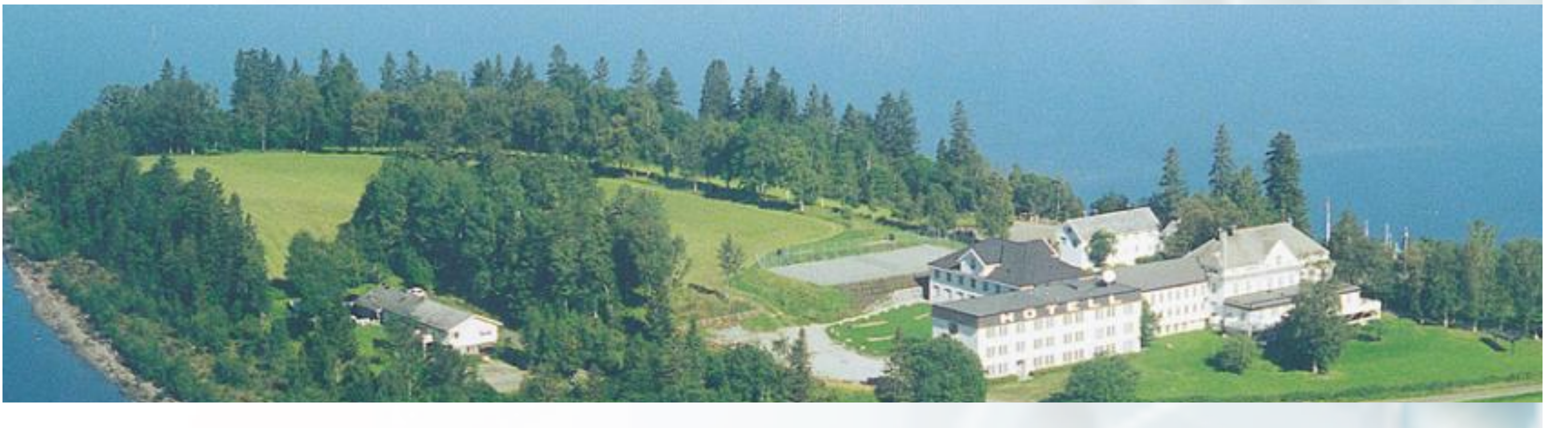
Program for the 2nd annual conference of NORBIS

September 14-16 2016

Selbusjøen hotell & gjestegård



NORBIS
The national research school in bioinformatics, biostatistics and systems biology

The Research Council of Norway

Contents **Page**

Program
2nd Annual Meeting of NORBIS, September 14-16 2016
Selbusjøen Hotell og Gjestegård

Wednesday 14. September

Chair: Kyrre Lekve (Simula Research Laboratory)

16:00 **Welcome** by Inge Jonassen (UiB)

16:15 **Predicting the organization and evolution of microbial metabolism** by Martin Lercher (Heinrich Heine University Dusseldorf)

17:30 **Poster session 1**
Odd number presenters

19:30 *DINNER*

Thursday 15. September

Chair: Ørnulf Borgan (UiO)

09:00 **Hypothesis Testing and Reproducability in Genomics** by Mette Langaas (NTNU)

09:50 **Student talks:**

A.
The Obesity-Cancer Connection: Malignant Cells and Metastasis due to Altered Metabolic Landscape of Obese Cancer Patients by Pouda Panahandeh (UiB)

10:05 **B.**
Combining omics data to predict the outcome of follicular lymphoma patients by Chloe B. Steen (UiO)

10:20 **C.**
CRISPR tools for XXI century genome engineering by Kornel Labun (UiB)

10:40 *BREAK (short feedback session)*

11:10 **D.**
Differential DNA methylation at conserved non-genic elements and transgenerational inheritance following mono(2-ethylhexyl) phthalate and 5-azacytidine exposure in zebrafish by Jorke Kamstra (NMBU)

11:25 **E.**
Enrichment of human specific differentially methylated regions in Schizophrenia by Niladri Banerjee (UiB)

11:40 **F.**
Two new software packages for finding enriched domains in ChIP-Seq data by Endre Bakken Stovner (NTNU)

12:00 *LUNCH (short feedback session)*

13:00 **Forum for supervisors:**
Some experiences and thoughts about PhD supervision and cultural differences by Rolv Bræk (NTNU)

Forum for students:
Group work, with emphasis on supervision

14:30 **Summary of group work**

15:00 *BREAK*

15:30 *TEAM-BUILDING ACTIVITY*

17:30 **Poster session 2**
Even number presenters

19:30 *DINNER (with poster prize)*

Friday 16. September

Chair: Gos Micklem (University of Cambridge)

09:00 **Keynote lecture** by Pål Sætrom (NTNU)

09:50 **Student talks**

G.
Transcriptome-Wide Discovery of Cell-Cycle Genes by Antonin Klima (NTNU)

10:05 **H.**
Sea anemones in the climate change research – transcriptome profiling approach by Ilona Urbarova (UiT)

10:20 **I.**
Prokaryotic classification – novel insight in 16S rRNA-based classification by Hilde Vinje (NMBU)

10:40 *BREAK (short feedback session), check out*

11:10 **Keynote lecture** by Ole Christian Lingjærde (UiO)

12:00 **NORBIS members – who are you?** by Gunnar Schulze (UiB)

12:15 **Closing remarks** by Manuela Zucknick (UiO)

12:30 *LUNCH*

14:00 *DEPARTURE by bus to Værnes and Trondheim (arrival at Værnes appr. 14:30 / Trondheim 15:30)*

Practical information

Bus to and from Selbusjøen

On September 14, a NORBIS bus will depart from St Olavs Hospital (Labsenteret / Finalebanen) at 13.30 and from Trondheim Airport Værnes (to the left outside the arrivals hall) at 14.30. Those of you travelling from Værnes, please enter the bus as soon as possible after 14.10, to allow departure no later than 14.30. The bus will arrive at Selbusjøen at 15.00. On September 16, the bus departs from Selbusjøen at 14.00 and will arrive at Trondheim Airport Værnes around 14.30 *at the earliest*. If you want to join the bus service to Selbusjøen, please let us know at contact-norbis@uib.no *as soon as possible* and arrange your flights accordingly. Selbusjøen can also be reached by public transport, see www.atb.no for options.

