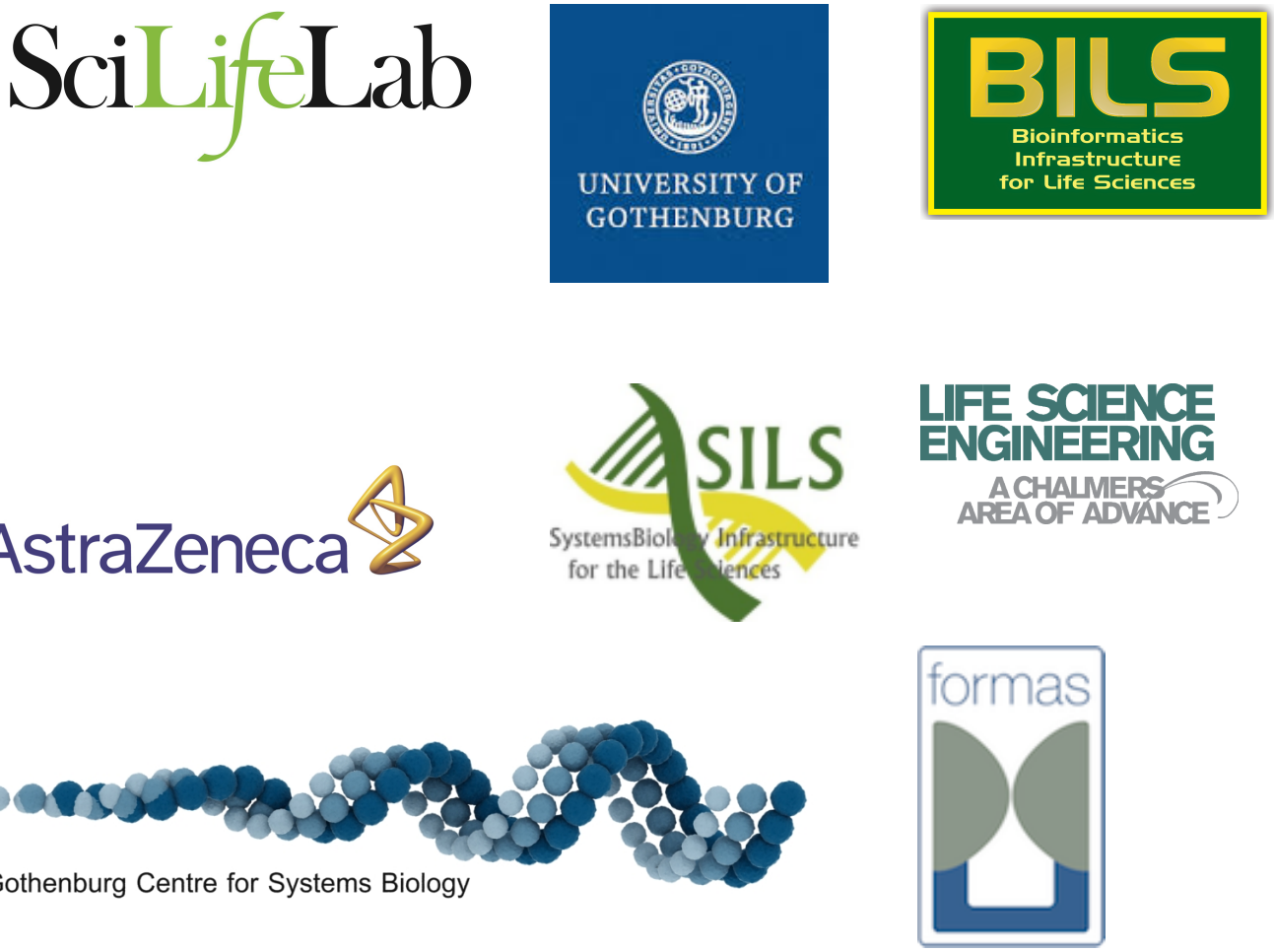
Program and Abstract Book

October 23-24, 2014
Conference Centre Wallenberg
Göteborg

Cover Page Photo: Patrik B (pebe)

Swedish Bioinformatics Workshop
23-24 October 2014

Conference Centre Wallenberg, Gothenburg, Sweden



This workshop is supported with an educational grant from AstraZeneca.

Welcome to the Swedish Bioinformatics Workshop 2014!

The organizing committee would like to welcome you to Göteborg for this yearly occasion that gathers researchers in bioinformatics from Sweden and abroad.

This year we anticipate over 100 participants to the workshop and we are thrilled at the idea to share this experience with you. In this two day event, an exciting program is scheduled ahead of us that features some exclusive highlights:

- First of all, three *keynote speakers*, top-notch European principal investigators in bioinformatics that will share their latest work with us.

- Second, the *workshops*, organized in parallel sessions and led by six expert researchers who will entertain a round table discussion on particularly compelling and intriguing bioinformatics questions.

- Finally and most importantly, *young researchers* in bioinformatics. SBW is devoted to give a chance to PhD students and postdocs to present their research to their fellows. We have *plenary and parallel sessions* for a total of 15 oral presentations as well as *poster presentations* for an additional 25 projects.

Hopefully, you will join us for the SBW dinner on October 23rd, that will be served at the River Café restaurant on the pier of Göta älv.

We are also looking forward to awarding a prize for the best oral and poster presentation in SBW.

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

Thanks for joining SBW 2014!

The organizing committee

Organizing Committee

The organization of SBW2014 is a joint collaboration between PhD students and postdocs from Chalmers & GU.

Fredrik	**Boulund**	**fredrik.boulund@chalmers.se**
Johan	Bengtsson-Palme	johan@microbiology.se
Francesco	Gatto	gatto@chalmers.se
Shaq	Hosseini	Shaghayegh.hosseini@chalmers.se
Anna	Johnning	anna.johnning@chalmers.se
Kaisa	Thorell	kaisa.thorell@gu.se
Leif	Väremo	varemo@chalmers.se

Content

Keynote Speakers

KS1. From single cells to whole genomes

Joakim Lundeberg, Gene Technology, KTH, Stockholm

Joakim Lundeberg is Professor in Gene Technology at KTH, Stockholm and Director of the Genomics platform at Science for Life Laboratory, Stockholm. His research focus is innovative technology development for DNA/RNA analysis.

KS2. The Biocode - DNA Sequencing the Earth

Dawn Field, NERC Centre for Ecology and Hydrology

Dawn Field is Head of the Molecular Evolution and Bioinformatics Group at the NERC Centre for Ecology and Hydrology. Her research is in comparative microbial genomics and microbial diversity using bioinformatics approaches and next-generation sequencing techniques. An area of interest is the function and evolution of the "unknown genome" which has yet to be characterized in (meta) genomic data sets.

KS3. Bioinformatics, genomics and metagenomics applied to farm animal health and food security

Mick Watson, Roslin Institute, University of Edinburgh

Mick Watson is Director of ARK-Genomics at the Roslin Institute, University of Edinburgh. His research focus revolves around the use of computational and mathematical techniques to understand genome function with an emphasis on transcriptomic analysis in systems of relevance to animal health and food security.

Workshops

Workshop 1. Fusion gene detection by molecular characterization of solid tumors

Moderator: Anders Edsjö (anders.edsjo@gu.se)

Workshop 2. Reference-free variant analysis --- determining microbial population structure from metagenomes

Moderator: Alexander Eiler (alexander.eiler@ebc.uu.se)

Workshop 3. From clinical observations to molecular functions

Moderator: Daniel Muthas, AstraZeneca (daniel.muthas@astrazeneca.com)

Abstract: The workshop will focus on how to best link clinical data (both literature/electronic health records/RWE) to underlying "omics"-data, and how this could serve as a changing paradigm in drug discovery.

Workshop 4. Future challenges of systems biology --- what is the way forward?

Moderator: Marija Cvijovic (marija.cvijovic@chalmers.se)

Workshop 5. Beyond the exome --- how can we determine the relevance of non-coding mutations in tumors?

Moderator: Erik Larsson (erik.larsson@gu.se)

Abstract: Today, there are efficient bioinformatical methods for uncovering mutations in coding regions that contribute to cancer. Far less is known about the putative contributions from non-coding sequences. The idea is to discuss bioinformatical strategies for assessing the possible importance of somatic alterations in non-coding DNA.

Workshop 6. Integrity and reliability of DNA sequences in biology

Moderator: Henrik Nilsson (henrik.nilsson@bioenv.gu.se)

Abstract: DNA sequences are a routine source of scientific information in biology. Increasing data volumes pave the way for automated solutions to sequence analysis, but these solutions do not always meet quality expectations. This workshop discusses some typical cases of compromised DNA sequences and tries to find solutions to these problems.

Oral Presentations

OP01. **An Automated De Novo Assembly Pipeline at NGI-Stockholm**
Presenting author: Francesco Vezzi (francesco.vezzi@scilifelab.se)

OP02. **Using Piper to analyze thousands of genomes**
Presenting author: Johan Dahlberg (johan.dahlberg@medsci.uu.se)

OP03. **Large-scale analysis of bacterial genomes and plasmids for antibiotic, biocide and metal resistance genes and their co-selection potential**
Presenting author: Chandan Pal (chandan.pal@gu.se)

OP04. **How bioinformatics designed a plasma and urine marker to detect kidney cancer**
Presenting author: Francesco Gatto (gatto@chalmers.se)

OP05. **Metagenomics reveal changes in antibiotic resistance gene abundance in gut bacterial communities resulting from international traveling**
Presenting author: Johan Bengtsson-Palme (johan.bengtsson-palme@gu.se)

OP06. **A kinetically constrained FBA model predicts shifts in Saccharomyces cerevisiae's metabolic strategy**
Presenting author: Avlant Nilsson (avlant@hotmail.com)

OP07. **A first truly systems level mechanistic model – unravelling the gene regulation of Th2 differentiation**
Presenting author: Mattias Köpsén (matt.kopsen@gmail.com)

OP08. **Clustering Metagenomic Contigs using CONCOCT**
Presenting author: Johannes Alneberg (johannes.alneberg@scilifelab.se)

OP09. **Correction of long PacBio reads using short Illumina reads**
Presenting author: Sylvie Tesson (sylvie.tesson@gu.se)

OP10. **The influence of glucose uptake on Snf1/Mig1 Signalling dynamics** at the single cell level
Presenting author: Niek Welkenhuysen (niek.welkenhuysen@gmail.com)

OP11. **A validated gene regulatory network and GWAS to identify early transcription factors in T-cell associated diseases**
Presenting author: Mika Gustafsson (mika.gustafsson@liu.se)

OP12. **Computational elucidation of the gene expression and regulation in insulin resistance based on the Cap Analysis of Gene Expression fat tissue profiling**
Presenting author: Olga Hrydziuszko (olga.hrydziuszko@gmail.com)

OP13. **Bioinformatic and structural approaches to study cytoskeletal proteins in Mycobacterium bovis BCG and M. marinum M**
Presenting author: Nicole Selzer (nicole.selzer@molbiol.umu.se)

OP14. **Systematic search for transcription-modulating somatic mutations in regulatory DNA regions across 505 tumor genomes**
Presenting author: Johan Fredriksson (johan.fredriksson@gu.se)

OP15. **HCSD: Human Cancer Secretome Database**
Presenting author: Amir Feizi (feizi@chalmers.se)

OP01. An Automated De Novo Assembly Pipeline at NGI-Stockholm

Francesco Vezzi[1] and National Genomics Infrastructure[1]

1. SciLifeLab, Department of Biochemistry and Biophysics, Stockholm University, 171 65 Stockholm, Sweden

Presenting author: Francesco Vezzi (francesco.vezzi@scilifelab.se)

As high-throughput sequencing data penetrates more and more areas of genomic study, many general and specialized software tools have been developed to facilitate analysis. In particular, the decreasing cost of sequencing and the increasing length of reads has prompted a growing number of researchers to sequence novel genomes with the aim of producing a *de novo* assembly. Despite the advances in both core sequencing technologies and related analysis software, *de novo* assembly is still an open problem, and efficient and accurate assembly of large and complex genomes remains a chimera.

The most challenging step continues to be the running of the *de novo* assembly software; however, to produce a high-quality sequence, this step alone is not sufficient. Several down- and upstream analyses must be performed (e.g. data QC, assembly evaluation, etc.) to sanitize and validate data using an array of different tools. Here we present an automated *de novo* assembly pipeline, NGI-deNovo, currently in use at the Swedish National Genomic Infrastructure (NGI). NGI-deNovo simplifies the production of a first draft assembly and can be used by a researcher lacking either experience in *de novo* assembly or the requisite computational infrastructure.

