



**BiB | Bioinformatics in Bergen**

**2017 2nd – 3rd October 2017**

**ABSTRACT BOOK**

**Ribosome Signatures Aid Bacterial Translation Initiation Site Identification**

*Adam Giess, Computational Biology Unit*

While methods for annotation of genes are increasingly reliable, the exact identification of the translation initiation site remains a challenging problem.

Since the N-termini of proteins often contain regulatory and targeting information developing a robust method for start site identification is crucial. 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 from footprints are adjusted to determine the specific codon being translated. Utilising these signatures in combination with nucleotide sequence information, we build a model capable of predicting translation initiation sites and demonstrate its high accuracy using N-terminal proteomics. Applying this to prokaryotic translatomes, we re-annotate translation initiation sites and provide evidence of N-terminal truncations and extensions of previously annotated coding sequences. These re-annotations are supported by the presence of Shine-Dalgarno sequences, structural and sequence based features next to N-terminal peptide evidence. Finally, our model identifies 61 novel genes previously undiscovered in the Salmonella enterica genome.

**Can DNA methylation in neonatal blood inform about mental disorders risk in adulthood?**

*Alfonso Buil Demur, Region H, Denmark*

Emerging evidence implicates altered DNA methylation in mental illness including autism, ADHD, bipolar disorder, major depressive disorder, anorexia and schizophrenia. However, it is unclear whether the DNA methylation changes observed to date are causative or reflect disease progression or treatment. The neonatal period is a time of rapid neurodevelopment during which alterations in DNA methylation may contribute to the risk of mental illness later in life. Hence, we explored whether differences in DNA methylation in neonatal blood taken at birth were associated with twin discordance for mental illnesses including autism, ADHD, affective disorder, anorexia, schizophrenia or bipolar disorder. A total of 597 pairs of twins (220 monozygotic) discordant for mental illness born between 1981 and 2005 were identified for methylomic comparison. Blood samples obtained from neonatal Guthrie cards were used for DNA extraction and genome-wide profiling of DNA methylation with the use of Infinium HumanMethylation450 BeadChip or EPIC array from Illumina. Quality control, data pre-processing and statistical analysis was performed using the minfi package in R. Data were normalized using the ssNoob method and adjusted for batch effects using the Combat tool. Blood cell composition was estimated using FlowSorted.CordBlood.450k. Using linear regression models and including potential confounders such as sex, blood cell composition and zygosity, we observed differentially methylated positions (DMPs) associated with mental illness. We observed significant DMPs (P<10-06) for ADHD (mapping to TET2, HNRNPH2, HMGN5), autism (mapping to ATP1B4), and anorexia (ITGB4, GJA3, NXN). Interestingly, there is an enrichment of DMPs mapping to genes in the dopaminergic and serotonergic synapse KEGG pathways including KCNJ5, PRKCA, CACNA1D, CREB5, and ALOX12 (P<0.05). Our data indicate that DNA methylation differences are quantifiable in neonatal blood from twins discordant for mental illness later in life and suggest that susceptibility to mental illness is conferred by dysregulated neurodevelopmental genes.

**Identification of molecular mechanisms following cannabis exposure in patients with psychosis**

*Anne-Kristin Stavrum, Department of Clinical Science, UiB*

Cannabis is the most widely used illegal drug worldwide. Some people use it recreationally, while others use it as a mean of self medication against diseases that induce pain, or for relief of symptoms related to for example anxiety, stress, epilepsy and ADD/ADHD. However, cannabis may also induce psychosis. The molecular mechanisms by which cannabis could induce psychosis are not well understood. There are strong indications that it causes DNA methylation changes, although no-one has specifically tested this at the genome wide level. We have profiled methylation patterns in blood samples from different people where some have been smoking cannabis. We have also transcriptomics information for the same individuals to characterise the effects of cannabis smoking on gene expression.