Student talks

We have invited nine PhD students to present their research project by giving a talk. The talks will be of a duration of 12 minutes with additional 3 minutes for questions and discussion. We will arrange a feedback committee for each session.

Poster sessions

We will arrange two poster sessions, one on Wednesday and one on Thursday. Abstracts with odd numbers will present on Wednesday, whereas abstracts with even numbers will present on Thursday. A committee will evaluate all posters and their presentation, and award a prize for the best poster during dinner Thursday evening.

Forum for supervisors

This year, Professor Rolv Bræk from Department of Telematics at NTNU will bring your attention to supervision in light of individual and cultural differences; that different forms of PhD supervision are suitable for different types of PhD students and phases in the PhD study. He will further discuss typical cultural differences and how they can be addressed to avoid cultural problems.

Team building activity

Outdoor challenges at the nearby Camp Eggen will help us get to know each other and work together as a team.

Expenses

NORBIS covers all expenses (travel, accommodation and food) for its PhD students. Master students and post docs may apply to have support. In return, we expect you to present your work by a poster or a student talk. Please keep all your receipts, and we will inform you how to get reimbursed after the meeting.

Abstracts for student talks

A.
The Obesity-Cancer Connection: Malignant Cells and Metastasis due to Altered Metabolic Landscape of Obese Cancer Patients

Pouda Panahandeh, PhD student, Department of Biomedicine, University of Bergen

Obesity, characterized as increased adipose tissue mass throughout the body, has shown to be significantly associated with incidence, relapse or mortality of cancer patients. Different cancer types are differently affected by high body mass index (BMI). Of these, BMI affects mortality rates in endometrial cancer the most. As the prevalence of overweight and obesity have dramatically increased worldwide in the past decades, studying the molecular biology of obesity-related cancer development is an important call for this study. Therefore, we aim to understand how cancer cells adapt and benefit the alteration in the metabolic landscape of obese cancer patients by altering gene-expression profile.

To this end, we obtained transcriptomic microarray profiles from paired primary and metastasis tumor nodules in a cohort of endometrial cancer patients. To identify differentially expressed obesity-induced/repressed metastatic genes, we performed a Spearman's rank correlation between log2 metastasis/primary gene expression ratio for each gene and corresponding patients' BMI values. A permutation test was performed to the Spearman's ranked correlation hypothesis, in which a p-value ≤ 0.005 were adjusted as significance. Amongst differentially expressed genes, the progestin and adipoQ receptor 4 (PAQR4) have been shown to be upregulated significantly (p-value ≤ 0.005). Depletion of PAQR4 using RNAi technology impedes growth of endometrial cancer cell lines. This might be related to a mitochondrial function and associated senescence. However, in vitro and in vivo studies will be conducted to further understand how PAQR4 interferes in mitochondrial functionality in the obese context.

B.
Combining omics data to predict the outcome of follicular lymphoma patients

Chloe B. Steen, PhD student, Group for Biomedical Informatics, Department of Informatics, University of Oslo

Introduction: Follicular Lymphoma (FL) is a cancer that affects B-lymphocytes, a type of white blood cells vital to our immune response. FL is an indolent disease, meaning that it often causes little or no symptoms for an extended time period. A subgroup of FL patients experience at some point in time a transformation to a more aggressive phenotype, with much poorer survival. Being able to predict FL transformation at the time of diagnosis is important, both to understand the mechanism and in order to provide the best possible treatment. In 2014, our group published a study that combined the analysis of DNA copy number data and gene expression data and identified multiple markers for FL transformation. So far, the study has not been validated in independent patient cohorts. To complicate matters, the original study was based on biopsies from patients that were treated before the introduction in the clinic of the monoclonal antibody rituximab, a drug that has profoundly affected the survival of many lymphoma patients. It is of great interest to establish whether the transformation predictor found in the study from 2014 is still relevant today with the current treatment regime.

Material and methods: We analyzed SNP6.0 allele-specific copy number data and Affymetrix gene expression array data from 80 FL patients from Norway and Sweden. The patients were included between 2002-2008 in a phase III study where they received treatment of rituximab, either alone or in combination with interferon-alpha.

Results: The 698 significantly correlated genes in-cis identified in the original study show positive correlations in the present data set. Furthermore, the magnitude of the gene signature scores are comparable in the two studies, and one out of the six gene signatures found in the original study significantly predicts transformation in the new patient cohort. In addition, across all six gene signatures there is good correspondence between the effect size found in the original study and in the current study.

Conclusion: The results from the rituximab era show good agreement with the findings published in the original study. Although standard treatment has changed for FL patients, the findings in Brodtkorb et al (2014) are still valid and should be further investigated.

C.

CRISPR tools for XXI century genome engineering

Kornel Labun, PhD student, Computational Biology Unit, Department of Informatics, University of Bergen

Few breakthroughs in recent years can rival that of the discovery and repurposing of CRISPR. Initially a defense mechanism of the bacterium Streptococcus pyogenes, CRISPR was repurposed as a molecular tool for inducing targeted mutations. Using a short guide RNA to specify the target loci CRISPR can be used to silence genes, activate genes, cleave DNA, introduce single point mutations or frameshifts to make knockout mutants. The standard CRISPR system induces a double-strand break which is repaired in the cell either by non-homologous end-joining pathway or homology directed repair.

While highly promising CRISPR is far from perfect and a number of studies are being conducted to improve efficiency and precision. To aid researchers in the design of CRISPR experiments we have developed two tools: 1) CHOPCHOP, a web-server tool for finding optimal guide RNAs. CHOPCHOP supports a large number of genomes and incorporates the latest findings in terms of guide specificity and efficiency. And 2) amplican, an R package to automate analysis of CRISPR induced mutations using high throughput sequencing for validation. Together these enables a researcher to do high-throughput screens and take their CRISPR experiments from conception to completion.