NGI-deNovo is divided into three steps: data validation, genome assembly, and assembly evaluation. In the first step, we check data quality, remove adaptor sequences, and build kmer plots to check data complexity and heterozygosity levels; if a closely-related sequence is available, reads are also aligned and alignment statistics are checked. In the second phase (*de novo* assembly), several different assemblers are selected based on the qualities of the genome being sequenced and are used to process the available libraries. Specifically, MaSuRCA, SPADES, and CABOG are preferred for bacterial genomes, while ALLPATHS, SOAPdenovo, and ABySS are used with complex genomes. The incorporation of optical mapping data (OpGen) is also supported. In the last phase (assembly evaluation), reads are aligned against the assemblies in order to compute GC and coverage plots. FRCurve is also employed to rank the assemblies. Finally, a report is generated to allow the researcher to easily compare assemblies and select the best one.

NGI-deNovo needs a global and a project configuration file: the first describes the tools available on the system, while the second describes the sequencing data and the tools/analyses to be run. NGI-deNovo is available at https://github.com/vezzi/de_novo_scilife . Up to now it has been employed in the construction and analysis of more than 100 bacterial and 8 complex genome assembly projects.

Time and place: Thursday, October 23, 13.00, Wallenbergsalen (Main hall)

OP02. Using Piper to analyze thousands of genomes

Johan Dahlberg[1]

1. Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala 75185, Sweden

Presenting author: Johan Dahlberg (johan.dahlberg@medsci.uu.se)

Whole genome sequencing is now feasible in large population-based and clinical cohorts. As part of the Swedish Genomes Program, thousands of whole human genomes will be sequenced at SciLifeLab in the coming years. In order to be able to process the large amount of projects, samples, and data, an automated analysis pipeline is needed.

I will present Piper, an analysis engine that is build on top of GATK Queue and implements a number of common next-generation sequencing analysis workflows, including, but not limited to best practice whole genome data processing. I will also present the challenges faced when moving to data analysis on a massive scale from a core facility perspective. I will explain how this has driven the design process of Piper and where we are at today in terms of implemented workflows and resource usage. Since Piper is an ongoing project, this also a call for feedback on which deliverables would be the most useful for the Swedish research community.

Piper is freely available here: https://github.com/NationalGenomicsInfrastructure/piper

Time and place: Thursday, October 23, 13.20, Wallenbergsalen (Main hall)

OP03. Large-scale analysis of bacterial genomes and plasmids for antibiotic, biocide and metal resistance genes and their co-selection potential

Chandan Pal[1], Johan Bengtsson-Palme[1], Erik Kristiansson[2] and D G Joakim Larsson[1]

1. *Sahlgrenska Academy, University of Gothenburg*
2. *Chalmers University of University*

Presenting author: Chandan Pal (chandan.pal@gu.se)

Background: Antibiotic resistance has become a major public health concern globally. The heavy use of antibiotics is the most important major driver behind this development. In addition, antibiotic resistance can sometimes be selected for by exposure to other chemicals, such as metals and antibacterial biocides. One way this can occur is if biocide/metal resistance genes are physically located within the same cell as the antibiotic resistance genes. If they are located on the very same mobile genetic element, such as on a plasmid, the resistance genes can be transferred together between bacteria via horizontal gene transfer.

Purpose: To identify which biocide/metal resistance genes that often occur together with antibiotic resistance genes on bacterial genomes and plasmids, and to evaluate if there are environments where bacteria with co-selection potential are particularly common.

Methods: We have analyzed 2539 completely sequenced bacterial genomes and 4582 plasmids retrieved from the NCBI refseq database. Similarity searches against a database of biocide/metal resistance genes (BacMet) and a database of mobile antibiotic resistance genes (RESQU) was performed with USEARCH. Metadata of the completed genomes were collected from original literature, PATRIC, and the KEGG database. Based on the source of isolation, the genomes were categorized into 12 different types of environments. Distributions of genomes carrying resistance genes, resistance plasmids and resistance genes were analyzed across different environments. Co-selection potential was evaluated for all bacterial genera and the different environments where those bacteria were isolated from. Gene-networks were built to analyze co-occurrence patterns of biocide/metal and antibiotic resistance genes on plasmids and chromosomes, and visualized using Cytoscape. Heatmaps were generated using R to cluster the biocide/metal resistance genes that occur together with antibiotic resistance genes most frequently.
Results and conclusions: Metal resistance genes are more ubiquitous across all environments than biocide and antibiotic resistance genes, and thus metals appear to have the broadest selection potential (disregarding any mutation-based resistance). Biocide and metal resistance genes are more common on chromosomes, whereas the majority of antibiotic resistance genes are found on plasmids. Thus, the most common co-selection situation between classes of chemicals is mediated via chromosomal metal resistance genes and plasmid borne antibiotic resistance genes. Genes conferring resistance to different metals are quite distinct in terms of which types of antibiotic resistance genes they co-occur with. For example, mercury resistance genes are closely linked to many antibiotic resistance genes, whereas silver, copper and arsenic resistance genes are linked to more distinct subsets of antibiotic resistance genes. Plasmid sizes between 16-128 Kb appeared to have the highest co-selection potential. Isolates from clinical and animal-associated environments seem to have a particularly high potential for co-selection.

Significance: This study provides a framework for assessing co-selection potential between biocides, metal and antibiotics in bacteria. As such, we may be in a better position to understand, and manage some aspects of antibiotic resistance development.

Time and place: Thursday, October 23, 14.00, Wallenbergsalen (Main hall)

OP04. How bioinformatics designed a plasma and urine marker to detect kidney cancer

Authors: Francesco Gatto[1], Intawat Nookaew[1], Marco Maruzzo[2], Anna Roma[2], Nicola Volpi[3], Umberto Basso[2], and Jens Nielsen[1].

Presenting author: Francesco Gatto (gatto@chalmers.se)

1. Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg, Sweden.
2. Medical Oncology Unit 1, Istituto Oncologico Veneto IOV - IRCCS, Padova, Italy.
3. Department of Life Science, University of Modena and Reggio Emilia, Modena, Italy.

The reprogramming of metabolism in cancer consists in deregulated energy metabolism and macromolecule biosynthesis tuned to sustain clone survival and proliferation. We and others have recently shown that unique genetic alterations in clear cell renal cell carcinoma (ccRCC) dictate an outstanding metabolic reprogramming, showing widespread repression of key metabolic pathways like the tricarboxylic acid cycle, glycerophospholipid and nucleotide metabolism and reliance on glycolysis and the pentose phosphate pathway.

Given the clinical potential of such unique changes for diagnosing ccRCC, we used genome-scale metabolic modeling and differential gene expression analysis to further characterize ccRCC metabolic reprogramming. Contrary to other six cancers here evaluated, only in ccRCC we discovered a coordinated regulation of the glycosaminoglycan (GAG) pathway. GAGs promote tumor migration and invasion, hence we speculated that such regulation could be indirectly detected in the plasmatic and/or urinary GAG profile of metastatic ccRCC patients. In line with our hypothesis, we found three GAG markers significantly elevated in metastatic ccRCC compared to control. These markers are uncorrelated with dietary information and their predictive power of the clinical outcome is statistically significant.

These results demonstrate ccRCC coordinately regulates GAG biosynthesis and GAG profiling in accessible fluids can compensate for the current lack of metastatic markers for ccRCC.

Time and place: Thursday, October 23, 14.20, Wallenbergsalen (Main hall)

OP05. Metagenomics reveal changes in antibiotic resistance gene abundance in gut bacterial communities resulting from international traveling

Johan Bengtsson-Palme[1], Martin Angelin[2], Mikael Huss[3], Sanela Kjellqvist[3], Erik Kristiansson[4], Helena Palmgren[2], DG Joakim Larsson[1] and Anders Johansson[2]

1. *Department of Infectious Diseases, University of Gothenburg, Sweden*
2. *Department of Clinical Microbiology, Umeå University, Sweden*
3. *Science for Life Laboratory, Stockholm University, Sweden*
4. *Department of Mathematical Statistics, Chalmers University of Technology, Sweden*

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

Antibiotic resistance is an immense, and rapidly growing, problem. Multi-resistant pathogens, which cannot be treated with available antibiotics, account for thousands of deaths every year. While there is a fair degree of understanding of resistance mechanisms and which genes that are responsible for certain resistance phenotypes, the understanding of how these genes are disseminated globally is limited. Despite travel being identified as a central risk factor in the spread of antibiotic resistance, little work has been done to investigate this transmission route for resistant bacteria. Most studies to date have used culturing or PCR-based approaches to investigate resistance towards specific antibiotics and/or specific bacterial species. Such studies provide a snapshot of the resistome associated with travelling, but the diversity of resistance genes that could be carried across the world while travelling remains unknown. To gain insight into the previously unseen resistance genes, we have analyzed fecal samples from Swedish medical students before and after long-range travelling; either to India or central Africa. None of the study subjects took antibiotics six months prior to or during the study. Total DNA was extracted from the bacterial communities in each fecal sample, and subjected to Illumina sequencing. The sequenced reads were mapped to the Resqu database of resistance genes. We have also investigated the taxonomy of the microbial communities to assess the changes that are related to international travel.
In total, we generated 1.6 terabases of sequence data for 35 individuals. At the group level, there was a slight increase in the relative abundance of resistance genes after travelling. For some categories (such as aminoglycoside, beta-lactam, sulfonamide and trimethoprim resistance genes) increases were significant, largely driven by a small set of resistance genes that have previously been encountered on several different mobile genetic elements. There was, however, a large degree of variation between individuals, and no clear distinction in the resistance genes patterns between travellers returning from India or Africa.

There were no significant changes of genus diversity before and after travel, neither in total, nor for the two geographical regions individually. However, there were indications that *Escherichia* and other enterobacteria were increasing in relative abundance after travel, especially for travellers to Africa, and the levels of Proteobacteria were elevated in the entire cohort.

We conclude that travel can contribute to increasing the relative resistance gene abundance in the gut. The overall bacterial diversity at the genus level, however, did not change, indicating that travel alone, in the absence of antibiotic treatment, is not associated with major microbiome composition changes. On the contrary, our findings indicate that the increased levels of antibiotic resistance genes were associated with a small fraction of gut bacteria. Other factors, such as antibiotics intake or hospital stay at the destination might play a major role in the dissemination process, and requires further attention in future studies.

Time and place: Thursday, October 23, 14.40, Wallenbergsalen (Main hall)

OP06. A kinetically constrained FBA model predicts shifts in Saccharomyces cerevisiae's metabolic strategy

Avlant Nilsson[1]

1. Chalmers University of Technology

Presenting author: Avlant Nilsson (avlant@hotmail.com)

Flux balance analysis (FBA) is a successful mathematical approach for predicting the metabolic activity of a cell. It makes use of the stoichiometry of the biochemical reactions and the rates of nutrient uptake. These relations are used to generate self consistent sets of metabolic fluxes, i.e. rates of metabolic conversion over the reactions. It often agrees well with experimental data to chose the set that has the highest growth rate.

One problem with the standard FBA approach is that it does not constrain the flux levels. In the living cell fluxes are limited by the fact that they are performed by a finite amount of enzymes. The enzyme levels are in turn limited by the metabolites available for enzyme production and a limited
space for enzymes to occupy. It has previously been shown that taking such limits in to account can improve the predictive powers of FBA.

A modified version of FBA has been developed that estimates the weight of the participating enzymes from their catalytic activity and the fluxes. The total enzyme weight can then be constrained to experimentally determined levels.

Applying the method to the central carbon metabolism of *Saccharomyces cerevisiae* accurately predicts the Crabtree effect i.e. the shift in metabolic strategy from respiration to fermento respiration at high growth rates. It also accurately predicts a significantly lower maximum growth rate for growth on galactose than on glucose. Since this can not be done with the standard FBA approach the findings suggests that limits in enzyme kinetics is the underlying cause of the phenotypes.

The modified version of FBA might be of use to metabolic engineers in predicting if a potential pathway might significantly decrease cell fitness. It could also be a useful tool for identifying other phenotypes caused by limits in enzyme kinetics.

Time and place: Thursday, October 23, 16.00, Wallenbergsalen (Main hall)

OP07. A first truly systems level mechanistic model – unravelling the gene regulation of Th2 differentiation

Mattias Köpsén[1], William Lövfors[1], Sören Bruhn[2], Gunnas Cedersund[1], Mikael Benson[1] and Mika Gustafsson[1]

1. *Linköping University*
2. *Karolinska Institute*

Presenting author: Mattias Köpsén (matt.kopsen@gmail.com)

Recent and ongoing revolutions in measurement technologies imply completely new possibilities for genome research: today, time-resolved, quantitative, and systems-level data are available. Nevertheless, without a corresponding revolution in methods for data analysis, these new data tend to drown researchers and doctors, rather than provide clear and useful insights. Such new methods are developed within the field of systems biology. Systems biology has two main approaches: mechanistically detailed and well-determined simulation models for small subsystems, and more approximative statistical models for the entire genome. However, there are few, if any, methods that combine the strengths of these two approaches. Herein, we present LASSIM, a new simulation-based approach, which can be applied to systems of the size of the entire genome. The superior performance of LASSIM is demonstrated in three examples: i) an example with simulated data shows that unlike traditional large-scale methods, LASSIM correctly identifies the true behavior between measured data-points, ii) LASSIM outperforms the winner of a previous DREAM challenge, the most competitive benchmarking approach available, iii) based on new data from TH2 differentiation, LASSIM identifies a first mechanistic model for the entire genome. The key predictions of this model are typically enriched for DNA bindings, which suggests that most predicted interactions are direct. Moreover, *in silico* knockdowns were experimentally validated. In summary, LASSIM opens the door to a new type of model-based data analysis: to models that combine the strengths of reliable mechanistic models with truly systems-level data.