We are now in the process of comparing the methylation patterns and gene expression of people that have smoked cannabis and people that have not. We are interested in identifying areas of the genome that have a changed expression and/or methylation pattern induced by cannabis. Our aim is to characterise the effect of cannabis at the molecular level and complement that with modelling in stem cells to understand the causality (and other mechanisms) linking cannabis use to the development of psychosis.

**Using sequence similarity networks for analyses and classification of N-terminal acetyltransferses**

*Bojan Krtenic, MBI*

The GNAT fold-containing N-terminal acetyltransferases (NATs) are members of a big and diverse, but poorly characterized group of enzymes. NATs catalyze the reaction of transferring an acetyl group from the cofactor Ac-CoA to the N-terminus of a substrate protein (N-terminal acetylation). This is one of the most common modifications in eukaryotic cells.

Despite the high structural conservation of the GNAT fold, universally shared among NATs, there are at least seven types of NATs described so far that differ in substrate specificity, catalytic strategies and cellular localization. It is essential to analyze the variation in catalytic mechanisms employed by the NATs in pursuance of understanding their functional diversity and predicting the specificity of uncharacterized members of the group. We use sequence similarity networks to analyse large datasets of NAT sequences in order to better characterize and understand the sequence-structure-function relationship of this diverse group of enzymes. Our work lays the groundwork for a more detailed classification of all NATs, thus aiding in discovering and annotating novel N-terminal acetyltransferases

**What do pathway databases contain and what is the consequence for pathway analysis?**

*Bram Burger, Computational Biology Unit (Computational Biology Unit), Department of Informatics, University of Bergen, Norway.*

In biomedical studies it is common practice to perform pathway analysis when a set of biological compounds of interest --- genes, proteins, metabolites, etc. ---

has been identified in order to gain insight into the disease of interest. Pathway analysis tools however rely on an accurate model of the underlying biological mechanisms, which is essential for the development of accurate and robust methods,

and is of great value for the interpretation of the resulting output. Currently only around half of the human proteins are annotated in the manually curated Reactome pathway database (reactome.org), where proteins that are easier to

detect and well-studied are more likely to be included. This project aims to help researchers make informed decisions when performing pathway analysis, and when interpreting the output. By using network based approaches, the current state of our knowledge regarding the proteins involved in biological pathways is described using the human complement

of the Reactome pathway database. The implications of the findings with regards to pathway analysis is also discussed.

**A year in the life of the minutiae in a changing Arctic Ocean**

*Bry Wilson, Universitetet i Bergen*

As the global climate changes, the higher latitudes are seen to be warming significantly faster and it is likely - if not already apparent - that the Arctic biome will experience considerable shifts in ice melt season length and permafrost thawing, leading to changes in photoirradiance and in the freshwater and terrigenous inputs to the marine environment.

The exchange of nutrients between Arctic surface and deep waters and their biogeochemical cycling throughout the water column is driven by the seasonality of some of the most extreme environmental changes on the planet. The impacts, however, of the current global climate transition period on the biodiversity and its continued nutrient cycling within the Arctic Ocean are not yet known. To determine seasonal variation in the microbial flora and fauna of the deep water column, samples were collected from a 1000m depth profile in the seas around the Western coasts of the Svalbard archipelago throughout the polar year. High-throughput sequencing of tag amplicon and shotgun metagenomes were used to monitor microbial diversity and function in both the epipelagic surface waters (defined by the diametric diurnal conditions of the polar summer and winter) and the relatively invariable and permanently dark mesopelagic depths. In epipelagic surface waters (<200m depth), seasonal diversity varied significantly, with light and the corresponding annual phytoplankton bloom pattern being the primary drivers of change during the late spring and summer months.