D.

Differential DNA methylation at conserved non-genic elements and transgenerational inheritance following mono(2-ethylhexyl)phthalate and 5-azacytidine exposure in zebrafish

Jorke Kamstra, PhD student, Department of Basic Science and Aquatic Medicine, Norwegian University of Life Sciences, Oslo

It is hypothesized that exposures during early development can cause latent effects that could be transferred over generations. Recent studies demonstrate environmentally induced transgenerational epigenetic effects linked to increased disease risk over generations, for instance on endocrine and reproductive systems. Some of those studies observed changes in the DNA methylome, indicating that effects can be inherited via DNA methylation. Here, we used a transgenerational set up using zebrafish as a vertebrate model organism to investigate the acute, latent and transgenerational effects on DNA methylation of early life exposure to the putative obesogenic phthalate metabolite mono(2-ethylhexyl) phthalate (MEHP) and the DNA methylation inhibitor 5-azacytidine (5AC). We performed genome wide DNA methylation analysis using reduced representation bisulfite sequencing (RRBS) on 6 dpf exposed embryos (30 and 10 μM for MEHP and 5AC, respectively). In addition, we performed bisulfite amplicon sequencing on 10 differentially methylated regions (DMRs) discovered by RRBS. RRBS analysis revealed 410 and 580 DMRs for MEHP and 5AC, respectively, with strong enrichment at conserved non-genic elements. Ingenuity pathway analysis of genes associated to these DMRs revealed involvement of MEHP in adipogenic and neuronal developmental processes. For 5AC, general stress responses and effects on embryonic development were enriched, as well as regulators for neuronal development. Bisulfite amplicon sequencing revealed persistent effects up to F2 at 2 and 6 out of the 10 analyzed loci, for MEHP and 5AC respectively. Our results reveal considerable effects on DNA methylation following exposures during early life in zebrafish to MEHP and 5AC at non-toxic concentrations. Enrichment at conserved non-genic elements implies a role of compound induced changes in DNA methylation, conserved throughout evolution. Persistent effects on methylation in F2 implies that DNA methylation changes can be inherited by multiple generations. This study will provide new knowledge about transgenerational epigenetics in zebrafish.

E.

Enrichment of human specific differentially methylated regions in Schizophrenia

Niladri Banerjee, PhD student, Department of Clinical Sciences, University of Bergen

Background: Schizophrenia is a heritable psychiatric disorder with a prevalence of 1 in every 100 individuals and persistence recorded throughout human history. One possible explanation for this persistence is the so-called evolutionary hypothesis that describes the disease as a by-product of human evolution. Recently, Gokhman et al. identified genomic regions that give a glimpse into the evolution of the human epigenome. They described a set of regions that are differentially methylated in modern humans by comparing with Neanderthal and Denisovan methylation levels. We tested if genetic variants in these human differentially methylated regions (DMRs) could be enriched for association with several complex traits, and especially schizophrenia.

Results: We tested enrichment of association for the human DMRs in 12 GWAS of various phenotypes. The human DMRs show enrichment of association only in the schizophrenia GWAS. The enrichment was consistent whether we consider single-nucleotide polymorphisms in linkage disequilibrium or just overlapping with DMR regions. We observed that only human DMRs had enrichment of association for schizophrenia; not Neanderthal or Denisovan DMRs. Finally, we show that the enrichment seen in human DMRs is comparable with the enrichment previously reported for other regions of human specific evolution: Neanderthal Selective Sweep and human DMRs were more enriched for association with schizophrenia than Human Accelerated Regions.

Conclusion: Methylation changes occurring over evolutionary time frame in humans occurred in regions that show enrichment of genetic variants leading to vulnerability to Schizophrenia. Other recent genomic studies show that evolution may have played a key role in making schizophrenia a part of the human story, we now show that regions of epigenetic variation may have contributed as well.

F.

Two new software packages for finding enriched domains in ChIP-Seq data

Endre Bakken Stovner, PhD student, Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim

ChIP-Seq is an experimental method used to analyze how a protein of interest interacts with DNA. The end result of a ChIP-Seq experiment is a pool of DNA fragments that binds to the protein. These fragments must be aligned and further analyzed to find so-called enriched regions - the genomic regions where the protein binds. Different methods are needed depending on whether the enriched regions are short and the signal is strong or the region is long and the signal diffuse. This poster introduces two methods appropriate for each type, namely triform2 and epic.

Triform2 is a software package for finding narrow, enriched regions such as binding sites for transcription factors. It does not make assumptions about the data or use fitted parameters. It achieves robustness by relying on general properties of the aligned ChIP-Seq data to find enriched regions. Epic is a complete reimplementation of the extremely popular SICER algorithm. Epic works by finding regions that are enriched according to a poisson model and then giving these an FDR-value based on the number of control (background) reads in the region. In addition to a much improved speed and memory efficiency epic includes many new features such as paired-end support and the possibility of analyzing multiple files together.

Both epic and triform2 contain novel facilities for making it very easy to 1) visualize the resulting enriched regions in a genomic context and 2) run more sophisticated statistical tests on the resulting enriched regions from different conditions. The programs have been developed with ease of installation and user-friendliness in mind.

https://github.com/endrebak/epic https://github.com/endrebak/triform2

G.

Transcriptome-Wide Discovery of Cell-Cycle Genes

Antonin Klima, PhD student, Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim

Your research problem is related to the cell cycle, yet nothing comes up significant? If only you could restrict your search to RNAs that are changing with the cell cycle. Suddenly, those Bonferronis, Students and the like would become flyweight opponents. But what portion of such RNAs has been found? And how do we find more?

Experience shows that, in bioinformatics, when lots of data just won't do, lots and lots of data will. We put together data from different sources, and spice it up with a few biologically-motivated algorithmic ideas to get a more complete overview of cell-cycle genes.