Time and place: Thursday, October 23, 16.20, Wallenbergsalen (Main hall)

OP08. Clustering Metagenomic Contigs using CONCOCT

Johannes Alneberg[1], Brynjar Smári Bjarnasson[2], Luisa W Hugerth[1], John Larsson[3], Jarone Pinhassi[3], Ino de Bruijn[4], Melanie Schirmer[5], Umer Z. Ijaz[5], Joshua Quick[6], Leo Lahti[7], Nick Loman[6], Anders Andersson[1] and Christopher Quince[5]

1. KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Stockholm, Sweden.
2. KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Stockholm, Sweden., Örebro Universitet
3. Linnéuniversitetet, Kalmar
4. Bioinformatics Infrastructure for Life Sciences (BILS), Stockholm, Sweden.
5. School of Engineering, University of Glasgow, Glasgow, UK.
6. Institute of Microbiology and Infection, University of Birmingham, Birmingham, UK.
7. Department of Veterinary Biosciences, University of Helsinki, Helsinki, Finland., Laboratory of Microbiology, Wageningen University, Wageningen, the Netherlands.

Presenting author: Johannes Alneberg (johannes.alneberg@scilifelab.se)

Metagenomic sequencing can successfully produce genome fragments (contigs) from complex microbial communities through assembly, but since the assembly does not reconstruct entire genomes, binning of contigs is necessary. Traditionally, unsupervised metagenomic binning programs have used sequence composition, exploiting k-mer biases between species, and/or coverage differences within a sample.

CONCOCT, Clustering cONtigs on COverage and ComposiTion, combines these two sources of information and extends them by using patterns of coverage distributions over multiple samples instead of a single one. Using Bayesian statistics to estimate the parameters of a Gaussian Mixture model, CONCOCT can automatically cluster contigs into genomes without the need to fix the number of clusters in beforehand. The performance of CONCOCT will be demonstrated on a real metagenomic data set from the Linneaeus Microbial Observatory time series in the Baltic Sea.

Time and place: Friday, October 24, 10.20, Wallenbergsalen (Main hall)

OP09. Correction of long PacBio reads using short Illumina reads

Sylvie VM Tesson[1], Tomas Larsson[1], Magnus Alm Rosenblad[1], Anders Blomberg[1], Anna Godhe[2] and Mats Töpel[1]

1. *University of Gothenburg, Department of Chemistry and Molecular Biology*
2. *University of Gothenburg, Department of Biological and Environmental Sciences*

Presenting author: Sylvie Tesson (sylvie.tesson@gu.se)

The reference genome of *Skeletonema marinoi* (estimated genome size 50Mb), a centric diatom present in Scandinavian waters, is presently sequenced at the University of Gothenburg using two long-read (LR) PacBio libraries and six short-reads (SR) Illumina libraries. The LR PacBio technique provides long DNA fragments (average 6 kb) distributed along the genome, with at least 20x coverage and potential insertion/deletion errors. The SR Illumina technique provides short paired-end reads (150bp) randomly disturbed over the genome with circa 200x coverage. SR Illumina libraries were used to correct sequencing errors in LR PacBio libraries before subsequent assembly. Four correction tools were tested: PacBioToCA, recommended by Pacific Biosciences, and three alternative tools: LSC, BLUE, and ECTools. We performed further analyses of the corrected LR reads to understand how each correction tool utilizes SR Illumina data and their role in providing better-assembled genomes.

Time and place: Friday, October 24, 10.40, Wallenbergsalen (Main hall)

OP10. The influence of glucose uptake on Snf1/Mig1 Signalling dynamics at the single cell level

Niek Welkenhuysen[1], Loubna Bendrioua[1], Riccardo Dainese[2], Mattias Backman[2], Ricardo Silva[3], Mattias Goksör[3], Marija Cvijovic[2] and Stefan Hohmann[1]

1. *CMB, Gothenburg University*
2. *Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg*
3. *Department of Physics, University of Gothenburg*

Presenting author: Niek Welkenhuysen (niek.welkenhuysen@gmail.com)

The uptake and metabolism of glucose in the yeast *Saccharomyces cerevisiae* triggers several signaling pathways and influences major cell processes which makes it a key nutrient for the yeast cell. The glucose repression pathway, with main components SNF1-complex and Mig1, regulates the energy homeostasis in the cell. Upon low concentration of glucose and other preferred carbon sources Snf1 is activated by phosphorylation, translocates to the nucleus and consequently phosphorylates transcription inhibitor Mig1. Phosphorylated Mig1 exits the nucleus and this allows transcriptional activation of genes that enable the cell to use less preferred carbon sources. The behavior and dynamics of this pathway is complex and we apply a multidisciplinary approach with microfluidics, microscopy and mathematical modelling to be able to elucidate the mechanism of the glucose repression pathway.

To study the dynamic properties of glucose signaling we apply single cell time-scale fluorescence microscopy on the Snf1-Mig1 system, whereby we observe shuttling of fluorescence tagged Mig1 in and out of the nucleus. To be able to control the environment of the external glucose concentration we use a three channel microfluidic device where cells are captured and arranged in an array with the help of optical tweezers. The external glucose concentration was shifted up or down and the response of the cell to this switch in glucose concentration was observed. Changes in the localization of Mig1 after upshift or downshift of glucose are rapid and happen within time scale of seconds. By using a gradient of concentrations it is observed that Mig1 relocalization is dependent on a threshold of glucose concentration and cell to cell variability increases near the threshold glucose concentrations. To study how the reduction in glucose uptake influences the threshold for Mig1 relocalization and cell to cell variability, we used yeast cells that where genetically modified in order to only contain either the low affinity hexose transporter Hxt1, the high affinity hexose transporter Hxt7 or the chimeric TM6*. These single transporter strains have a progressively reduced glucose uptake in comparison with the wild type yeast strain. To elucidate the factors that influence the cell to cell variability we have developed both general deterministic and a nonlinear mixed effect model to be able to describe the single cell data and get a measure of cell to cell variability.

In conclusion, we study glucose repression with an approach that brings together several scientific fields. We have been able to show that the Snf1-Mig1 system has a threshold value for glucose concentration changes and that it reacts rapidly on changes in glucose concentration. With the help of mathematical modelling we are in process of studying how the uptake of glucose influences the Snf1-Mig1 system and what causes cell to cell variability.

Time and place: Friday, October 24, 11.00, Wallenbergsalen (Main hall)

OP11. A validated gene regulatory network and GWAS to identify early transcription factors in T-cell associated diseases

Mika Gustafsson[1], Danuta Gawel[1], Sandra Hellberg[2], Aelita Konstantinell[1], Daniel Eklund[2], Jan Ernerudh[3], Antonio Lentini[1], Robert Liljenström[1], Johan Mellergård[4], Hui Wang[1,5], Colm E. Nestor[1], Huan Zhang[1], Mikael Benson[1]

1. The Centre for Individualised Medicine, Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
2. Clinical and Experimental Medicine, Division of Clinical Immunology, Unit of Autoimmunity and Immune Regulation, Linköping University, Sweden
3. Department of Clinical Immunology and Transfusion Medicine and Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden
4. Department of Neurology and Department of Clinical and Experimental Medicine, Linköping University, Sweden
5. Department of Immunology, MD Anderson Cancer Centre, Houston, Texas, USA

Presenting author: Mika Gustafsson (mika.gustafsson@liu.se)

The identification of early regulators of disease is important for understanding disease mechanisms, as well as finding candidates for early diagnosis and treatment. Such regulators are difficult to identify because patients generally present when they are symptomatic, after early disease processes. Here, we present an analytical strategy to systematically identify early regulators by combining gene regulatory networks (GRNs) with GWAS. We hypothesized that early regulators of T-cell associated diseases could be found by defining upstream transcription factors (TFs) in T-cell differentiation. Time-series expression profiling identified upstream TFs of T-cell differentiation into Th1/Th2 subsets enriched for disease associated SNPs identified by GWAS. We constructed a Th1/Th2 GRN based on integration of expression, DNA methylation profiling and sequence-based predictions data using LASSO algorithm. The GRN was validated by ChIP-seq and siRNA knockdowns. GATA3, MAF and MYB were prioritized based on GWAS and the number of GRN predicted targets. The disease relevance was supported by differential expression of the TFs and their targets in profiling data from six T-cell associated diseases. We tested if the three TFs or their splice variants changed early in disease by exon profiling of two relapsing diseases, namely multiple sclerosis and seasonal allergic rhinitis. This showed differential expression of splice variants of the TFs during relapse-free asymptomatic stages. Potential targets of the splice variants were validated based on expression profiling and siRNA knockdowns. Those targets changed during symptomatic stages. Our results show that combining construction of GRNs with GWAS can be used to infer early regulators of disease.

Time and place: Friday, October 24, 11.20, Wallenbergsalen (Main hall)

OP12. Computational elucidation of the gene expression and regulation in insulin resistance based on the Cap Analysis of Gene Expression fat tissue profiling

Olga Hrydziuszko[1], Carsten O. Daub[1], Albin Sandelin[2], Mikael Rydén[3] and Peter Arner[3]

1. *Department of Biosciences and Nutrition, Karolinska Institutet*
2. *The Bioinformatics Centre, Department of Biology & Biotech Research and Innovation Centre, University of Copenhagen*
3. *Unit of Endocrinology with the Lipid laboratory, Department of Medicine, Karolinska Instituet*

Presenting author: Olga Hrydziuszko (olga.hrydziuszko@gmail.com)

Insulin resistance (IR) is a condition in which cells fail to respond to a normal action of insulin, a hormone that plays a key role in regulating glucose metabolism and homeostasis. IR has been found to be a strong predictor of the development of type 2 diabetes yet its exact molecular mechanisms remain unclear. Here, we are using a broad spectrum of bioinformatics techniques based on the Cap Analysis of Gene Expression (CAGE) fat tissue profiling to improve our understanding of gene expression and regulation in IR.

Obese (n=52) and non-obese (n=9) individuals were enrolled in the study and their IR was measured using the gold standard euglycemic clamp. Fat tissue biopsies were taken after an overnight fast before the clamp and following the clamp test. These were used for CAGE profiling, in which short sequences tags originating from the 5' end of RNA transcripts are isolated and sequenced.

Mapping back the sequencing tags to the reference genome provides a powerful method to determine transcription start sites (TSSs) on the genome-wide scale. Over 500 000 CAGE defined TSSs were identified with 139 689 being expressed by at least 5 counts in at least 20% of the samples. Using enhancer's properties to produce bidirectional, exosome-sensitive, un-spliced RNAs, 1199 highly expressed enhancer candidates were computationally predicted. Preliminary data analysis showed a clear insulin response in the fat tissue after 2 hours of euglycemic clamp where TSSs associated with known insulin response genes changed expression in the expected directions, incl. PDK4, CTGF, PFKFB3 and PNPLA3. Motif activity response analysis that combines prediction of the TF binding sites with mathematical modelling of changes in gene expression showed that transcription factors incl. EGR1, TFAP2B and RREB1 were likely driving these changes. Furthermore, we found more than 100 genes that were specifically affected (down- or up-regulated) in the obese individuals and associated with insulin signalling pathway and examples of enhancer candidates likely regulating some of these genes (e.g. PIK3R1)

These preliminary results demonstrate the applicability of CAGE profiling combined with computational approaches to investigate molecular mechanisms of IR at the genome-wide scale in a clinically derived fat tissue samples.

Time and place: Friday, October 24, 11.40, Wallenbergsalen (Main hall)

OP13. Bioinformatic and structural approaches to study cytoskeletal proteins in Mycobacterium bovis BCG and M. marinum M

Nicole Selzer[1], Ala Javadi[1], Linda Sandblad[1] and Christer Larsson[1]

1. Umeå University, Department of Molecular Biology, 90187 Umeå

Presenting author: Nicole Selzer (nicole.selzer@molbiol.umu.se)

M. tuberculosis is a rod shaped bacteria, which causes the human diseases tuberculosis. *Mycobacterium marinum* M (Mm) and *Mycobacterium bovis* BCG (Mb) function as model organisms for *Mycobacterium* and their genomes were recently sequenced. *Streptomyces coelicolor* (Sc) is a filamentous bacterium and is established as model organisms for polar growth. Both genera *Mycobacterium* and *Streptomyces* belong to the order Actinomycetales and we investigate their relationship on proteomic level, focusing on cytoskeletal like proteins involved in growth and polarity.