In the mesopelagic ocean deeps (>200m), seasonality subsequently had much less effect on biodiversity. Interestingly, species richness consistently increased down through the water column, with the deepest darkest waters containing the greatest diversity. The phenomenon of the polar phytoplankton blooms, followed by the successional explosion of heterotrophic bacterial populations, also seemingly spurs the annual disappearance of the ancient and chemolithoautotrophic marine Archaea from surface waters. During the winter darkness, these venerable microbes feast upon the summer fruits of the phytoplankton photosynthesis, replenishing surface waters with nutrients which fuel the next spring bloom, essentially yin to the phytoplankton's yang. However, should suggested models of a freshening Arctic be correct, surface Arctic basin waters in a warming world may become increasingly stratified,

such that the vertical flux of nutrients between deeper waters and the epipelagic zone may be much reduced primary productivity would consequently be lessened and this annual biogeochemical cycle, so essential for Arctic Ocean productivity, would inevitably be disrupted.

Given the significance of the annual phytoplankton bloom pattern on biodiversity in Arctic waters, any changes to bloom dynamics resulting from accelerated global warming will likely have major impacts on surface marine microbial communities, those impacts inevitably trickling down into deeper waters...

**Parent-of-origin effects in children with attention deficit hyperactivity disorder.**

*Dinka Smajlagic, Department of Clinical Science*

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common and most heritable childhood-onset neuropsychiatric disorders, characterized by multifaceted genetics.

To date, genetic studies of ADHD focused on additive effects only, explaining just a fraction of its heritability. Thus, we aimed at examining parent of origin effects (POE) together with maternal and additive effects, providing novel insight into the complex genetic architecture of ADHD. We compiled parent-offspring data collected through the Psychiatric Genomics Consortium and the Norwegian Mother and Child Cohort, consisting of 2060 trios and 328 duos.

Additional parent-offspring data is being added. ADHD was diagnosed based on DSM-IV, ICD-10 and child behavior checklist. POE, maternal and additive genetic effects are being evaluated using multinomial modelling implemented in EMIM software. We explored our signals in the light of known imprinted genes (POE) and the largest ADHD genome-wide association study (GWAS) in children (N=17666). Gene based analyses are being performed using MAGMA software. Heritability estimates and genetic correlations of the examined effects are being calculated using LD score regression. Our preliminary results indicate the presence of non-additive genetic effects in the development of ADHD. Our preliminary strongest imprinting signal is located in CALD1 locus (rs11980823, effect=0.77, SE=0.14, P=1.21E-07).

This gene also revealed strong association signal in the previously reported large-scale childhood ADHD GWAS (rs79846815, P=2.03E-06). CALD1 plays essential role in nerve regeneration, a function previously implicated in a number of neuropsychiatric disorders. Our preliminary gene-based analyses of the known imprinted genes revealed strong association with TP73 locus (P=0.0034), encoding a transcription factor implicated in disorders of nervous system (e.g. neuroblastoma). Additional hits were noted in the non-coding RNA genes, adding to the recent observations in neuropsychiatric genetics of gene regulation playing a pivotal role in the development of disorders of mental health. In conclusion, this is the first and the largest genome-wide parent-offspring study in ADHD, exploring its non-additive genetic effects by detecting and distinguishing between POE (imprinting), maternal and child effects.

As we increase our sample size, we will provide estimates of such effects as well as those of their heritability and genetic correlations.

**Towards an integrative local network analysis of omics data**

*Eileen-Marie Hanna, Computational Biology Unit, UiB*

In view of the availability of large-scale biological data, of diverse types, the project aims at performing local network analysis of experiments through multiple-source data integration.

Hypothetically, combining data from different databases, would lead to a more comprehensive view of interconnections among genes/proteins in a network. In this context, the output of local analysis will be significantly differentially expressed star-shape modules corresponding to direct perturbations of central nodes and their neighbors.