Naturally, the data don't give themselves up easy. Not only do they come from different teams, but also from different tissues, experimental designs, and platforms. Perhaps that is why everyone thus far has rather created their very own precious dataset to work on. In fact, so did we. We show how we have tackled the merging process, harmonizing the differences among the datasets, and how the novel algorithmic ideas have helped us connect the datasets.

Perhaps soon also your cell-cycle related research (psst.. cancer) will light up more significant results?

H.

Sea anemones in the climate change research – transcriptome profiling approach

Ilona Urbarova, PhD student, Department of Medical Biology, University of Tromsø - The Arctic University of Norway

Climate change represents significant stress for all marine species, but some are able to cope with the stress better than others. Ocean acidification research shows mainly negative effects of low pH on marine organisms and decrease in pH is predicted to cause biodiversity loss. However, climate change studies are usually performed as short-term laboratory experiments including only one species, which does not capture naturally occurring long-term changes in the whole ecosystem.

We sampled adult polyps of the sea anemone Anemonia viridis from three various locations at a natural acidification site in Italy. Here, the seawater pH gradient from about 8.2 to 6.5 is created by the release of CO2 bubbles from an underwater volcanic vent site. We assessed the whole transcriptome profile of the sea anemones from the three locations of decreasing pH by performing poly-A enriched transcriptome and small RNA sequencing. Interestingly, small RNAs appear to be involved in the response to ocean acidification conditions. Furthermore, our differential expression analysis of protein-coding genes revealed that A. viridis is extremely stressed at low seawater pH. However, previous reports from the same location based on morphological and physiological assessment concluded that A.viridis is thriving at low seawater pH. Therefore, we suggest transcriptome profiling as more comprehensive and highly sensitive approach in monitoring the current state of an animal upon environmental changes.

I.
Prokaryotic classification – novel insight in 16S rRNA-based classification

Hilde Vinje, PhD student, Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås

DNA sequencing is a high prioritized area in science, because it provides us with the most basic information of all: The DNA sequence of nucleotides. These makes it able to identify and survey genomes as well as explore microbes and the dynamics of microbial communities. Over the last 10 years a new sequencing technology have revolutionized the field; Sequencing has gone from low throughput and high costs, too new high-throughput genome sequencing. Both the cost and the time-efficient have been reduced significantly, generating an enormous amount of sequencing data available.
This amount of sequence data gives us a solid foundation to understanding the biologically phenomena, but it has also introduced an increased need for time and computer-efficient methods to extract information from data such as these.

The 16S ribosomal RNA (rRNA) gene serves as a molecular marker for investigating microbes or microbial community composition and structure and huge amount of available 16S RNA sequences provide us with a rapidly increasing reference framework for comparison and identification of new sequence data.

The aim of my PhD project has been to improve classification of Prokaryotes, both by presenting new insight concerning the tree of life and by rigorously testing and presenting different classification methods.

Here I will present some of the problems we face today, while considering taxonomic classification of prokaryotes, and suggestions for improvement supported by the results from the project.

Abstracts for posters

1.
The rainfall plot: its motivation, characteristics and pitfalls

Stefania Salvatore, Post doc, Group for Biomedical Informatics, Department of Informatics, University of Oslo

A visualization referred to as rainfall plot has recently gained popularity, mostly to visualize the distribution of somatic mutations in cancer along the genome. The plot shows location of points along one dimension (x-axis) versus each point's distance from the preceding point (y-axis).

Despite its frequent use, the motivation for applying this particular visualization for somatic mutations and the appropriate interpretation of visual characteristics have never been explicitly described. We show that the rainfall plot can have a useful purpose, as a way of combining information on frequency and clustering (line plot and histogram) in a single plot. However, the productive use of rainfall plots require a precise understanding of basic properties of these plots, as well as potential pitfalls in interpretation. Furthermore, if the only purpose of using the plot is to show an increased frequency of mutations in particular regions of the genome, a standard line plot of frequency would show the same information in a more direct manner.

We here describe and exemplify basic properties of rainfall plots, helping the correct interpretation of both the frequency and clustering spectrum of the inherent information. Also, we describe various potential pitfall and the situations in which they may arise, as well as quantifying their effect in such exemplary situations.

2.
Connecting Obesity and Breast Cancer

Xiaozheng Liu, Master student, Department of Biomedicine, University of Bergen

Obesity is known to be associated with the initiation, progression and ultimately survival of breast cancer patients, but little is known about the specific gene expression changes that are deregulated by obesity and how these affect patient mortality. Here, we propose to identify genes and pathways that correlate with both body mass index (BMI) and the survival time of breast cancer patients.
To this end, we have obtained microarray data sets and disease specific survival rates of 261 breast cancer patients. Using permutation-based Spearman’s rank-order correlation, we have identified genes that are regulated in a BMI specific manner and that predict patient disease specific survival. Next, using bioinformatics tools we have performed GSEA pathway analysis and gene enrichment analysis of these genes, in order to extract signaling pathways that play important roles in obesity-induced breast cancers and which can be used for further downstream studies.

3.
The extreme phenotype sampling design

Thea Bjørnland PhD student, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim

Extreme Phenotype Sampling (EPS) (response-dependent genotyping, selective genotyping), is a sampling design used for genetic association studies for continuous traits when the number of individuals that can be genotyped is restricted. Extreme phenotype individuals are chosen for genotyping under the assumption that the power to detect causal genetic variants is greater in an extreme phenotype sub-cohort compared to a randomly sampled subcohort. For example, if the aim is to investigate genetic association with waist circumference, the individuals in the highest and lowest quartiles of a population can be sampled for genotyping. In practice, case-control methods are often used for association testing; genotype frequencies are compared between the low and high extreme phenotype groups. The potential power gained by the extreme sampling can be lost under simplications such as treating a continuous outcome as a binary variable. We present valid statistical tests and models for the EPS design with a continuous outcome, focusing on association with common genetic variants. Statistical methods for two EPS designs are of relevance; first the phenotype and non-genetic information is obtained, along with genetic information, only for extreme phenotype individuals. Second, the genetic information is obtained for extreme individuals, but phenotype and covariate information is obtained for a larger sample. In the latter design, likelihood methods for missing at random covariate data are applied, assuming a parametric model for the genetic covariate. Through simulation studies and using real data, we show that tests for genetic association with a phenotype that are based on EPS data can have higher power than tests based on random samples.