On the proteomic level the genera *Mycobacterium* and *Streptomyces* show homologies among intermediate filament like proteins. The two coiled coil and filament forming proteins FilP and DivIVa are involved in growth and morphogenesis in hyphal tips of *S. coelicolor*. Sc-DivIVa and Sc-FilP putative homologue protein sequences are retrieved from *M. bovis* BCG and *M. marinum* M proteomes and analyzed. Homologous protein sequences for Sc-DivIVa are established for *M. bovis* BCG and *M. marinum* M. DivIVa is essential in *S. coelicolor* and we would like to test if it is also essential for rod shaped bacteria. However, Sc-FilP has a putative homologue in *M. bovis* BCG, but not in *M. marinum* M. A potential protein sequence for Mm-FilP might have been present, but was reduced as not essential protein.

Mb-FilP has been cloned and expressed in *E. coli*. The protein was successfully purified and form filamentous cytoskeletal structures in vitro. The protein was analyzed by negative staining electron microscopy. First observations of Mb-FilP protein show a filament forming structure similar to higher eukaryotic cytoplasmic intermediate filaments. Sc-FilP has earlier been shown as an intermediate filament like protein with a striated pattern. We would like to study the 3D organization of *Streptomyces* and *Mycobacterium* strains by electron tomography and use antibodies for immuno-localisation on bacterial serial sections and to visualize the protein within the bacteria by electron microscopy. An observation of the spatial localization of DivIVa and FilP within *M. bovis* BCG and *M. marinum* M will give further answers on protein function, cell stability and the mechanisms behind polarity establishment and maintenance among the Actinobacteria. This would also demonstrate putative protein evolution.

Time and place: Friday, October 24, 14.45, Wallenbergsalen (Main hall)

OP14. Systematic search for transcription-modulating somatic mutations in regulatory DNA regions across 505 tumor genomes

Johan Fredriksson[1], Lars Ny[1], Jonas Nilsson[1] and Erik Larsson[1]

1. University of Gothenburg

Presenting author: Johan Fredriksson (johan.fredriksson@gu.se)

Few studies have considered the putative roles of somatic mutations outside of coding sequences for tumor development, a rare exception being the recent uncovering of activating mutations in TERT regulatory DNA. While suggestive of a general mechanism for oncogene activation, this hypothesis remains untested. Here, we map somatic mutations in 505 high-coverage whole tumor genomes, spanning across 14 cancer types, and systematically screen for associations between mutations in regulatory regions and changes in RNA levels as determined by RNA-seq. We identify recurrent promoter mutations in several genes, including DPH3 and PLEKHS1, but find that TERT mutations are exceptional in also showing a strong and significant association with altered expression. Detailed analysis of TERT across cancers shows that the strength of this association is highly variable and strongest in copy-number stable cancers such as low-grade glioma and thyroid carcinoma. We additionally propose that TERT promoter mutations control expression of the nearby CLPTM1L. Our analysis provides a detailed pan-cancer view of TERT transcriptional activation in tumors and, importantly, precludes a frequent role for promoter mutations in oncogene activation beyond TERT.

Time and place: Friday, October 24, 15.05, Wallenbergsalen (Main hall)

OP15. HCSD: Human Cancer Secretome Database

Amir Feizi[1], Amir Banaei-Esfahani[1] and Jens Nielsen[1]

1. Chalmers University of Technology

Presenting author: Amir Feizi (feizi@chalmers.se)

Cancer secretome is a field to explore the factors secreted in tumor microenvironment by stem cells, non-stem cells and the surrounding stroma. Many of these factors recently were linked to the hallmark of cancer that are reliant on cell-cell signaling and cancer secretome is getting to be a critical field since it plays a deterministic role in cancer progression and it has a potential of the research to find a key therapeutic target for many cancer. Therefore, the rate of data generation in the field is growing quickly and querying these data challenges many researchers in the field. Consequently, to boost the research in the field creating a database for cancer secretome data is highly necessary. We developed Human Cancer Secretome Database (HCSD) to fulfill this need. HCSD contains >150000 measurements generated in the field through >20 qualitative and quantitative studies on 15 cancer type. It has a simple and user friendly query system based on gene names, data type, and cancer type as three main query fields. The results are visualized in a highly interacting way beside all the annotation, cross reference, secretory features for each gene. To our belief, developing HCSD can be a very important bioinformatic solution to boost the research on cancer secretome and tumor microenvironment.

Time and place: Friday, October 24, 15.25, Wallenbergsalen (Main hall)

Poster Presentations

P01. **Application of random graphs in neural modelling**
Presenting author: Fioralba Ajazi (fioralba@maths.lth.se)

P02. **Structural basis for copy-number alterations in cancer**
Presenting author: Babak Alaeimahabadi (babak.alaeimahabadi@gu.se)

P03. **A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences**
Presenting author: Fredrik Boulund (fredrik.boulund@chalmers.se)

P04. **Tentacle: distributed quantification of genes in metagenomes**
Presenting author: Fredrik Boulund (fredrik.boulund@chalmers.se)

P05. **Pan-cancer expression profiling of non-coding RNAs**
Presenting author: Niklas Dahr (niklas.dahr@gu.se)

P06. **An Automated RNA-Seq Analysis Pipeline at NGI-Stockholm**
Presenting author: Philip Ewels (phil.ewels@scilifelab.se)

P07. **A Fully Automated Infrastructure for NGS Analysis**
Presenting author: Mario Giovacchini (mario.giovacchini@scilifelab.se)

P08. **Elucidating co-selection of genes conferring resistance to biocides and metals with antibiotic resistance genes in Swedish wastewater treatment plants**
Presenting author: Rickard Hammarén (gusrickaha@student.gu.se)

P09. **Predicting antibiotic resistance profile in five carbapenem-resistant isolates using whole genome sequencing**
Presenting author: Anna Johnning (anna.johnning@chalmers.se)

P10. **A Hierarchical Bayesian Model for Ranking of Genes in Metagenomics Based on Differential Abundance**
Presenting author: Viktor Jonsson (viktor.jonsson@chalmers.se)

P11. **Evaluation of Statistical Methods for Comparative Metagenomics**
Presenting author: Viktor Jonsson (viktor.jonsson@chalmers.se)

P12. **Genome wide identification of active enhancers in a Illumina CAGE data set**
Presenting author: Wenjing Kang (wenjingkh@gmail.com)

P13. **FocalScan: Scanning for altered genes in cancer based on coordinated DNA and RNA change**
Presenting author: Joakim Karlsson ((joakim.karlsson@gu.se)

Poster Presentations

P14. **Evaluation of Candidate Genes for Prolificacy in Cameroon Native Goats**
Presenting author: Patrick Kenfack Wouobeng (pat3k.wouobeng@gmail.com)

P15. **Identifying microRNAs from small RNA sequencing data in the absence of a reference genome**
Presenting author: Jacqueline Nowak (jacqueline.nowak@ki.se)

P16. **A method to identify epigenetic signatures using Nimblegen arrays**
Presenting author: Balaji Rajashekar (balaji@ut.ee)

P17. **Homology modeling and docking study of 3 oxoacyl (acyl carrier protein) synthase II protein of Neisseria meningitides**
Presenting author: Vaibhav Sabale (vaibhavsabalebi@gmail.com)

P18. **Dissecting the transcriptome through strand specific RNA sequencing**
Presenting author: Benjamín Sigurgeirrson (benjamin.sigurgeirsson@scilifelab.se)

P19. **Role of non coding RNAs in stress response in budding yeast**
Presenting author: Agata Smialowska (agata@chalmers.se)

P20. **Identification of Novel Therapeutic Targets Against Methicillin Resistant Staphylococcus aureus using Computational Biology Methods**
Presenting author: Reaz Uddin (mriazuddin@iccs.edu)

P21. **Kiwi: a tool for integration and visualization of network topology and gene set analysis**
Presenting author: Leif Väremo (varemo@chalmers.se)

P22. **Computational protein modelling based on limited NMR data**
Presenting author: Maryana Wånggren (maryana@chalmers.se)

P23. **Complementing tissue characterisation by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium**
Presenting author: Nancy Yu (nancy.yu@scilifelab.se)

P24. **Estimating copy number alteration in neuroblastoma: comparison of exome sequencing data and SNP microarrays**
Presenting author: Malin Östensson (malin.ostensson@gu.se)

P25. **HierBin - A method for detailed functional annotation and quantification in metagenomes**
Presenting author: Tobias Österlund (tobiaso@chalmers.se)

P01. Application of random graphs in neural modelling

Tatyana Turova[1] and Fioralba Ajazi[1]

1. Lund University

Presenting author: Fioralba Ajazi (fioralba@maths.lth.se)

Random graphs are an important tool used to model structure and dynamics of real networks, in particular, neural networks. In general a network is a collection of objects connected to each other in some fashion. In neural network the nodes represent neurons, and the edges model the dendrites and axons which receive and transmit impulses. Questions related to the evolution of neural networks have received a lot of attention in the literature. Questions on how the structure of neural networks affects the functioning of a network are studied both from the theoretical and from the computational points of view.

The computational tools and the theoretical results can give to experimental analysis the correct help to approach the questions regarding the activation function of neurons, in particular they can give some fundamental conclusions on the evolution in time of one of the most complex network that we can study in nature. The models based on random graphs theory studied by Turova, Villa, Cabessa, Van der Hofstad, Kozma, deal with such questions.

We introduce a model for an inhomogeneous random graph, where the probability of edges also depends on the distance between vertices. We investigate the degree distribution. We find for which parameters of the model the degree of a vertex converges in distribution as the size of a graph goes to infinity and we find the limiting distribution in some special cases.

Simplification allows us to analyze rigorously the model, and study the role of all involved parameters. Having this basic analysis, in the future work we shall increase the complexity of the model, as e.g., consider higher dimensions and more parameters. One particularly interesting question is the evolution in time of the probabilities of the connections.

P02. Structural basis for copy-number alterations in cancer

Babak Alaeimahabadi[1] and Erik Larsson[1]

1. Gothenburg University

Presenting author: Babak Alaeimahabadi (babak.alaeimahabadi@gu.se)

Tumor genomes are mosaics of chromosomal copy-number alterations (amplifications and deletions), some of which some are under selection to activate or inactivate specific oncogenes or tumor suppressors. However, the structural basis of these alterations is poorly understood.

Whole-genome sequencing (WGS) presents an opportunity to better characterize copy-number changes at the structural level categorized as deletions, tandem duplications, insertions, inversions, translocations and even more complex rearrangements such as a deletion with an inversion or insertion at the breakpoint. We have tried to do this by analyzing hundreds of tumor genomes made available in the public domain by The Cancer Genome Atlas (TCGA) consortium.

We have evaluated softwares for mapping structural alterations in tumor genomes based on WGS data using a benchmarking method and have settled on Meerkat. We applied this tool on deep coverage WGS data for around 500 tumor/normal pairs. Next, we are going to provide a detailed overview of structural basis of copy number alteration in cancer.

P03. A novel method to discover fluoroquinolone antibiotic resistance (qnr) genes in fragmented nucleotide sequences

Fredrik Boulund[1], Anna Johnning[1], Mariana Buongermino Pereira[1], D.G. Joakim Larsson[2] and Erik Kristiansson[1]

1. *Department of Mathematical Statistics, Chalmers University of Technology and University of Gothenburg*
2. *Department of Infectious Diseases, Institute of Biomedicine, University of Gothenburg*

Presenting author: Fredrik Boulund (fredrik.boulund@chalmers.se)

Broad-spectrum fluoroquinolone antibiotics are central in modern health care and are used to treat and prevent a wide range of bacterial infections. The recently discovered *qnr* genes provide a mechanism of resistance with the potential to rapidly spread between bacteria using horizontal gene transfer. As for many antibiotic resistance genes present in pathogens today, *qnr* genes are hypothesized to originate from environmental bacteria. The vast amount of data generated by shotgun metagenomics can therefore be used to explore the diversity of *qnr* genes in more detail.

In this paper we describe a new method to identify *qnr* genes in nucleotide sequence data. We show, using cross-validation, that the method has a high statistical power of correctly classifying sequences from novel classes of *qnr* genes, even for fragments as short as 100 nucleotides. Based on sequences from public repositories, the method was able to identify all previously reported plasmid-mediated *qnr* genes. In addition, several fragments from novel putative qnr genes were identified in metagenomes. The method was also able to annotate 39 chromosomal variants of which 11 have previously not been reported in literature.

The method described in this paper significantly improves the sensitivity and specificity of identification and annotation of *qnr* genes in nucleotide sequence data. The predicted novel putative *qnr* genes in the metagenomic data support the hypothesis of a large and uncharacterized diversity within this family of resistance genes in environmental bacterial communities. An implementation of the method is freely available at http://bioinformatics.math.chalmers.se/qnr/.