**How to science in 3 simple steps**

*Eivind Valen, Computational Biology Unit*



**Gene-specific correlation of mRNA and protein levels in colorectal cancer cell lines**

*Fatemeh Zamanzad Ghavidel, UIB*

The central dogma of molecular biology describes the translation of genetic information from mRNA to protein. However, comparative reports have shown that the correlation between mRNAs and protein levels can be relatively weak or moderately positive and they vary between experiments and organisms. Several biological factors, e.g., post-translational modifications were identified which influence the correlation. Therefore, considering only gene's mRNA level might be far from perfect in predicting its protein expression level. In this work, we will examine a comprehensive correlation study between the protein profiles (9000 protein measurements) of 50 colorectal cancer cell lines and their corresponding gene expression levels from two public data bases: Genetech and Cancer Cell Line Encyclopedia (CCLE). Further more we will extend our initial analysis to explore predicting protein levels from extra features of mRNA and genetic data.

**Assessing the coding potential of RNAs outside annotated gene regions**

*Gunnar Schulze, 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.

In addition, recent studies suggest that a proportion of non-coding RNAs are translated or harbour small micropeptides and could potentially assume a coding or dual (coding/non-coding) role. Despite these findings however, to date a large proportion of RNAs still remain unexplored regarding this hidden potential. In this 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.

**How aromatic amino acids aid in peripheral protein membrane binding**

*Hanif M. Khan, MBI, Computational Biology Unit*

Peripheral proteins are membrane proteins those bind membrane reversibly and transiently. Binding of peripheral proteins to the cellular membrane is important for different cellular events, for example, cell signaling. Modulations of such cellular events need detailed understanding of the mechanism.

Hence, elucidating atomic detail of such binding mechanism is necessary where molecular modeling techniques can be useful. My work is focused on obtaining a generalized description of the binding mechanism of peripheral proteins using molecular modeling. One of the important goals along the way is to differentiate the role of aromatic amino acids in such binding processes. To this extent, I am also involved in methodological advancements for accurate descriptions of such interactions in different time and length scales.

**The RNA structurome**

*Katarzyna Anna, UiB*

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.

**Computational method for deconvolving transcriptome data from heterogeneous samples**

*Konstantina Dimitrakopoulou, Computational Biology Unit (Computational Biology Unit) & Centre for Cancer Biomarkers (CCBIO)*

Gene expression data are often generated from human or experimental animal samples that are heterogeneous consisting of multiple cell/tissue types.

However, the heterogeneity and the variability in cell type composition affect deeply standard transcriptome analysis tools from capturing the real differences between samples coming from different physiological states and the development of disease predictive models. To address this, we developed a computational method for estimating the number of participating cell/tissue types,

the mixture proportions for each sample and the cell/tissue type-specific gene expression profiles directly from the mixture data. We evaluated the performance of our method against other current approaches based on benchmark and cancer expression datasets.

**Identification, quantification and visualization of genuine mutations induced with genome editing techniques**

*Kornel Labun, UiB*

With the exceptional volume of research being generated using CRISPR, methods for robust validation of generated mutants are urgently needed. Resource-efficient verification of experiments is often carried out through PCR followed by sequencing, but not all studies using this approach follow good practices of experimental design.

Unfortunately, misidentification of natural variation between reference genome and the organism of interest often lead to false positives. Moreover, quantification of genuine variants carries many subtle complications e.g. read merging, primer-dimer filtering and cut site restriction. Here we show that use of controls accounts for organism specific differences and argue for the strict removal in the case group of all position specific mutations identified in the control group, instead of adjusting frequency levels. We optimize the quantification step and presentation of deletions and insertions in an informative, yet transparent way. Furthermore, we provide an R package amplican that scales to genome-wide screens, performs alignment of the amplicon reads, normalizes gathered data, calculates multiple statistics (e.g. cut rates, frameshifts) and presents results in form of aggregated reports. Data and statistics can be broken down by experiments, barcodes, user defined groups, guides and amplicons allowing for quick identification of potential problems. These advances can minimize inaccurately interpreted genome editing experiments and allow biological scientists to effectively analyze the sequencing data for themselves.

**Copy number variation calling in trios**

*Ksenia Lavrichenko, Computational Biology Unit*

We discuss the details and culprits of genomic copy number variation (CNV) calling and quality control in the Norwegian MoBa cohort of trios (father, mother, child).