4.
Atlantic salmon populations reveal adaptive divergence of immune related genes

Erik Kjærner-Semb, PhD student, Institute of Marine Research, Bergen

Populations of Atlantic salmon display highly significant genetic differences with unresolved molecular basis. These differences may result from separate postglacial colonization patterns, diversifying natural selection and adaptation, or a combination. Adaptation could be influenced or even facilitated by the recent whole genome duplication in the salmonid lineage which resulted in a partly tetraploid species with duplicated genes and regions.

In order to elucidate the genes and genomic regions underlying the genetic differences, we conducted a genome wide association study using whole genome resequencing data from eight populations from Northern and Southern Norway. From a total of ~4.5 million sequencing-derived SNPs, more than 10 % showed significant differentiation between populations from these two regions and ten selective sweeps on chromosomes 5, 10, 11, 13-15, 21, 24 and 25 were identified. These comprised 59 genes, of which 15 had one or more differentiated missense mutation. Our analysis showed that most sweeps have paralogous regions in the partially tetraploid genome, each lacking the high number of significant SNPs found in the sweeps. The most significant sweep was found on Chr 25 and carried several missense mutations in the antiviral mx genes, suggesting that these populations have experienced differing viral pressures. Interestingly the second most significant sweep, found on Chr 5, contains two genes involved in the NF-KB pathway (nkap and nkrf), which is also a known pathogen target that controls a large number of processes in animals.

The results of our GWAS study show that natural selection acting on immune related genes has contributed to genetic divergence between salmon populations in Norway. The differences between populations may have been facilitated by the plasticity of the salmon genome. The observed signatures of selection in duplicated genomic regions suggest that the recently duplicated genome has provided raw material for evolutionary adaptation.

5.
Genetics of insomnia; towards developing a genetic risk score

Daniela Bragantini, PhD student, Department of Neuroscience, Norwegian University of Science and Technology, Trondheim

Chronic insomnia is a significant health problem, affecting ca. 7% of the Norwegian population. In spite of its significance and costs both on society and individual level little is done about its genetic basis. The main aim of this project is to identify genetic risk factors for insomnia by investigating Single Nucleotides Polymorphisms (SNPs) in a series of sleep related genes and to develop a Genetic Risk Score (GRS) that can be used to improve diagnosis, treatment and prognosis of insomnia. In order to do so, we will select a sample of 3200 cases and controls from the HUNT3 study. We will assign subjects to either group according to their answer to sleep related questions in the HUNT questionnaire. We will genotype the subjects for ca. 150 common SNPs on genes relevant for sleep and neurotransmitter systems in the brain. Priority will be given to variations in genes related to modulation of circadian system in mammals, such CLOCK, TIMELESS, PER1,2, 3, CRY1 and CRY2, GSK3B, CSNK1D and CSNK1E. Tag SNPs in these genes has been found using the software Haploview and Tagger. Association analyses will be conducted using the software PLINK and R. The GRS will be calculated as the sum of the number of risk alleles across the associated SNPs, weighted for their effect size.

6.
Identification and classification of non-transcribed genomic features involved in gene regulation in salmonid species genomes

Teshome Mulugueta, PhD student, Department of Animal and Aquacultural Sciences, Norwegian University of Life Sciences, Ås

Background: Transcription factor (TF) motifs are important, integral parts in a functional genomics. Accordingly, they are important to annotate in order to gain a full understanding of the mechanisms underlying regulation and ultimately of the phenotype. Little effort has been made to identify and characterize TF motifs in salmonid species genomes. The Atlantic salmon genome has been released.

Understanding the regulation of gene expression is an important part of the next phase of its genomics research.

Methodology: A number of algorithms have been developed for in silico identification and characterization of cis-regulatory regions. We studied and tested various motif finding algorithms. Most of the algorithms studied reported overrepresented and conserved motifs. It is challenging to compare different motif finding tools; we selected tools based on their performance and information rich results. DNA sequences in promoter regions were extracted and scanned for identification and classification of non-transcribed and transcribed genomic features involved in gene regulation in Atlantic salmon.

Result: We have developed a database of DNA and RNA motifs and other related important genome information for salmonids. The database can be used for various purpose, including, but not limited to,: 1) for validation and characterization of transcriptional regulatory elements for carefully chosen biologically validated data sets, 2) testing specific hypotheses such as whether regulatory sub-functionalization between gene duplicates in the partially tetraploid salmon genomes are driven by sequence divergence of TF binding motifs, and 3) identifying and comparing evolutionarily conserved noncoding sequences among salmonid species.

7.
A Random Forest Approach to Translation Initiation Site Prediction Using 5' Ribo-Seq Read Length Patterns

Adam Giess, PhD student, Computational Biology Unit, Department of Informatics, University of Bergen

Ribosome profiling reads show distinct patterns of read length distributions around translation initiation sites. These patterns are typically lost in standard ribosome profiling analysis pipelines, when reads are shifted to determine which codon is represented by the footprint. We build a model that captures the unique signature around translation initiation sites and demonstrate its high accuracy in predicting start codons using N-terminal proteomics. Using this model we re-annotate the translation initiation landscape of Salmonella enterica serovar Typhimurium SL1344 and Escherichia coli str. K-12 substr providing evidence of truncations and elongations of annotated genes.