P04. Tentacle: distributed quantification of genes in metagenomes

Fredrik Boulund[1], Anders Sjögren[1] and Erik Kristiansson[1]

1. Department of Mathematical Statistics, Chalmers University of Technology and University of Gothenburg

Presenting author: Fredrik Boulund (fredrik.boulund@chalmers.se)

In metagenomics, microbial communities are sequenced at increasingly high resolutions generating data sets with billions of DNA fragments. Novel methods that efficiently can process the growing volumes of sequence data are necessary for accurate analysis and interpretation of existing and upcoming metagenomes.

Here we present Tentacle, a novel framework for distributed gene quantification in metagenomes. Tentacle is distributed and implemented using a master-worker approach where DNA fragments are streamed via a network and processed in parallel on worker nodes. Evaluations show that Tentacle scales very well with increasing computing resources. Tentacle is both modular and extensible and can easily be adapted to different applications in metagenomics and integrated into existing workflows. Tentacle is written for Linux in Python 2.7 and is published as open source under the GNU General Public License (v3). Documentation, tutorials, installation instructions, and source code is freely available online at: http://bioinformatics.math.chalmers.se/tentacle.

P05. Pan-cancer expression profiling of non-coding RNAs

Niklas Dahr[1] and Erik Larsson[1]

1. Biomedicine, Sahlgrenska

Presenting author: Niklas Dahr (niklas.dahr@gu.se)

The large, understudied class of long RNA-transcripts (lncRNA) that do not code for proteins have traditionally been considered to be 'junk' or transcriptional noise. Interest into these have been growing lately as more and more of them have been implicated in epigenetic processes (eg. XIST) and regulation of protein coding genes. Their functions appear to be heterogenous but it is clear that at least a subset of them are functional and vital. Utilizing a large dataset publicly available from The Cancer Genome Atlas (TCGA), we are investigating the expression levels of lncRNAs across 24 different cancer-types with the aim of finding novel and functional lncRNA-genes.

P06. An Automated RNA-Seq Analysis Pipeline at NGI-Stockholm

Philip Ewels[1]

1. Science for Life Laboratory, NGI Stockholm

Presenting author: Philip Ewels (phil.ewels@scilifelab.se)

In recent years, next-generation RNA-Sequencing has become the standard for transcriptome analysis. Accordingly, the NGI-Stockholm sequencing node processes a great deal of RNA-Seq samples for users (nearly 200 projects and 4000 samples since 2012) and offers a Best Practice bioinformatics analysis for users. The development of the field has not slowed in this time, and we are keen to keep the analysis that we offer up to date and relevant. To address this we are launching a focussed update of our RNA-Sequencing analysis pipeline.

The key aims of this work fall into three points: refinement of the pipeline into specialised routines tailored to different types of RNA-Sequencing libraries (eg. mRNA-Seq and smRNA-Seq); the addition of further analysis tools and restructing of the pipeline core to give greater transparency.

Here, we describe the steps taken in the planning and restructuring of the pipeline. This process may serve as a model for other groups intending to develop similar routines. We describe which tools we intend to use, and why. Finally, we explore future applications and pipeline scalability. We welcome any feedback and suggestions.

P07. A Fully Automated Infrastructure for NGS Analysis

Mario Giovacchini[1]

1. Science for Life Laboratory, NGI Stockholm

Presenting author: Mario Giovacchini (mario.giovacchini@scilifelab.se)

As Next-Generation Sequencing continues to revolutionize scientific fields from personalized medicine to population-scale genetics, one of the largest challenges facing researchers and core sequencing centers has become how to analyze the enormous quantities of data being produced: Illumina's newest sequencing platform produces nearly two terabases of information per run which must be organized and run through various computationally-intensive steps. The practical challenges of organizing and tracking these quantities of information are formidable, and bioinformaticians are often forced to spend as much or more time wrangling and processing their data as they spend interpreting results.

At the National Genomics Infrastructure in Sweden we have developed a cohesive set of software tools that automatically organize and process sequencing data as they come off the sequencers, producing meaningful analysis results without any human intervention. These tools are designed to work in a modular way, allowing bioinformaticians to direct the data through the bioinformatic pipelines of their choice. Selected result metrics can be tracked using the supplied database tools and the status of every process is made available via a REST-capable API and accompanying web frontend. Automatic report generation, data delivery, and a robust logging/notification system allow bioinformaticians to become involved only when required, allowing them to spend less time shuffling files and more time answering biological questions.

The NGI pipeline is currently used and developed at the Science for Life Laboratory in Stockholm and Uppsala, Sweden. As a fully public and open-source project, we welcome both additional users and developers who are interested in automating their analysis workflows.

P08. Elucidating co-selection of genes conferring resistance to biocides and metals with antibiotic resistance genes in Swedish wastewater treatment plants

Rickard Hammarén[1], Johan Bengtsson-Palme[1], Chandan Pal[1], Carl-Fredrik Flach[1], Erik Kristiansson[2] and DG Joakim Larsson[1]

1. *Department of Infectious Diseases, Sahlgrenska Academy, University of Gothenburg*
2. *Department of Mathematical Sciences, Chalmers University of Technology*

Presenting author: Rickard Hammarén (gusrickaha@student.gu.se)

Antibiotics constitute our core line of defense against bacterial infections and are also fundamental for our ability to perform advanced surgery and treat cancer. However, in the latest years, an antibiotic resistance crisis has emerged, causing increased mortalities and substantial costs. There is increasing evidence that the resistance we see in pathogens did not initially appear in the clinical setting, but that environmental bacteria have contributed to the resistance gene pool shared among pathogens today. Even if much evidence points towards an environmental origin for most mobile resistance determinants, the picture of their dissemination routes from the environment into human pathogens is much less clear. Wastewater treatment plants have been suggested as a possible hotspot which provides an opportunity for environmental bacteria to meet human pathogens and transfer genes between them. In this environment, bacteria are under a constant selection pressure from different chemicals that come with wastes from households, industrial and agricultural sources, and end up in wastewater treatment plants, including antibiotics, metals and biocides. It is therefore reasonable to assume that the risk for co-selection of resistance genes towards antibiotics, metals and/or biocides is high in the wastewater treatment process, particularly as those resistance genes are often situated on the same mobile genetic elements.

In this project, we are investigating the abundance and diversity of resistance genes related to antibiotics, biocides and metals in sludge and water from different steps of the treatment process in three Swedish wastewater treatment plants. The DNA from the samples has been sequenced using Illumina shotgun metagenomic sequencing.

In addition to analyzing at the frequencies of resistance genes, we are also studying their genetic context, by assembling the short sequencing reads from each sample type into longer contigs. In this way, we aim to characterize e.g. plasmids carrying resistance genes, to see whether different resistance genes are physically connected on the same piece of DNA, which would substantially increase their potential for co-selection.

P09. Predicting antibiotic resistance profile in five carbapenem-resistant isolates using whole genome sequencing

Anna Johnning[1], Nahid Karami[2], Erika Tång Hallbäck[2], Ingegerd Adlerberth[2], Erik Kristiansson[1]

1. *Department of Mathematical Sciences, Chalmers University of Technology, Göteborg*
2. *Institute of Biomedicine, Department of Infectious Diseases, University of Gothenburg, Göteborg*

Presenting author: Anna Johnning (anna.johnning@chalmers.se)

Antibiotic resistance is a global threat to human health and its prevalence is increasing and spreading over the world. The traditional method of determining the susceptibility of a bacterial infection is to grow the isolates in the presents of different types of antibiotics. This is usually done in a sequential manner – starting off by testing a smaller set of antibiotics and then adding more when the isolate proves multi-resistance – making full antibiotic resistance profiling of a multi-resistant infection a complex process. Whole genome sequencing could, therefore, offer a complete typing of multi-resistant isolates which would benefit the health of the patient. In this project we have sequence the genomes of five carbapenemase positive isolates of different species taken from different patients at the Sahlgrenska hospital in Gothenburg. All isolates showed resistance to a wide range of antibiotic classes, and was sequenced using Illumina MiSeq. Draft assemblies of each genome was done *de novo* using SPAdes, and annotated for mobile antibiotic resistance genes and chromosomal mutations in antibiotic target genes. From the acquired resistance mechanisms and the intrinsic resistance profile of each species, we could predict a susceptibility profile. The preliminary results presented here show a good cohesion between the predicted phenotype and the determined antibiotic resistance pattern. We conclude that whole genome sequencing could be used as a tool for antibiotic resistance profiling.

P10. A Hierarchical Bayesian Model for Ranking of Genes in Metagenomics Based on Differential Abundance

Viktor Jonsson[1], Olle Nerman[1] and Erik Kristiansson[1]

1. Mathematical Sciences, University of Gothenburg and Chalmers University ofTechnology, Sweden

Presenting author: Viktor Jonsson (viktor.jonsson@chalmers.se)

Metagenomics is a growing research field within ecology and medicine where entire communities of microbes are studied on the genome level. The majority of bacteria in the environment are unculturable and therefore difficult to study individually. Metagenomics does not rely on cultivation and is suitable for analysis of bacteria in their natural communities. The aim is to gain insight into such communities by observing differences in the abundance of genes between environmental conditions.

The statistical challenge lies in finding the genes which have a large enough difference between conditions to be statistically significant. However, the discrete and overdispersed nature of the data makes analysis harder and methods based on normality assumptions become suboptimal. In addition, the number of samples is often small while the number of genes present in a bacterial community is vast. This creates the problem of finding a few truly differentially abundant genes in a sea of noise.

We present a novel statistical model for inference in metagenomics. It is based on a generalized linear model with a canonical log-link but we extend this to include robust moderation of gene-specific variances. The moderation is achieved by putting a hierarchical structure on the variances and assuming that the gene-specific variability is sampled from a global distribution. The model is implemented in a Bayesian framework and the analyses rely on Markov Chain Monte Carlo (MCMC) sampling. Model performance has been evaluated on both simulated and real datasets. The evaluation shows that it has better performance than standard methods, including ordinary generalized linear models, t-tests and variance stabilizing transforms. We conclude that hierarchical Bayesian modeling can substantially increase the power of statistical inference in metagenomics.

P11. Evaluation of Statistical Methods for Comparative Metagenomics

Viktor Jonsson[1], Olle Nerman[1] and Erik Kristiansson[1]

1. Mathematical Sciences, University of Gothenburg and Chalmers University of Technology, Sweden

Presenting author: Viktor Jonsson (viktor.jonsson@chalmers.se)

Metagenomics is the study of microbial communities in their natural state. In comparative metagenomics samples are compared to identify differentially abundant genes and other features between conditions. The data considered is high-dimensional, discrete and highly overdispersed. Several statistical methods, ranging from Fisher's exact test and t-tests to generalized linear models and bootstrapping, have been proposed. However, no comprehensive evaluation has been performed to see which best suits metagenomic data. We present preliminary results from a comparison of several commonly used methods for statistical analysis of metagenomic data. We evaluate the performance of the methods with respect to number of samples, gene abundances, and effect sizes. In order to conserve the variance structure of real metagenomic data the test-data is created by resampling large metagenomic datasets. This study will provide guidance on the choice of statistical methods for future metagenomic studies.

P12. Genome wide identification of active enhancers in a Illumina CAGE data set

Wenjing Kang[1]

1. Karolinska Institutet

Presenting author: Wenjing Kang (wenjingkh@gmail.com)

Cap Analysis of Gene Expression (CAGE) has been recently recognized as a powerful tool to identify genomic locations of active enhancers in a genome wide scale. The behind idea is based on the discovery that active enhancers produce bidirectional, exosome-sensitive and short un-spliced capped RNAs which can be captured and measured by CAGE. The available CAGE-oriented method for active enhancer prediction and the atlas of enhancer loci identified by the method are both derived from the CAGE libraries (within FANTOM5 consortium) sequenced by HeliScope platform. However their applicability to identify active enhancers in the CAGE libraries sequenced by a more commonly used Illumina platform has not yet been reported. Here we aim at predicting the genomic locations of the active enhancers transcribed in the white adipose tissue (WAT) based on the Illumina CAGE data using two methods termed Enhancer Prediction (EP) and Enhancer Intersection (EI). Both methods focus on identifying the bidirectional and bimodal CAGE signal (enhancer-like pattern), which is a signature feature of active enhancers. In particular, EP is a *de novo* computational prediction method characterized with whole genome-wide detection of the loci with enhancer-like pattern; EI is based on the atlas of already identified enhancer candidates (within FANTOM5 consortium).

As a result, we found 47,350 and 38,554 active enhancer candidate loci for EP and EI methods respectively. In these loci, 4 types of CAGE signals were detected: the enhancer-like signal and other 3 noise signals, e.g. forward or reverse unidirectional CAGE signal and bidirectional noise signal. To optimize the specificity of both methods, we particularly designed and simultaneously used two filters that are specific for these noise patterns. Following the elimination of the noise loci (false positives), 5,976 uniquely identified enhancer candidates were obtained in WAT via EI and EP methods and further assessed in the biological context of obesity in WAT. Our results demonstrated that active enhancer candidates can be identified in Illumina CAGE data using the CAGE-oriented methods and enhancer loci source within FANTOM5 consortium. In order not to lose any biological information, a superset of enhancer candidates identified by either method is highly recommended.