The calling has been done with PennCNV, an established algorithm for predicting CNVs from microarray data. One HapMap sample and a number of duplicate samples between the two batches allow for rough estimates of sensitivity and specificity.

**Do changes in the immunoscore reflect biologic treatment effect of novel agents in treatment naive high-grade serous ovarian carcinomas? IMPACT**

*Liv Cecilie V. Thomsen, CCBIO, University of Bergen*

Authors: Liv Cecilie V. Thomsen a,b, Shamundeeswari Anandan a,b , Katrin Kleinmanns a, Sonia Gavasso a,c, Jorn Skavland a, Geir Bredholt a, Emmet Mc Cormac a, Bjorn Tore Gjertsen a,c and Line Bjorge a,b

a Centre for Cancer Biomarkers CCBIO, Department of Clinical Science, University of Bergen

b Department of Gynecology and Obstetrics, Haukeland University Hospital

c Department of Neurology, Haukeland University Hospital

d Hematology Section, Department of Internal Medicine, Haukeland University Hospital

Introduction

High-grade serous ovarian carcinomas (HGSOC) is the most common (>70%) and lethal subtype of epithelial ovarian cancer. In HGSOC degree of tumor reduction during primary surgery constitutes one treatment cornerstone, with improved disease outcomes if all visible intraabdominal tumor tissue is removed. To avoid significant amount of residual disease postoperatively, women with advanced intraabdominal disease are offered diagnostic laparoscopic evaluation of the tumor operability as well as histologic diagnosis verification. Result of the evaluation is either a) debulking surgery after two weeks where complete tumor removal is feasible, followed by chemotherapy, or, where appropriate debulking cannot be achieved, b) neoadjuvant chemotherapy followed by surgery. Currently personalized medicine is in increasing demand, requiring improved identification of which patients who will benefit from specific treatment options. Unfortunately a well defined pathophysiology of HGSOC is still lacking, and implementation of personalized medicine therefore not feasible. Immunoscore is a relatively new concept based on immune reactions in the tumor microenvironment and is increasingly utilized to describe the interplay between tumor and host immune system. Most clinical cancer-related studies investigating effect of medications are performed on patients after they have received at least one course of chemotherapy. How the treatment impacts tumor biology and the immune system in untreated patients has rarely been examined. The aim of this study is to identify how selected agents affect the immunoscore in treatment naive patients with HGSOC.

Methods

In the current Window-of-Opportunity study the immunoscore will be examined in a prospectively enrolled cohort of 32 women who proceed to have debulking surgery after laparoscopy. Starting the day of laparoscopic surgery, these participants will be treated with selected agents for 10-14 days. The treatment will be discontinued at the time of tumor reductive surgery. During both surgical procedures tissue samples from matching sites will be obtained. Blood, saliva and urine will be sampled at inclusion, after laparoscopy and before the main operation. Cytometry by Time of Flight (CyTOF) is a method that enables researchers to perform single cell proteomics by combining atomic mass spectrometry and reporter elements. We plan to compare the composition of immunological cell types in the microenvironment of the tumor and the host at time of diagnostic laparoscopy using a study-specific CyTOF panel of 38 antibodies representing the main classes of immune cells including T- and B-leukocytes, neutrophils, lymphocytes, macrophages, as well as subtypes of T-lymphocytes and tumor cell-specific markers. The high-dimensional data generated through the biologic examination will be analyzed using software in Cytobank and R. The focus will be on distinguishing populations present in the samples and the relative distribution of these, comparing samples taken before and after intake of the study medicine. Further, as revelation of phenotypically related cell subsets would be of great value, attempts to identify such will be made.

Discussion

Currently we lack both specific identification of which women with HGSOC that will benefit from a particular treatment agent, as well as knowledge on how medicines affect the immunoscore in a primary setting. The result is suboptimal management of the disease in many women, unnecessary side effects for those without treatment effect, and faulty use of hospital resources. Through this study we hope to detect changes in the immunoscore that correlates with effects of the administered medications. Such findings could help establish novel regimes for patients with HGSOC in a primary treatment setting.