8.
Taming the beast

Diana Domanska, Post doc, Group for Biomedical Informatics, Department of Informatics, University of Oslo

Towards a systematic framework for the discovery, accurate annotation, curation and expression analysis of microRNAs and their isoforms in cancer. MiRNAs are key regulators of animal development and because they play critical roles in many human diseases such as cancer, miRNAs have been the most intensively studied molecules in the last decade. Much is known about miRNA evolution and their biogenesis - especially in vertebrates. However, this knowledge is currently not systematically applied in miRNA studies and accordingly many bioinformatics pipelines have very high false positive and false negative rates. Indeed there are literally thousands of published 'novel miRNAs' that do not fulfill well established annotation criteria, and thus about two third of the 'miRNAs' currently listed in miRBase are not bona fide miRNAs.

We argue that this rapid increase in noise not only fails to provide new insights in to miRNA evolution and functionality, it negatively affects the detection of signal in an already highly redundant and complex miRNAtarget regulatory network. We present a new set of bioinformatics tools that is focused on miRNAs: their prediction ('MirMiner'), expression analysis ('MirAthon'), curation ('MirGeneDB_2.0') and automatic validation of published miRNAs ('MIRror'). We compare prediction performance, false-discovery rates and isoform detection with other currently used tools and reanalyze a set of published data sets on human cancer samples. Finally we present the progress on MirGeneDB, including the curation of new species, new features and a better, more user-friendly interface.

9.

Functional Classification of Long Non-coding RNAs by Ribosome Profiling

Gunnar Schulze, PhD student, Computational Biology Unit, Department of Informatics, University of Bergen

Long non-coding RNAs (lncRNAs) present a class of non-coding transcripts that, despite lacking a canonical coding sequence (CDS), can exhibit features similar to coding genes such as alternative splicing and tissue-specific expression patterns. Since only 2% of the human genome is coding, while at least 80% is estimated to be transcribed, the characerization of (long) non-coding RNA species and their potential function is a major aim to (re)-evaluate which proportion of the genome can be considered functional. So far, lncRNAs have been implicated in a range of biological functions including transcription factor binding, chromosome looping, impriting, dosage compensation and as miRNA-precursors and versatile scaffolding molecules. Recent studies also suggest that a proportion of non-coding RNAs are translated and could potentially assume a coding or dual (coding/non-coding) role. Despite these findings however, the larger proportion of lncRNAs remains uncharacterized to date. In our current work we develop a computational pipeline to assess the coding potential of long non-coding RNAs based on genome-wide RNA-seq and ribosome profiling data. We identify novel open reading frames (ORFs) within lncRNA transcripts in a variety of human tissues and cell-types and develop a classifier to evaluate their coding potential as compared to canonical CDS. We anticipate that the classification of lncRNAs by their coding potential will aid in their functional characterization currently conducted by the FANTOM consortium as well as in assigning their role in tissue differentiation and disease.

10.
Deriving Rules for microRNA Regulation from Patterns in RNA Sequencing Data

Anna Tarsia, Master student, Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim

MicroRNAs (miRNAs) are single stranded non-coding RNA molecules, approximately 22 nucleotides (nts) in length, that can regulate gene expression in animals, plants and some viruses. This study explores the process of miRNA biogenesis with the aim to build a predictive model for the miRNA strand selection. During the miRNA biogenesis, the new miRNA strands are excised from longer double-stranded regions of RNA (precursor miRNA). A strand is chosen to join the silencing complex that will affect the production of proteins. The other one is degraded (passenger strand or miRNA* strand). The thermodynamic stability of the duplex, as well as other factors, appear to play an important role in this decision, but the mechanism behind the strand selection is still not fully understood. Recent studies have for example shown that for some miRNAs, both strands are included with equal probability. This suggests the existence of other mechanisms to control the selection of mature miRNAs.

There are recognized properties to detect functional miRNAs, like expression level and short reads. We hypothesize that the miRNAs most likely to be selected are those highly expressed and with a major number of short reads, and ultimately confirm this hypothesis. The miRNA strands with those properties will downregulate at least some of the predicted targets. This will result in a shift of the correlation distribution. Short reads can align to the miRNA sequences with different levels of accuracy (offset) and in different positions (start/end). Indeed, an additional method explores miRNAs with short reads positioned at the start of the sequences compared to the others.

11.
Trans-generational effects of gamma-radiation on the Expression small-ncRNA (miRNA, 5´tRNA-half) in early Zebrafish embryos.

Leonardo Martin, PhD student, Department of Basic Science and Aquatic Medicine, Norwegian University of Life Sciences, Oslo

To correlate gamma-radiation induced changes with gene expression and phenotypic effects. Zebrafish (~6 month old) were exposed to 10 mGy h-1 of γ-radiation (6 Gy total dose) during gametogenesis for 28 days. Exposed F0 fish were kept under standard conditions at NMBU-VetBio zebrafish facility and bred to generate F1, F2 and F3 embryos . Embryos were sampled at 5.5 hours post fertilization (late blastula / early gastrula) and RNA was isolated for ncRNA-seq (F1; Illumina platform, Novogene, China) and RT-qPCR profiling (F2 and F3). The dataset of ncRNA sequences (18-35nt), comprising an average of 10 million clean reads per library after adapter trimming (one library per RNA sample) with a quality score >30 (Phred score). The bioinformatic pipeline comprise (i) Trimgalore! for quality checking and adaptor trimming; (ii) mirDeep2 for miRNA expression analysis. (iii) Bowtie for mapping and alignment of other ncRNA such as piRNA and 5´tRNA-halfs. (iv) Seqmonk for visualization and counting of reads in mapping files and (v) EdgeR (Bioconductor) for differential expression analysis.

The further bionformatic analyses is on-going and results from this study will be presented and compared aiming at uncovering trans-generational effects from the IR exposure.

