ABSTRACT BOOK

Summer School in Bioinformatics & NGS Data Analysis

CBiES, Jachranka, Poland
September 10-17, 2017

#NGSchool2017
`https://ngschool.eu/2017`

Organized by

- International Institute of Molecular and Cell Biology in Warsaw
- Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava
- Department of Information Technologies, Masaryk University in Brno
- Institute of Genetics, Hungarian Academy of Sciences in Szeged

Contributions

- Leszek Pryszcz: main organiser, `https://ngschool.eu/`, Lecture notes and more
- Organisation & scientific committee: Broňa Brejová, German Demidov, Alina Frolova, Katarzyna Kędzierska, Bogumił Konopka, Łukasz Kiełpiński & Tomáš Vinař
- IIMCB Admin, Grant, Finance & PR Units
- Centrum Badań i Edukacji Statystycznej GUS: accommodation & boarding
- PTBi: help with obtaining the grant from Polish Ministry of Science and Higher Education
- Ewa Ramotowska #NGSchool logo

Supporters This activity is financially supported by grants from **International Visegrad Fund** (Visegrad Grant No. 21710381), **Polish Ministry of Science and Higher Education** (842/P-DUN/2017) and **International Institute of Molecular and Cell Biology in Warsaw**.

Contents

Programme

We'll have **morning (9-13)** and **afternoon (14-18)** sessions with coffee breaks around 11:00 and 16:00. In the evenings, we'll have lecture-only sessions, shot-talks, discussions and some more relaxed activities. Breakfast will be served from 8:00, lunch at 13:00 and dinner around 19:00. Workshops last 4 hours and consists of theoretical introduction and practical exercises.

	Day 0: Sunday	
15:00	**Introduction to Bioinformatics & NGS**	*Broňa Brejová & Tomáš Vinař*
18:00	**Welcome, dinner & Shot talks #1**	*Leszek Pryszcz*
	Day1: Monday	
9:00	**Data visualisation**	*Przemysław Biecek*
14:00	**Molecular data integration**	*Jacek Marzec*
20:00	**Shot talks #2**	*Leszek Pryszcz*
	Day2: Tuesday	
9:00	**Workflows & pipelines**	*Alina Frolova*
14:00	**Genome & transcriptome assembly**	*Andrey Prjibelski*
20:00	**Beta & Bit games**	*Przemysław Biecek*
	Day3: Wednesday	
9:00	**Detection of structural variations**	*Tomasz Gambin*
14:00	**Introduction to Statistics**	*German Demidov*
14:00	**RNA Structure Probing**	*Łukasz Kiełpiński*
17:30	**Open science - discussion**	*Paweł Szczęsny*
20:00	**Chilling evening: BBQ #1**	
	Day4: Thursday	
9:00	**Functional genome annotation**	*Marina Marcet-Houben*
14:00	**ChIP-seq**	*Aliaksei Holik*
20:00	**Team-working**	*Panagiotis Theodorakis*
	Day5: Friday	
9:00	**Single-cell RNA-seq analysis**	*Davis McCarthy*
14:00	**Differential expression analysis**	*Davis McCarthy*
20:00	**Genome structure & function**	*Noam Kaplan*
	Day6: Saturday	
9:00	**Single-cell Hi-C data analysis**	*Aleksandra Galitsyna*
14:00	**ATAC-seq**	*Katarzyna Kędzierska*
14:00	**Microbial genomics**	*Adam Witney*
20:00	**Chilling evening BBQ #2**	
	Day7: Sunday	
10:00	**Recap & farewell**	*Leszek Pryszcz*