**Refining Pathway Search using the Reactome Knowledgebase**

*Luis Fransisco Hernandez Sanchez, KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen*

From the 20,214 reviewed human proteins in UniProt, 10,844 are annotated in Reactome, which are participants of 9,297 biochemical reactions, composing 2,134 pathways.

Reactome also anno-tates possible PTM sets of a protein, for specific processes or at specific subcellular locations. On average, a protein in Reactome is participant of 6.96 reactions and 9.16 pathways, and has 2.28 different PTM sets known. However, if we consider a protein with a specific PTM set to be a dis-tinct entity, then each such entity only participates in 5.97 reactions and 8.32 pathways. Therefore, when only one possible set of modifications is mapped, the ambiguity in resulting pathways and reactions is reduced, improving search results to be better representative of the molecular pro-cesses involved. This allows the search for relevant pathways to be more specific and reduce the time analyzing the search results.

**Bergen Open Source -- an informal meeting space for local software devs and open source enthusiasts**

*Matus Kalas, Computational Biology Unit (Computational Biology Unit)*

Bergen Open Source [1] is an initiative for exchanging experiences and inspiration between software developers in and around Bergen.

The term Open Source does not relate only to open-source software - with respect to which we aren't purists - but even more to the open fashion of sharing one's tricks and failures, and learning from the others'. Bergen Open Source started in the previous school-year of 2016-2017, with monthly meetings in form of informal workshops or micro-hackathons, including short lectures, demos, and discussions.

These meetings are hosted by Computational Biology Unit, the Computational Biology Unit at UiB, which has traditionally been a productive community of open-source software, and is thus a very adequate host. Reciprocally, the Bergen Open Source initiative aims at helping bioinformaticians overcoming obstacles while creating succesful software and impactful results. Not limited to bioinformatics, Bergen Open Source invites both software devs and enthusiasts throughout the domains of science, society, art, technology, and leisure.

Beyond software, we exchange also knowledge about sharing and opening other intellectual works and notions, such as scientific results, organisation of work, or reputation and community building. Sharing is caring [1] http://groups.google.com/group/bergenopensource

**Using normal modes analysis to characterize the flexibility of protein tunnels and channels**

*Pierre Bedoucha, Computational Biology Unit - UiB*

Protein 3D structures are tightly related to protein functions. There are withal missing links between protein

structure and function and studies have shown that protein dynamics is one of them. The transport of compounds of various sizes through cell membranes is partly ensured by transmembrane proteins, such as channels or carriers.

Their 3D structure is often characterized by an opening providing a route through which ions or small molecules will be traveling. During the transport process these tunnels may change geometry, adjusting their access and permeability to regulate the protein's function. These structural changes are happening on time scales that are often too long to be captured by molecular dynamics simulations. We propose to use Normal Mode Analysis (NMA) to investigate the repercussion of protein intrinsic dynamics on the properties of tunnels in proteins 3D structure.

NMA has indeed been repeatedly shown to be an efficient and reliable method to describe slow and large amplitude movements in proteins ( Fuglebakk et al., BBA, 2015). We combine NMA with the use of CAVER (Chovancova et al., PCBI, 2012), a tool for the analysis and visualization of protein tunnels and cavities. We have developed a computational framework linking CAVER and NMA results. We have validated it on a few proteins for which we have compared the use of coarse-grained (CG) and all atoms representations to model channel flexibility.

The use of CG representations implies residue side-chain reconstruction as cavities are highly sensitive to side chain positions so we have also introduced this in our pipeline. We confirmed that NMA is a fitting mean to investigate tunnel and channel plasticity and thus can be used as an input for investigation tools such as CAVER.