P13. FocalScan: Scanning for altered genes in cancer based on coordinated DNA and RNA change

Joakim Karlsson[1] and Erik Larsson[1]

1. Sahlgrenska akademin

Presenting author: Joakim Karlsson (joakim.karlsson@gu.se)

Accumulation of gene copy number changes is one of the driving mechanisms of cancer growth. FocalScan identifies genomic regions where many tumors show simultaneous increases in DNA copy-number amplitude (CNA) and RNA expression (or conversely for DNA deletions). The FocalScan score is based on the dot product of log2 CNA and RNA changes. This puts equal weight to the two variables, but requires coordinated changes in both to achieve a positive score. FocalScan can be used as a complement to e.g. GISTIC (which attempts to identify driver genes only with respect to CNA) in regions of interest, or alone to find regions of focal CNA alteration with consistent change in RNA.

The score follows intuition: If a given genomic region displays both elevated CNA and RNA levels in only some tumors, this would yield a medium score. If elevated levels of both are detected in many tumors, a high score would be assigned. Many tumors showing increased CNA levels and highly increased RNA expression would result in a very high score. If RNA expression is unchanged across tumors, however, this would result in a low score even if CNA levels are elevated.

In addition to the ”basic” score described above, FocalScan also calculates a ”high pass filtered” score. The latter effectively subtracts large (> 10 Mbp) segments from the score, such as chromosome arm-level events, leaving only focal/small alterations.

The algorithm is “non-gene centric”: The genome is scanned at high (500 nt) resolution by dividing chromosomes into small (1000 nt) overlapping tiles. RNA-seq data is used to quantify transcription in each tile. FocalScan can therefore pinpoint e.g. novel non-coding RNAs that are absent in gene annotations.

FocalScan may be most useful for studying amplifications, but regions with coordinated CNA and RNA reduction will also score favorably. Negative scores mean that DNA and RNA are anti-correlated (not common).

P14. Evaluation of Candidate Genes for Prolificacy in Cameroon Native Goats

Patrick Kenfack Wouobeng[1], Kouam Simo Jaures[1], Felix Meutchieye[1], Manjeli Yacouba[1] and Morris Agaba[2]

1. *Department of Animal Science, FASA-University of Dschang*
2. *Biosciences in eastern and Central Africa ILRI Hub, Kenya*

Presenting author: Patrick Kenfack Wouobeng (pat3k.wouobeng@gmail.com)

The aim of the study was to understand genetic factors which may affect the litter size variability in native goats and eventually to develop a test kit for females with high potential of kidding trait. The bone morphogenetic protein receptor 1B (BMPR-1B) gene and growth differentiation factor-9 (GDF-9) were studied as candidate genes for the prolificacy of goats. According to mRNA sequence of ovine BMPR-1B and GDF-9 genes, ten pairs of primers and 1 pair of primers were designed respectively to detect single nucleotide polymorphisms (SNPs) of exon 1 to exon 10 of the BMPR-1B gene and exon 2 of the GDF-9 gene in both high prolificacy goat (12 animals) and low prolificacy goat (12 animals) by polymerase chain reaction (PCR)-single strand conformation polymorphism (SSCP) method. So far only the products amplified by primers P1 of the unique exon of GDF-9 and primers P2 (exon 2), P3 (exon 3), P10 (exon 10) of BMPR-1B displayed polymorphisms. For primer P1, four (4) genotypes were detected. These results preliminarily showed that 2 of the detected loci of the GDF-9 gene had a significant effect in the amino acid products in both group of animals. For primers P2, P3 and P10, a total of ten (10) mutations were detected. But none of them causes a change in the amino acid products. Further laboratory and statistical analysis are still going on to process the results and conclude.

P15. Identifying microRNAs from small RNA sequencing data in the absence of a reference genome

Jacqueline Nowak[1,2]

1. *Karolinska Institutet*
2. *University of Postdam*

Presenting author: Jacqueline Nowak (jacqueline.nowak@ki.se)

MicroRNAs are small, single-stranded RNAs with an average length of 22-24 nucleotides. Pairing with their target mRNA, they control the cleavage of mRNAs and are therefore an important part in post-transcriptional processing. To identify microRNAs from small sequencing data, the standard approach is to map the data against the genome sequence. Since this is not applicable to species without any reference genome, this work concentrates on methods in identifying microRNAs without the need of the genome. The model that was used here was *Notophthalmus viridescens*, or red-spotted newt, known for its ability to regenerate limbs, brain and lenses. The genome of this organism has not yet been sequenced.

To identify microRNAs, the sequencing data of *Notophthalmus viridescens* was sorted into piles, each pile representing sequences that are similar to each other and assemble to a contig. From these piles, profile HMMs were computed and firstly mapped to the miRBase mature and hairpin databases to identify known microRNAs found in other organisms. In addition, the profile HMMs generated from the small RNA piles were used to find reverse complement pairs. Identification of such pairs increases the likelihood for these sequences to be genuine microRNAs. This would allow us to identify novel and perhaps newt-specific microRNAs that are not available in any sequence databases.

By applying the HMM approach to *N. viridescens*, we found several microRNAs that were found with this approach but not identified when using direct alignment of the data to the microRNA databases. For the measurement of similarities between profile HMMs, we used small RNA sequencing data of *Xenopus tropicalis*, or Western clawed frog, a model organism with a sequenced reference genome as well as well-annotated microRNAs, as a benchmarking control and indicator for accuracy and reliability of this method. For both the frog and the newt datasets we were able to identify complementary microRNA pairs, giving a first indication for novel microRNAs.

The methodology developed from this work can be used for studies in other species without a reference genome, giving a first insight into which microRNAs can be found. Furthermore, our results indicate that using profile HMMs to identify complementary sequence pairs in small RNA sequence datasets has the potential to be a powerful approach in identifying novel microRNAs without the need for a reference genome.

P16. A method to identify epigenetic signatures using Nimblegen arrays

Balaji Rajashekar[1]

1. University of Tartu

Presenting author: Balaji Rajashekar (balaji@ut.ee)

We developed a method to identify epigenetic signatures using DNA methylation, gene expression and Histone marks data from Nimblegen arrays. The mouse skin cancer dataset was used in our method. The analyzed data contained three technically replicated measurements from gene expression, DNA methylation and four different histone marks. The main goal was to identify genetic marks that show cancer progression in the three stages.

All the experimental data were analyzed in our analysis pipeline. The data were VSN normalized in Ringo package. We selected probes found in the 2000bp downstream and 500bp upstream around transcription start site for DNA methylation and Histone marks data. The expression was estimated for each gene. We identified differential significant methylated transcripts by comparing counts of probes with M-value>1 within each transcript between the stages of disease. Comparison between the stages was done using the one-sided Fisher test, using 0.00005 as the p-value threshold. We extracted the set of differentially expressed genes that contained at least one transcript with significantly increased or decreased methylation level during the cancer progression (we call them driver differentially methylated genes) and checked, whether those genes contain transcripts with significantly increased or decreased levels of histone modification.

We identified 19 and 38 significant driver differentially methylated genes from the first two and the last two stages respectively. These significant genes were correlated with histone marks allowed us to identify patterns in histones marks H3K4me3 and H3K9ac which act as repressors. In few cases the histone mark H3K27me3 acted as enhancer.

P17. Homology modeling and docking study of 3 oxoacyl (acyl carrier protein) synthase II protein of Neisseria meningitidis

Vaibhav Sabale[1] and Arun Ingale[1]

1. Department of Biotechnology, School of Life Sciences, North Maharashtra University, P O Box 80 Umavi Nagar, Jalgaon 425001 (MS), India

Presenting author: Vaibhav Sabale (vaibhavsabalebi@gmail.com)

Meningitis is inflammation of the protective membrane covering the brain and spinal cord known collectively as the meninges. The bacterial meningitis disease has repeatedly caused outbreak worldwide. *Neisseria meningitidis* (NM) is major causative agent of bacterial meningitis. The "3 oxoacyl (acyl carrier protein) synthase II" (Beta-Ketoacyl Acp synthase II) enzyme which involved in fatty acid biosynthesis of *Neisseria meningitidis*. This enzyme in fatty acid synthesis and target for discovery of novel antibacterial agent. In this study *in silico* analysis was done by using bioinformatics tools and software. Modeller (V 9.12) use for the 3D structure modeling of 3 oxoacyl (acyl carrier protein) synthase II using wild type *E. coli* fabF (KASII) (PDB ID- 2GFW) protein 3D structure. The quality and validation of model obtained was performed using the structural Analysis and Verification Server (SAVES). Drugbank database used for the ligand selection. Hex online server and Argus Lab software use for the DOCKING of Protein 3D structure and ligand compound. The docking study data showed good fit Root Mean Square Difference (RMSD). This study aims to understand functional aspect to development of novel drug against Meningitis using *in silico* approach.

P18. Dissecting the transcriptome through strand specific RNA sequencing

Benjamín Sigurgeirsson[1], Olof Emanuelsson[1] and Joakim Lundeberg[1]

1. *Royal Institute of Technology (KTH)*

Presenting author: Benjamín Sigurgeirrson (benjamin.sigurgeirsson@scilifelab.se)

Strand specific RNA sequencing is rapidly replacing conventional cDNA sequencing as an approach for assessing information about the transcriptome. Alongside improved laboratory protocols the development of bioinformatical tools is steadily progressing. In the current procedure the Illumina TruSeq library preparation kit is used, along with additional reagents, to make strand specific libraries in an automated fashion which are then sequenced on Illumina HiSeq 2000. By the use of freely available bioinformatical tools we show, through quality metrics, that the protocol is robust and reproducible. We further highlight the practicality of strand specific libraries by comparing expression of strand specific libraries to non-stranded libraries, by looking at known antisense transcription of pseudogenes and by identifying novel transcription. Additionally, two sequence aligners, Tophat2 and STAR, are compared through the metrics of mapping speed and mapping yield.

P19. Role of non coding RNAs in stress response in budding yeast

Agata Smialowska[1], Petri-Jaan Lahtvee[1] and Jens B. Nielsen[1]

1. *Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg*

Presenting author: Agata Smialowska (agata@chalmers.se)

The recent observation that most DNA in eukaryotic genomes is transcribed has sparked a debate on novel aspects of RNA biology and has transformed our understanding of gene, chromosome and genome regulation. The development of parallel sequencing technologies has allowed unbiased investigation of transcriptomes, which has led to discovery of plethora of unexpected RNAs that have no protein coding capabilities, present in all organisms studied so far. Non-coding RNA (ncRNA) genes produce functional RNA molecules, which rather than encoding protins, interact with other RNA molecules, transcription machinery or chromatin to modulate gene expression. They are involved in numerous cellular processes such as gene regulation, chromatin domain formation and genome stability.

NcRNAs seem to be particularly important for roles that require highly specific nucleic acid recognition without complex catalysis, such as in directing co- and post-transcriptional regulation of gene expression or in guiding RNA modifications. Recent studies indicate that ncRNAs are important players in modulating responses to changing environmental conditions. In this study, we present evidence that ncRNAs are invlolved in regulation of the stress response in budding yeast. We analysed transcriptomes of wild type *Saccharomyces cerevisiae* subject to temperature, osmotic and cell membrane stress. We show that ncRNAs are enriched in loci involved in stress regulation, and we discuss the implications of this finding.

P20. Identification of Novel Therapeutic Targets Against Methicillin Resistant Staphylococcus aureus using Computational Biology Methods

Reaz Uddin[1]

1. Dr. Panjwani Center for Molecular Medicine and Drug Research, International Center for Chemical and Biological Sciences, University of Karachi

Presenting author: Reaz Uddin (mriazuddin@iccs.edu)

Several available antibiotics are not much effective because of the acquired resistance developed by multidrug resistant bacteria (MDR). Methicillin Resistant *S. aureus* (MRSA) is one of the MDR pathogens notorious for its widespread infection around the world. The high resistance showed by MRSA reflects a serious concern and efforts should be made for the discovery of better therapeutics. In order to achieve this objective, we have performed an *in silico* comparative metabolic pathways analysis of the host (i.e. *Homo sapiens*) and the pathogen (i.e. MRSA). We have identified 22 non-homologous metabolic pathways unique to the pathogen MRSA (i.e. absent in human host). Consequently, enzymes from unique metabolic pathways were retrieved via KEGG metabolic pathway database and compared with the enzymes from the host by a manual comparison between them.

As a result, we have shortlisted few unique and essential enzymes that could be drug targets against MRSA. Those identified drug targets could be useful for an effective drug discovery phase. We also have searched the PDB database using the unique and essential enzymes from MRSA as queries. We shortlisted at least five enzymes for which there was no corresponding deposition in PDB reflecting that their crystal structures were not solved yet! We chose one enzyme (i.e. Glutamate synthase) out of five so that its structure could be solved by Homology modeling. The modeled structure was well validated by multiple tools. It is concluded that the comparative *in silico* analyses of multidrug resistant strain MRSA provides an effective approach for the identification of novel antibiotic targets.