12.
The RNA structurome

Katarzyna Anna Chyzynska, PhD student, Computational Biology Unit, University of Bergen

An inherent property of mRNA is its formation of complex structures. Certain regions of transcripts are more prone to fold tightly than others, in order to hinder or facilitate translation. Recent advances in deep sequencing methods coupled with in vivo structural probing, such as SHAPE-Seq, provide new insights into transcriptome dimensionality – the so-called 'RNA structurome'. Based on such data from early zebrafish development we explore the structural profiles of transcripts and their fluctuations at different time points in development. Coupling the structural data with sequence information, RNA-Seq and ribosome profiling, we set out to decipher intricate connections between linear mRNA sequence, its secondary structure and translational dynamics.

13.
Biological network analysis

Zhaoran Zhou, PhD student, Computational Biology Unit, Department of Informatics, University of Bergen

Salmon lice (sea lice) are the major pathogens affecting the global salmon farming industry. Each year the sea lice causes multi-million dollar commercial losses to the industry world-wide, and they also affect the wild stocks of salmonids. Nowadays sea lice have developed resistance against widely used drugs. It is evident that understanding the growth, reproduction and drug resistance of sea lice is of crucial importance. Thus a fundamental question is to explain the gene functions of sea lice. In this study, we attempted to gain an insight into the gene function of sea lice by construction weighted coexpression network based on the RNA-seq data from 8 different life stages. The RNA-seq data are transformed using Voom from R pacakge limma, which opens access for RNA-seq analysts to a large body of methodology developed for microarrays. Genes with little expression variations across different stages are filtered.

We utilized an R package known as Weighted Gene Coexpression Network Analysis (WGCNA) as a main tool to construct co-expression network in measure of the correlation between gene expressions. We further identify different gene sets (modules) which are highly correlated with each life stage. To measure the topological properties among modules, we made a network using modules as nodes and calculated the intermodular adjacencies. The hypergeometric test is performed for each module to find the significant GO annotation and modules network was visualized with Cytoscape. Using the known genes involved in Chitin pathway, we analyzed each modules, and we identified some genes which might be crucial in the chitin pathway although they have very few annotations now. Collectively, our methods may imply a potential role of a set of genes in the chitin pathway as well as biological growth and developments of sea lice.

14.
Use of resampling to improve accuracy of data analysis in time course experiments

Mathias Bockwoldt, PhD student, Department of Arctic and Marine Biology, University of Tromsø - The Arctic University of Norway

Time series of expression data in higher eukaryotes are costly and only a limited number of data points can be measured as the animals have to be sacrificed at every data point. Thus, often only few replicates are available. As the data points furthermore stem from different animals, a time series of expression data does not comprise true biological replicates, but rather a random assembly of time points from a larger number of biological replicates. This high variability leads to very noisy datasets and makes the identification of rhythmic transcripts difficult and error-prone. To improve the identification of circadian regulated genes while maintaining the minimal required sample number, we used data resampling. In tests with artificial time course series, the number of false positives could be greatly reduced while maintaining the number of true positives. These findings were confirmed with several common algorithms to identify rhythmic data series: ARSER, Haystack, JTK Cycle, and Biodare FFT-NLLS. Using the resampling approach, we were able to increases the accuracy of circadian expression data analysis and showed the importance of replicates for time course data analysis. Using our approach 3-4 replicate time series are sufficient to achieve a high accuracy in the detection of oscillating transcripts. We furthermore point out that averaged datasets that are often generated in circadian datasets, are not well suited for expression data analysis. The results are, however, largely depending on the algorithm used. Thus different algorithms should be considered depending on how datasets have been generated.

15.
Temporal meta-omic analyses of a syntrophic biogas-producing community

Benoît J. Kunath, PhD student, Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås

Biogas reactors are a source of renewable energy, which can be used for heat or as transportation fuel. The anaerobic digestion of biomass to methane requires a complex microbiological process combining the metabolic capacities of several physiological groups of microorganisms that can hydrolyze organic substrates, ferment sugars, oxidize short-chain fatty-acids and convert fermentation products to methane. To deal with anoxic and energetically limited conditions, the microorganisms must have close interactions and reach efficient metabolic coordination, also called syntrophy. Understanding how microbes within a community interact and cooperate to convert biomass to biogas is still limited, some reasons being the high species complexity within environmental ecosystems and the unculturability of most of them.

Here we use culture dependent and independent techniques to describe the functioning and interactions of different phylotypes within a minimalistic biogas-producing microbiome. Enrichment techniques produced a functional community that is composed of only six species and is able to comprehensively utilize a broad spectrum of lignocellulosic substrates. Metagenomic analysis combining the PacBio CCS long-read sequencing and the Illumina HiSeq sequencing technologies enabled the partial reconstruction of each species as well as the identification of a strain-level complexity (up to 6 additional strains). Quantitative metaproteomics identified ~4000 proteins per time point over the lifetime of the community, and these proteins were matched against the reconstructed genomes. Integration of these temporal analyses with corresponding metadata revealed complementary saccharolytic enzymes and mechanisms, as well as syntrophic phylotypes centered on key metabolites. This relatively simple community can be used to unravel the key relationships at the strain-level and to contribute to worldwide efforts to understand microbial lignocellulose conversion and biomass production.

16.
Metagenomics analysis of Bergen COPD microbiome pilot study

Yaxin Xue, PhD student, Computational Biology Unit, Department of Informatics, University of Bergen

The course of chronic obstructive pulmonary disease (COPD) are intricately linked to the microbiomes. However, we still have limited knowledge to the complex relationship. Bergen COPD microbiome study is a longitudinal study aiming to investigate the hypothesis that composition of airway and lung microbiomes is the missing link between risk factor exposure and disease development. To date we collected 168 samples and then performed a pilot study. We evaluated the impact of laboratory and reagent contamination, compared different bronchoscopic sampling techniques, and investigated the microbiome diversity within/between different sampling positions and disease groups.