**Computational investigation of the membrane association of N-terminal acetyltransferase 60 (Naa60).**

*Qaiser Waheed, Department of molecular biology, computational biology unit*

N-terminal acetyltransferases (NATs) are members of GCN5-related N- acetyltransferase (GNAT) superfamily, known for playing catalytic role in N- terminal acetylation of the proteins.

Recent experimental studies have shown that Naa60 (NatF) is the only candidate from the N-terminal acetyltransferase family, which is found to have Golgi membrane association (Asknes et.al, 2104). This study confirmed that only the C-terminal part of NatF (Naa60182-242) is important and sufficient for the membrane association. As the resolved structure is missing the C-terminal part (Naa60185-242), we modeled this part based on the secondary structure predictions (PSIPRED and JPRED), which give two alpha helices in this region. These helices were analyzed by helical wheel and found to contain amphiphatic characteristic by containing a hydrophobic face on one side and a hydrophilic face on the other side. The modeled structure was then used to investigate the membrane association by molecular dynamics simulations. Results of an implicit membrane model (IMM1) simulations show that Naa60185-242 associate with the membrane at the interfacial position and both helices keep their predicted hydrophobic face partially anchored in the hydrophobic core of the membrane. The energy contribution from the anchored amino acids is favorable. To further investigate the detailed interactions we performed atomistic simulations using the Charmm force field. Naa60185-242 was docked on three membrane compositions, two anionic (POPC:PI4P 92:8, POPC:POPS 75:25) and one neutral POPC:PSM:CHOL 70:20:10. Results for all three compositions will be discussed in terms of density distribution and detailed interactions showing a favorable association for anionic membranes. In vitro assays and cellular localization experiments verified that Naa60 is a peripheral membrane protein.

**Allelic heterogeneity across psychotic disorders and related phenotypes**

*Tatiana Polushina, Department of clinical science*

Major mental illnesses have been shown to overlap at the clinical and genetic levels. The genetic overlaps have been so far explored at the single genetic variants level, but very few studies have explored how independent variants within a locus could contribute to the genetic overlaps.

In our study, we intend to increase the information captured from GWASs by focusing on allelic heterogeneity, i.e. the contribution of several independent markers within one genetic locus, within a trait and across related traits. We observed allelic heterogeneity within and across traits. 147 genomic regions were associated with independent markers (not in LD) across several traits.

We have established a map of genetic overlaps for these clusters across psychiatric disorders and relevant phenotypes (brain volumes, cognitive and personality traits). The strongest overlaps were observed in pairs: schizophrenia - educational attainment and schizophrenia - bipolar disorder.We have established a pipeline for identification of allelic heterogeneity across different phenotypes. Several of the GWAS included were too limited in power to provide significant hits yet,

and will need bigger samples to yield more significant results. Our approach is complementary to studies that explore genetic overlap at the single marker level.

This improves our understanding of the impact of genetic factors in main psychotic disorders and related phenotypes, and could help to direct functional studies later.

**RareVariantVis2: a suite for analysis of rare genomic variants in whole genome sequencing data**

*Tomasz Stokowy, UiB*

Second version of RareVariantVis package aims to provide comprehensive information about rare variants for your genome data. It annotates, filters and presents genomic variants (especially rare ones) in a global, per chromosome way. Large structural variants, including copy number variants are also supported. Package accepts variants directly from variant caller - for example GATK or Speedseq. Output of package are lists of rare variants, together with adequate visualization and annotation. Visualization of variants is performed in two ways - standard, that outputs png figures and interactive, that uses JavaScript d3 package. Interactive visualization allows to analyze trio/family data, for example in search for causative variants in rare Mendelian diseases, in point-and-click interface. The package includes homozygous region caller, CRISPR guide prediction for discovered variants and allows to analyse whole human genomes in less than 30 minutes on a desktop computer. Finally, outputs from RareVariantVis is supported by other variant analysis tools, including http://gene.iobio.io. RareVariantVis disclosed novel causes of several rare monogenic disorders, including non-coding causative variant in keratolythic winter erythema.