P21. Kiwi: a tool for integration and visualization of network topology and gene set analysis

Leif Väremo[1], Francesco Gatto[1] and Jens Nielsen[1]

1. Systems biology, Department of Chemical and Biological Engineering, Chalmers University of Technology

Presenting author: Leif Väremo (varemo@chalmers.se)

The analysis of high-throughput data in biology is aided by integrative approaches such as gene-set analysis. Gene-sets can represent well-defined biological entities (e.g. metabolites) that interact in networks (e.g. metabolic networks), to exert their function within the cell. Data interpretation can benefit from incorporating the underlying network, but there are currently no optimal methods that link gene-set analysis and network structures. Here, we present Kiwi, a new tool that processes output data from gene-set analysis and integrates them with a network structure such that the inherent connectivity between gene-sets, i.e. not simply the gene overlap, becomes apparent. In two case studies, we demonstrate that standard gene-set analysis points at metabolites regulated in the interrogated condition. Nevertheless, only the integration of the interactions between these metabolites provides an extra layer of information that highlights how they are tightly connected in the metabolic network.Kiwi is a tool that enhances interpretability of high-throughput data. It allows the users not only to discover a list of significant entities or processes as in gene-set analysis, but also to visualize whether these entities or processes are isolated or connected by means of their biological interaction.

P22. Computational protein modelling based on limited NMR data

Maryana Wånggren[1], Martin Billeter[2] and Graham J.L. Kemp[1]

1. *Chalmers University of Technology*
2. *University of Gothenburg*

Presenting author: Maryana Wånggren (maryana@chalmers.se)

Introduction

Molecular dynamics simulation methods and simulated annealing are commonly used when calculating model protein structures from NMR data. These computational methods can require large amounts of data, possibly from many different NMR experiments. While some experiments are relatively straightforward and are routinely performed when studying a new protein, other experiments
are relatively time-consuming and expensive. The aim of this project is to develop a protein modelling program that can be used together with data from straightforward NMR experiments to obtain accurate model structures quickly, reducing or even eliminating the need for more expensive and complex multidimensional NMR experiments that require alternative isotope labelling to be done and take longer to perform.

Methods

We explore the conformational search space using the zipping and assembly method, which is a dynamic programming algorithm used here to construct longer fragments from pairs of shorter ones. Solutions are filtered using a variety of information from NMR experiments, including distance restrictions due to NOEs, torsion angle ranges predicted by TALOS+ from chemical shifts data, and secondary structure information based on HN-HN NOEs and predicted main chain torsion angles. Additional constraints are inferred from information about disulphide bridges and general knowledge of protein conformation.

Results

We are currently testing our method with a range of proteins by generating ensembles of structures based on only secondary structure information and disulphide bridges. The resulting models are compared with experimentally determined structures from the Protein Data Bank, and models obtained using simulated annealing methods and the same limited input.

P23. Complementing tissue characterisation by integrating transcriptome profiling from the Human Protein Atlas and from the FANTOM5 consortium

Nancy Y Yu[1], Björn M Hallström[2], Linn Fagerberg[2], Fredrik Ponten[3], Hideya Kawaji[4], Piero Carninci[5], The FANTOM Consortium[5], Alistair RR Forrest[5], Mathias Uhlén[2] and Carsten O Daub[1]

1. *Science for Life Laboratory, Karolinska Institutet*
2. *Science for Life Laboratory, Kungliga Tekniska högskolan*
3. *Department of Immunology, Genetics and Pathology, Science for Life Laboratory, Uppsala Universitet*
4. *RIKEN Preventive Medicine and Diagnosis Innovation Program, Japan*
5. *RIKEN Center for Life Science Technologies (CLST), Japan*

Presenting author: Nancy Yu (nancy.yu@scilifelab.se)

Understanding the normal state of human tissue transcriptome profiles is essential for recognizing tissue disease states and identifying disease markers. Recently, the Human Protein Atlas and the FANTOM5 consortium have each published extensive transcriptome data for human tissue, cell line, and primary cell samples using RNA-Seq and CAGE technologies, respectively. We performed the first large-scale complex tissue transcriptome comparison between datasets produced by two distinct technologies. We found that overall gene expression correlation was high between the 22 corresponding tissues analyzed ($R > 0.8$). For genes ubiquitously expressed across all tissues, the two datasets seem highly comparable with small differences indicating the need to update current gene models. Gene expressions specific to or highly enriched in single tissues also confirm the comparability, with differences likely attributed to methods used for tissue dissection leading to variation in cell type proportions. Our results show that while RNA-Seq and CAGE tissue transcriptome datasets serve well as reference data independently, the two datasets are actually very complementary and more values can be extract by combining the two datasets.

P24. Estimating copy number alteration in neuroblastoma: comparison of exome sequencing data and SNP microarrays

Malin Östensson[1], Susanne Fransson[1], Anna Djos[1], Niloufar Javanmardi[1], Per Kogner[2] and Tommy Martinsson[1]

1. Department of Medical and Clinical genetics, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden
2. Childhood Cancer Research Unit, Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden

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

Neuroblastoma show high degree of clinical heterogeneity (from tumors with fatal outcome to cases of spontaneous regression). Analysis of recurrent chromosomal aberrations such as losses of 1p, 11q, gain of 1q, 17q and/or MYCN amplification are currently used for patient stratification and definition of therapeutic strategy. Different analysis techniques for detection of segmental abnormalities include FISH, CGH-microarrays or multiplex ligation-dependent probe amplification (MLPA). However, as next-generation sequencing (NGS) becomes available for clinical use, this could also be used for assessment of copy number alterations simultaneously with mutational analysis.

We compare genomic profiles generated through the bioinformatical tool Control-FREEC on Exome sequencing (ES) data with profiles generated from Affymetrix 250K or 50K SNP-microarrays on 20 NB tumors of different stages. ES was performed by paired-end sequencing on Illumina instrumentation after DNA enrichment with Agilent SureSelect All Exome. The sequencing were performed at three separate occasions with median raw coverage of 91X, 127X and 340X respectively. Neuroblastoma tumors were normalized with either corresponding constitutional DNA or normal control DNA for NGS data while microarrays were normalized against healthy control DNA. Gross genomic changes were extracted through visual and/or ratio inspection from the 20 neuroblastoma tumors.

We constructed a web-based application using the R package Shiny, which visualizes the ControlFREEC results chromosome by chromosome or for the whole genome in one figure. The application also uses Fused Lasso Signal Approximator (FLSA) to detect segmental changes and breakpoints.

128 larger segmental aberrations and 63 numerical aneuploidies were detected through SNP-microarrays. Discrepancies between the methods were detected in three cases; one segmental loss detected through microarray was not confirmed by NGS-profiling, one smaller deletion was only detected through NGS-profiling and in one instance both methods indicated 2p-gain but at different positions. Otherwise the results were concordant. Furthermore, through ES data we detected ALK-mutations in four patients, ATRX-deletions in two patients and chromothripsis of chromosome 5 in one patient.

Exome sequencing could be used for diagnosis of neuroblastoma tumors combining mutational screening with detection of common chromosomal/segmental aneuploidies such as 2p-gain, 17q-gain, 11q-deletion, MYCN-amplification and alterations of ALK or ATRX.

P25. HierBin - A method for detailed functional annotation and quantification in metagenomes

Tobias Österlund[1], Viktor Jonsson[1] and Erik Kristiansson[1]

1. Department of mathematical sciences, Chalmers University of Technology and University of Gothenburg, Sweden

Presenting author: Tobias Österlund (tobiaso@chalmers.se)

Metagenomics aims for sequencing and identification of DNA from communities of microorganisms in their natural environment. One main challenge in metagenomic data analysis is the lack of reference sequences. As an example it has been estimated that only 0.05% of all bacterial species has a fully sequenced reference genome. Instead metagenomic methods rely on *de novo* assembly of sequencing reads followed by gene prediction methods, typically relying on finding functional domains using PSSMs or HMMs (e.g. PFAM, TIGRFAM and COG). However, this procedure results in a broad classification of gene products since such domains often classifies many proteins into the same functional domain. The human gut gene catalog, for instance, is predicted to contain more than 5 million genes, but the number of PFAM domains is only around 10,000.

Here we present a new method for functional annotation and quantification in metagenomes that gives a more detailed functional description of the metagenome. The method involves a supervised classification step where known functional domains are identified in the assembled metagenome, followed by an unsupervised classification step where all sequences belonging to a domain are clustered based on sequence similarity to create subclusters. The abundance of each subcluster is then quantified by mapping the sequence reads to the subcluster sequence. The last step is a differential analysis step where the abundance of each subcluster is compared between metagenomes or between conditions.

We apply our analysis method both to previously published metagenomic datasets from the human gut as well as new metagenomic data from environmental samples. We show, using resampling methods, that our method performs better in predicting differential abundant genes than methods using only functional domains for gene prediction and detects differences at higher resolution that would be invisible at lower resolution. The method for functional annotation and quantification is available as Python and R scripts.

Travel Instructions and maps

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.
Note: Central Station is not in the following map.

Nearest tram stop

The closest public transport stop is Medicinaregatan. Tram 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. Please refer to the map above for walking directions.

How to reach Central Station

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

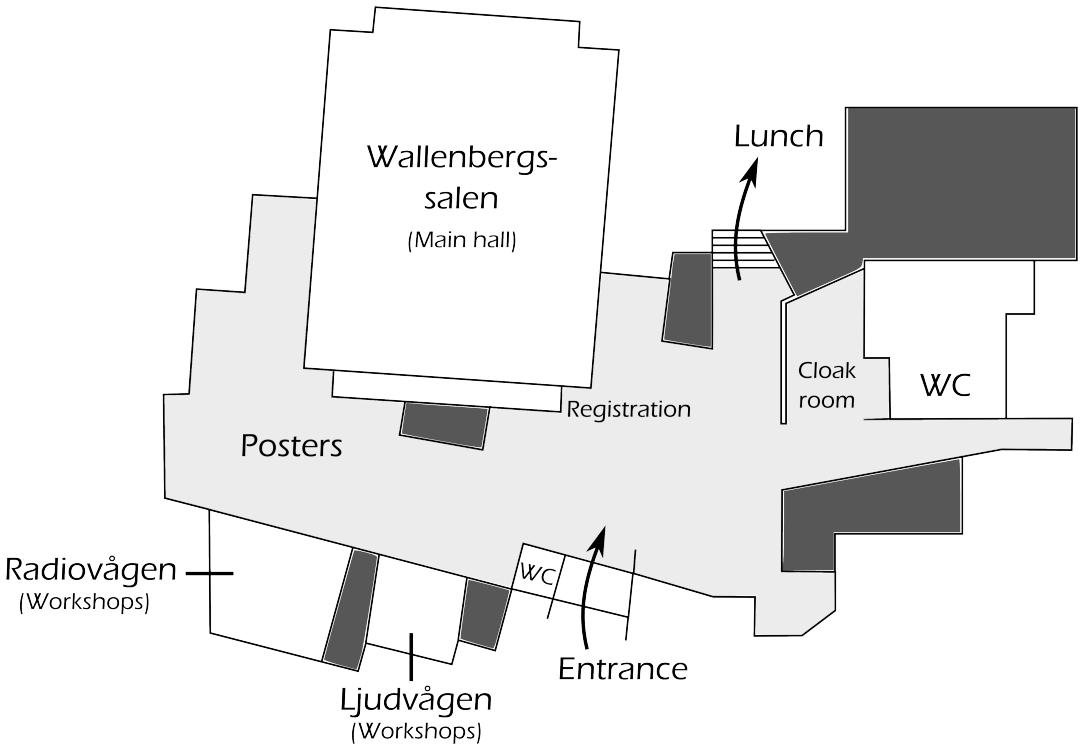
Where to get Västtrafik tickets

Buy your tickets at one of Västtrafik shops or at ticket sales points, Pressbyrån. Note: You cannot pay with cash on the trams or buses. If you register your mobile phone number, you can pay by SMS.
http://www.vasttrafik.se/biljetter-priser/enkelbiljett/nytt-sms-system/registrera-sms/
(page only in Swedish).