Our results indicated that possible contaminant had an effort on taxonomic distribution, revealed that contaminations should be considered and processed before downstream analysis. Meanwhile, three sampling techniques: protected bronchoalveolar lavage (PBAL), protected sterile brushes (PSB), small volume lavage (SVL), were evaluated; PBAL was more informative than the others. Samples obtained from seven different positions of each participant were sequenced: oral wash (ow), first part of PBAL from right middle lobe (svho1), second part of PBAL from right middle lobe (svho2), PSB from left upper lobe (bove), PSB from right lower lobe (boho), SVL from left upper lobe (svve), and negative control (ktrl). Diversity analysis were performed and compared, results of svho1 showed an interesting correlation between microbiome communities with disease development.

17.
Flexibility effect on protein functionally relevant channels. From protein channels dynamical characteristics to prediction

Pierre Bedoucha, PhD student, Computational Biology Unit, Department of Molecular Biology, University of Bergen

Proteins 3D structures are tightly related to protein functions. There is withal a missing link between protein structure and function and studies have shown that protein dynamics stand tying the two.

As a token of protein 3D structures, channels in proteins are tightly related to function. Cell membrane transport proteins utilize their channels to ensure small substrates through the cell membrane to carry out a specific physiological goal.

We aim at investigating the repercussion of protein intrinsic dynamics on channels' shape and size via the analysis and visualization solution CAVER. We will develop a reliable method with coarse-grained elastic network models to model channels' flexibility. Subsequently, we will implement a tool rendering our model available to other users. At last we will demonstrate the effect of flexibility on channels in a dataset of transmembrane proteins.

Name	Position	Institution	Department	Abstract
Adam Giess	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	7
Anna Tarsia	Master student	Norwegian University of Science and Technology	Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine	10
Antonin Klima	NORBIS board PhD student	Norwegian University of Science and Technology	Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine	G
Benoit Kunath	PhD student	Norwegian University of Life Sciences	Department of Chemistry, Biotechnology and Food Science	15
Charitra Kumar Mishra	Researcher	University of Bergen	Computational Biology Unit, Department of Informatics	
Chloe Steen	PhD student	University of Oslo	Group for Biomedical Informatics, Department of Informatics	B
Christine Stansberg	NORBIS coordinator	University of Bergen	Computational Biology Unit, Department of Informatics	
Daniela Bragantini	PhD student	Norwegian University of Science and Technology	Department of Neuroscience	5
Diana Domanska	Post doc	University of Oslo	Group for Biomedical Informatics, Department of Informatics	8
Endre Bakken Stovner	PhD student	Norwegian University of Science and Technology	Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine	F
Erik Kjærner-Semb	PhD student	Institute of Marine Research	Group for Reproduction and Development Biology	4
Gos Micklem	NORBIS SAB Professor	Cambridge University	Department of Genetics	
Gunnar Schulze	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	9
HIlde Vinje	PhD student	Norwegian University of Life Sciences	Department of Chemistry, Biotechnology and Food Science	I

Ilona Urbarova	PhD student	University of Tromsø, The Arctic University	Department of Medical Biology	H
Ines Heiland	Professor	University of Tromsø, The Arctic University	Department of Arctic and Marine Biology	
Inge Jonassen	NORBIS director Professor	University of Bergen	Computational Biology Unit, Department of Informatics	
Jan Terje Kvaløy	NORBIS board Professor	University of Stavanger	Department of Mathematics and Natural Sciences	
Jorke Kamstra	PhD student	Norwegian University of Life Sciences	Department of Basic Science and Aquatic Medicine	D
Katarzyna Anna Chyzynska	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	12
Kornel Labun	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	C
Kyrre Lekve	NORBIS SAB Deputy Managing Director	Simula Research Laboratory		
Leonardo Martin	PhD student	Norwegian University of Life Sciences	Department of Basic Science and Aquatic Medicine	11
Manuela Zucknick	NORBIS co-director Associate Professor	University of Oslo	Department of Biostatistics	
Martin Jakt	NORBIS board Researcher	Nord University	Faculty of Biosciences and Aquaculture	
Martin Lercher	Professor	Heinrich-Heine-Universität Düsseldorf	Department of Computer Sciences	
Mathias Bockwoldt	NORBIS board PhD student	University of Tromsø, The Arctic University	Department of Arctic and Marine Biology	14
Mette Langaas	NORBIS co-chair Professor	Norwegian University of Science and Technology	Department of Mathematical Sciences	
Michael Dondrup	Researcher	University of Bergen	Computational Biology Unit, Department of Informatics	
Niladri Banerjee	PhD student	University of Bergen	Department of Clinical Sciences	E

Ole Christian Lingjærde	NORBIS chair Professor	University of Oslo	Group for Biomedical Informatics, Department of Informatics	
Phil Pope	Researcher	Norwegian University of Life Sciences	Department of Chemistry, Biotechnology and Food Science	
Pierre Bedoucha	PhD student	University of Bergen	Computational Biology Unit, Department of Molecular Biology	17
Pouda Panahandeh	PhD student	University of Bergen	Department of Biomedicine	A
Pål Sætrom	Professor	Norwegian University of Science and Technology	Department of Computer and Information Science and Department of Cancer Research and Molecular Medicine	
Rolv Bræk	Professor	Norwegian University of Science and Technology	Department of Telematics	
Stefania Salvatore	Post doc	University of Oslo	Group for Biomedical Informatics, Department of Informatics	1
Teshome Mulugueta	PhD student	Norwegian University of Life Sciences	Department of Animal and Aquacultural Sciences	6
Thea Bjørnland	PhD student	Norwegian University of Science and Technology	Department of Mathematical Sciences	3
Torgeir R. Hvidsten	NORBIS board Professor	Norwegian University of Life Sciences	Department of Chemistry, Biotechnology and Food Science	
Xiaokang Zhang	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	
Xiaozheng Liu	Master student	University of Bergen	Department of Biomedicine	2
Yaxin Xue	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	16
Zhaoran Zhou	PhD student	University of Bergen	Computational Biology Unit, Department of Informatics	13
Ørnulf Borgan	NORBIS SAB Professor	University of Oslo	Department of Mathematics	

NORBIS

contact-norbis@uib.no
norbis.no