**A Comparative Analysis of Feature Selection Methods for Biomarker Discovery in Study of Toxicant-treated Atlantic Cod (Gadus morhua) Liver**

*Xiaokang Zhang, Computational Biology Unit, University of Bergen*

Biomarker discovery is extraordinarily important in gene expression analysis in context of toxicant exposure. Among gene selection methods, differential expression analysis is often applied because of its simplicity and interpretability. But it treats genes individually, disregarding the correlation between them.

So some multivariate feature selection methods are proposed for biomarker discovery. We compared three methods that stem from different theories, namely Significance Analysis of Microarrays (SAM) which finds out the differentially expressed genes, minimum Redundancy Maximum Relevance (mRMR) based on information theory, and Characteristic Direction from a geometrical aspect, according to the stability and classification accuracy. The stability of one feature selection method is measured based on the overlap of selected features from different sampling steps. Using the subsets of selected features from 3 feature selection methods, we trained 4 classifiers, namely Random Forest, Support Vector Machine,

RIDGE regression, LASSO, and then test the prediction accuracy to see how well the subsets can improve it. Based on these two aspects, we studied the performance of 3 feature selection methods. Tested on the gene expression data from two toxicant exposure experiments, we found that GeoDE is more stable, and can give higher prediction accuracy in low-dose condition.

**An iterative approach for reconstructing ribo-somal genes in metatranscriptomics**

*Yaxin Xue, Computational Biology Unit*

Technological advances in metatranscriptomic approaches (MTS) have enabled a deep understanding of the structure and function of microbial communities.

Whole MTS (total RNA) provides a unique opportunity to investigate active microbial community from all three domains of life (rRNA) and function (mRNA) simultaneously. 16S rRNA reconstruction is a critical step to analyze the characterization of microbial community. However, current tools are developed for amplicon and met-agenomic data, which lack of ability to handle massive volume of dataset and high complexity in MTS. Here we introduce a novel Iterative Approach for Reconstructing Ribosomal genes (IARR) in MTS. It addresses the challenges by combining sampling, assembly, iteration and mapping steps.

We apply the approach to several simulated microbial communities, show that our tool can recover more rRNA genes with less false positive results and time/memory usage, comparing with several specially designed rRNA reconstruction tools.

**Online pipeline to store, analyze, and visualize proteomics data**

*Yehia Farag, Computational Biology Unit, Barsnes Group*

The amount of publicly shared proteomics data has grown exponentially over the last decade as the solutions for sharing and storing the data have improved.

However, the use of the data is often limited by the manner of which it is made available. There are two main approaches: download and inspect the proteomics data locally, or interact with the data via one or more web pages. The first is limited by having to download the data and thus requires local computational skills and resources, while the latter most often is limited in terms of interactivity and the analysis options available. A solution is to develop web-based systems supporting distributed and fully interactive visual analysis of proteomics data. The use of a distributed architecture makes it possible to perform the computational analysis at the server, while the results of the analysis can be displayed via a web browser without the need to download the whole dataset. Here we proposed a solution by developing web interface that interact with Galaxy server that allow users to store and analyze their proteomics data using SearchGUI and PeptideShaker tools and visualize the results through multiple connected interactive visual displays in a web-based setting.

**Co-expression module analysis on salmon-lice network**

*zhaoran zhou, Computational Biology Unit, UiB*

Lepeophtheirus salmonis (Sea louse) is the major pathogen affecting global salmon farming industry.

As a non-model organism, the parasitical mechanism of sea lice is practically unknown. In this study, we constructed 4 parasitism co-expression network based on the gene expression data from sea lice and salmon fish of 4 different time points. Network modules containing genes from both fish and lice samples were found. Several well-preserved modules across all time points were also observed. In addition, networks constructed from highly resistant fish samples were compared with counterparts of lowly resistant fish.

Differently correlated modules between the 2 fish groups were identified. Our analysis results would provide new insights into parasitical mechanisms of sea lice.