Find dinner venue

The dinner takes place in RIVER CAFÉ Restaurant. You can take the tram 6 from Medicinaregatan to Järntorget and from there you walk to Rosenlund Färjeläge. Take the boat from Rosenlund and get off at Eriksbergs Färjeläge. Please note that you have to have a Västtrafik ticket for the boat.
If you are not interested in taking boat, it is possible to go by bus to Eriksbergstorget and walk from there to the restaurant. Please refer to the map on page 48 for directions.
Address: Dockpiren, Eriksberg, Göteborg, 41764, Sweden
Phone:+46 31 51 00 00

Map of Conference Centre Wallenberg



Attendee List

First Name	Last Name	Email	Affiliation
Tobias	Abenius	tobbe@tobbe.nu	University of Gothenburg
Anna	Abramova	anna.abramova@cmb.gu.se	University of Gothenburg
Sofia	Adolfsson	sofia.adolfsson@med.lu.se	Lund University
Mohammad Tanvir	Ahamed	tanvirm@student.chalmers.se	Chalmers University of Technology
Fioralba	Ajazi	fioralba@maths.lth.se	Lund University
Rozita	Akrami	rozita.akrami@wlab.gu.se	University of Gothenburg
Babak	Alaeimahabadi	babak.alaeimahabadi@gu.se	University of Gothenburg
Johannes	Alneberg	johannes.alneberg@scilifelab.se	KTH Royal Institute of Technology
Ahmed	Arslan	ahmed.arslan@biw.kuleuven.be	KU Leuven
Arghavan	Ashouri	gusashoar@student.gu.se	University of Gothenburg
Nurselin	Ates	gusatenu@student.gu.se	University of Gothenburg
Mattias	Backman	backman.91@hotmail.com	University of Gothenburg
Mauricio	Barrientos	mauricio.barrientos@ki.se	Karolinska Institute
Daisy	Belfield	drjbelfield@gmail.com	AstraZeneca
Graham	Belfield	graham.belfield@astrazeneca.com	AstraZeneca
Johan	Bengtsson-Palme	johan@microbiology.se	University of Gothenburg
Fanny	Berglund	fannyb@chalmers.se	Chalmers University of Technology
Joydeep	Bhadury	joydeep.bhadury@gu.se	University of Gothenburg
Magnus	Bjursell	Magnus.Bjursell@astrazeneca.com	AstraZeneca
Fredrik	Boulund	fredrik.boulund@chalmers.se	Chalmers University of Technology
Moritz	Buck	murumbii@gmail.com	Uppsala University
Mathias	Cardner	mathias.cardner@gmail.com	University of Gothenburg
Dongfeng	Chen	dongfeng.chen@gu.se	University of Gothenburg
Marcus	Claesson	m.claesson@ucc.ie	University College Cork, Ireland
Marija	Cvijovic	marija.cvijovic@chalmers.se	Chalmers University of Technology
Johan	Dahlberg	johan.dahlberg@medsci.uu.se	Uppsala University
Niklas	Dahr	niklas.dahr@gu.se	University of Gothenburg
Marcela	Davila	marcela.davila@gu.se	University of Gothenburg
Pierre	De Wit	pierre.de_wit@bioenv.gu.se	University of Gothenburg
Anders	Edsjö	anders.edsjo@gu.se	Sahlgrenska University Hospital
Alexander	Eiler	alexander.eiler@ebc.uu.se	Uppsala University
Martin	Engqvist	martin.engqvist@chalmers.se	Chalmers University of Technology
Phil	Ewels	phil.ewels@scilifelab.se	NGI Stockholm
Amir	Feizi	feizi@chalmers.se	Chalmers University of Technology
Dawn	Field	fiedawn@gmail.com	NERC Center for Ecology and Hydrology
Alma	Forsman	alma.forsman@gu.se	University of Gothenburg
Johan	Fredriksson	johan.fredriksson@gu.se	University of Gothenburg
Manuel	Garcia	gmanuel@chalmers.se	Chalmers University of Technology
Francesco	Gatto	gatto@chalmers.se	Chalmers University of Technology
Pouyan	Ghaffari Nouran	ghaffari@chalmers.se	Chalmers University of Technology
Homa	Ghiais	homa.gh11@gmail.com	

First Name	Last Name	Email	Affiliation
Mario	Giovacchini	mario.giovacchini@scilifelab.se	NGI Stockholm
Mika	Gustafsson	mika.gustafsson@liu.se	Linköping University
Rickard	Hammarén	gusrickaha@student.gu.se	University of Gothenburg
Shaghayegh	Hosseni	shaghayegh.hosseini@chalmers.se	Chalmers University of Technology
Olga	Hrydziuszko	olga.hrydziuszko@gmail.com	Karolinska Institute
Elisabeth	Israelsson	elisabeth.israelsson@astrazeneca.com	AstraZeneca
Mathias	Johansson	math.joh.90@gmail.com	University of Gothenburg
Anna	Johnning	anna.johnning.gu@gmail.com	Chalmers University of Technology
Viktor	Jonsson	viktor.jonsson@chalmers.se	University of Gothenburg
Jekaterina	Jutkina	jekaterina.jutkina@gu.se	University of Gothenburg
Jonatan	Kallus	kallus@chalmers.se	Chalmers University of Technology
Wenjing	Kang	wenjingk@kth.se	KTH Royal Institute of Technology
Fredrik	Karlsson	frekar@chalmers.se	Chalmers University of Technology
Joakim	Karlsson	j.w.karlsson@gmail.com	University of Gothenburg
Fatemeh	Kaveh	fatemeh.kaveh@medisin.uio.no	Oslo University
Eduard	Kerkhoven	eduardk@chalmers.se	Chalmers University of Technology
Mehmood	Khan	malagori@kth.se	KTH Royal Institute of Technology
Niklas	Korsbo	niklas.korsbo@gmail.com	Lund University
Erik	Kristiansson	erik.kristiansson@chalmers.se	Chalmers University of Technology
Mattias	Köpsén	matt.kopsen@gmail.com	Linköping University
Petri	Lahtvee	lahtvee@chalmers.se	Chalmers University of Technology
Tiange	Lang	langtiange@xtbg.org.cn	University of Gothenburg
Britta	Langen	britta.langen@gu.se	University of Gothenburg
Anna	Larsson	anna.larsson@chalmers.se	Chalmers University of Technology
Erik	Larsson	erik.larsson@gu.se	University of Gothenburg
John	Larsson	john.larsson@scilifelab.se	Linnaeus University
Julie	Lorent	julie.lorent@ki.se	Karolinska Institute
Joakim	Lundeberg	joakim.lundeberg@scilifelab.se	KTH Royal Institute of Technology
Erik	Ländström	erik.landstrom@hotmail.se	University of Gothenburg
A K M Firoj	Mahmud	aakk0004@student.umu.se	Umeå University
Erik	Malmberg	erik.malmberg@clinchem.gu.se	University of Gothenburg
Arfa	Mehmood	arfa.mehmood@utu.fi	University of Turku
Sayyed Auwn	Muhammad	auwnm@kth.se	KTH Royal Institute of Technology
Daniel	Muthas	daniel.muthas@astrazeneca.com	AstraZeneca
Jens Christian	Nielsen	jens.c.nielsen@chalmers.se	Chalmers University of Technology
Avlant	Nilsson	avlant@hotmail.com	Chalmers University of Technology
Henrik	Nilsson	henrik.nilsson@bioenv.gu.se	University of Gothenburg
Jacqueline	Nowak	janowak@uni-potsdam.de	University of Potsdam
Anders	Nygaard	anders.benteson-nygaard@hioa.no	Oslo and Akershus University College
Ogonna	Obudulu	ogonna.obudulu@slu.se	Swedish University of Agricultural Sciences
Lisa	Olsson	olsson.lisa@gu.se	University of Gothenburg
Laura	Orellana	laura.orellana@scilifelab.se	KTH Royal Institute of Technology
Mattias	Ormestad	mattias.ormestad@scilifelab.se	Science for Life Laboratory

First Name	Last Name	Email	Affiliation
Chandan	Pal	chandan.pal@gu.se	University of Gothenburg
Mariana	Pereira	mariana.pereira@chalmers.se	Chalmers University of Technology
Jens	Persson	jenez1990@gmail.com	University of Gothenburg
Balaji	Rajashekar	balaji@ut.ee	University of Tartu
Nils	Rudqvist	nils.rudqvist@radfys.gu.se	University of Gothenburg
Millie	Rådjursöga	millie.radjursoga@gu.se	University of Gothenburg
Ladan	Saadat Vafa	nasturtiumladan@gmail.com	
Vaibhav	Sabale	vaibhavsabalebi@gmail.com	North Maharashtra University
Johanna	Sagemark	johanna.sagemark@astrazeneca.com	AstraZeneca
Kemal	Sanli	kemal.sanli@bioenv.gu.se	University of Gothenburg
Emil	Schüler	emil.schuler@gu.se	University of Gothenburg
Nicole	Selzer	nicole.selzer@molbiol.umu.se	Umeå University
Partho	Sen	sarathi@chalmers.se	Chalmers University of Technology
Sviatlana	Shashkova	sviatlana.shashkova@gu.se	University of Gothenburg
Sanea	Sheikh	sanea.sheikh@ebc.uu.se	Uppsala University
Benjamín	Sigurgeirsson	benjamin.sigurgeirsson@scilifelab.se	KTH Royal Institute of Technology
Anders	Sjögren	anders.sjogren@chalmers.se	Chalmers University of Technology
Patrik	Smeds	patrik.smeds@medsci.uu.se	Uppsala University
Agata	Smialowska	agata@chalmers.se	Chalmers University of Technology
Johan	Spetz	johan.spetz@gu.se	University of Gothenburg
Babett	Steglich	Babett.Steglich@ki.se	Karolinska Institute
Thomas	Svensson	thomas.svensson@scilifelab.se	Science for Life Laboratory
Venkat	Talla	tallavenkat.vit@gmail.com	Uppsala University
Sylvie	Tesson	sylvie.tesson@gu.se	University of Gothenburg
Subazini	Thankaswamy	subazini@chalmers.se	Chalmers University of Technology
Kaisa	Thorell	kaisa.thorell@gu.se	University of Gothenburg
Reaz	Uddin	mriazuddin@iccs.edu	University of Karachi
Benjamin	Ulfenborg	benjamin.ulfenborg@his.se	University of Skövde
Francesco	Vezzi	francesco.vezzi@scilifelab.se	Science for Life Laboratory
Leif	Väremo	varemo@chalmers.se	Chalmers University of Technology
Mick	Watson	mick.watson@roslin.ed.ac.uk	Roslin Institute
Niek	Welkenhuysen	niek.welkenhuysen@gu.se	University of Gothenburg
Emma	Wijkmark	wijkmark@chalmers.se	Chalmers University of Technology
John	Wilson	jawilson@illumina.com	Illumina, AB
Maryana	Wånggren	maryana.wanggren@gmail.com	Chalmers University of Technology
Nancy	Yu	nancy.yu@scilifelab.se	Karolinska Institute
Yan	Zhou	yan.zhou@ki.se	Karolinska Institute
Yongjin	Zhou	yongjin@chalmers.se	Chalmers University of Technology
Malin	Östensson	malin.ostensson@gu.se	University of Gothenburg
Tobias	Österlund	tobiaso@chalmers.se	Chalmers University of Technology

swedish bioinformatics workshop 2014

BILS Bioinformatics Infrastructure for Life Sciences

SciLifeLab

formas

LIFE SCIENCE ENGINEERING A CHALMERS AREA OF ADVANCE

SILS SystemsBiology Infrastructure for the Life Sciences

UNIVERSITY OF GOTHENBURG

	Thursday October 23	Friday October 24
9:00		Keynote: Dawn Field Main Hall
15		
30	Registration and Fika Wallenberg conference center – Foyer	
45		
10:00		Fika
15		
30	Opening address Main Hall	Presentations Main Hall Johannes Alneberg Sylvie Tesson
45		
11:00	Keynote: Joakim Lundeberg Main Hall	Presentations Main Hall Niek Welkenhuysen Mika Gustafsson Olga Hrydziuszko Workshop 3 Radiovågen Astra Zeneca Workshop 4 Ljudvågen Marija Cvijovic
15		
30		
45		
12:00	Lunch The Lyktan Restaurant	Lunch The Lyktan Restaurant
15		
30		
45		
13:00	Presentations Main Hall Francesco Vezzi Johan Dahlberg	Keynote: Mick Watson Main Hall
15		
30		
45		
14:00	Presentations Main Hall Chandan Pal Francesco Gatto Johan Bengtsson-Palme Workshop 1 Radiovågen Anders Edsjö Workshop 2 Ljudvågen Alexander Eiler	Sponsor presentation: BILS, WABI & SciLifeLab Main Hall
15		
30		Fika
45		
15:00	Fika	Presentations Main Hall Nicole Selzer: Johan Fredriksson Amir Feizi Workshop 5 Radiovågen Erik Larsson Workshop 6 Ljudvågen Henrik Nilsson
15		
30	Sponsor presentation: Astra Zeneca Main Hall	
45		
16:00	Presentations Main Hall Avlant Nilsson Mattias Köpsén	
15		
30		
45		
17:00	Poster session Wallenberg conference center – Foyer	
15		
30		
45		
18:00		Closing address and awards Main Hall
15		
30		
45		
19:00	Conference dinner River Café	
15		
30		
45		

AstraZeneca

Gothenburg Centre for Systems Biology

This workshop is supported with an educational grant from AstraZeneca