Nucleo-mitochondrial interactions in Candida albicans

Adam Kolondra

University of Warsaw, Poland

All eukaryotic cells contain residual genomes in their mitochondria. Our objective is to gain insights into the mechanisms of nucleo-mitochondrial interactions in the yeast C. albicans. We propose that they will reflect the co-evolution of the nuclear and mitochondrial genome. Mitochondrial ribonucleases are involved in general mitochondrial RNA turnover. A mitochondrial exoribonuclease encoded by the PET127 gene is responsible for degradation of intronic sequences and unspliced precursors. In S. cerevisiae it plays a role in mitochondrial RNA maturation and degradation via its putative 5'-3' exoribonucleolytic activity, which has never been directly demonstrated. RNA-seq of the pet127Δ mutant revealed changes of steady-state level of several mitochondrial mRNAs, and suggested that Pet127p could also degrade intron-containing RNAs. This is the first demonstration that Pet127p plays a role in degradation of introns and partially spliced precursors, and is thus an important contributor to RNA surveillance and turnover. A similar RNA-seq study is also planned for the main mitochondrial exoribonuclease – Dss1p. Noncoding mitochondrial transcripts of unknown function. During the transcriptomic study of C. albicans mitochondrial transcriptome, in addition to transcripts mapping to the annotated genes of the mitochondrial genome, we found significant transcriptional activity in the unannotated noncoding region of the inverted repeats. We also detected antisense reads clustering near the ends of transcripts or putative processing sites. We plan to study the function of these transcripts. In these projects we need a strong bioinformatics background to analyse quantitative and qualitative changes in mtRNA processing. Work in S. cerevisiae is particularly challenging, as the mtDNA contains only 13% of GC in 25 islands, making read mapping difficult. Studying the regulatory response of the nuclear genome to the dysfunction of the OXPHOS and the retrograde pathway. We compare the nuclear transcriptome of respiratory-deficient mutant strains with wild-type C. albicans using quantitative high-throughput RNA sequencing in order to identify transcripts induced or repressed in response to mitochondrial dysfunction. This requires individual bioinformatics approach allowing us to detect properly annotated transcripts, study functional enrichment, identify motifs in promoter sequences, and describe the mechanisms of up or downregulation of identified transcripts.

Microbial Genomics

Adam Withey

SGUL, London, UK

NGS is transforming clinical microbiology by enabling the prediction of drug resistance profiles in a much faster and robust way, leading to more appropriate, personalised treatment plans for patients. The result is better clinical outcomes for the patients, potentially reducing the risk of the development of antimicrobial resistant and reducing overall healthcare costs. In addition the high resolution obtained by NGS enables accurate tracking of infectious disease isolates, thus enabling doctors and public health officials to track and intervene rapidly in outbreak scenarios.

In this workshop you will hear of some real examples where NGS has been used to direct patient care and track active outbreaks. You will then perform the exact analysis yourselves, using sequence analysis tools to predict the drug resistance profiles of Tuberculosis isolates from a set of patients and then predict the transmission network between these patients.

Impact of Copy Number Variations on obesity in dogs

Adrian Grzemski

Poznan University of Life Sciences, Poland

Obesity is becoming a very serious problem in Western societies. More than 30% of Americans are obsessed and Europe is catching up very fast. With obesity there comes a great risk of heart diseases and strokes, diabetes, high blood pressure and many others. This problem is shared by man's best friend – dog. Percent of obese dogs in USA is very similar to what we observe in humans. Dogs are considered to be a great model to study human conditions. According to Online Mendelian Inheritance In Animals database there are 402 naturally occurring genetic diseases that has their counterpart in humans.

Despite years of intensive research our understanding of obesity is greatly unsatisfactory. To this date only 97 loci were associated with Body Mass Index variation and they explain less than 3% of it. Knowledge of obesity in dogs is even more obscure. Only 4 genes were officially connected with obesity. In case of both species variants associated with this disease were mostly short variants like Single Nucleotide Polymorphisms and Insertions/Deletions.

Class of variants that were overlooked in the context of obesity are structural variants. The special type of those variants are Copy Number Variants, CNVs for short. The copy number variant is the region of DNA that exists in the variable copy number compared to reference genome and is at least one thousand base pairs long. There are reports of such variants associated with human obesity, but there are no reports about such regions in dogs. The aim of my PhD will be to associate the risk of developing the obesity with the occurrence of copy number variations.

In my project I will analyze the data from Whole Genome Sequencing of 50 dogs, from same breed – Labrador Retriever, half obesed and half with normal weight. In this data I will detect all type of variation, Short Variants and Copy Number Variants, and try to find some statistically significant associations. Most promising variants will be send for validation on larger sample. Additionally I will describe the CNVs' distribution and length, their relative position to genes, and compare this information to human databases. Like in my previous projects, my work is purely in silico, any molecular analysis will be performed by other members of my Department or by third-party companies.

Correlation of Dcx protein expression and migration of patients derived glioblastoma cells, using chick embryo's CAM assay with gelatin-sponge

Adrian Odrzywolski

Department of Biochemistry and Molecular Biology, Medical University of Lublin, Poland

Glioblastoma multiforme is one of the deadliest diseases affecting mankind with an average survival of 15 months despite aggressive, multimodal treatment. This dismal prognosis arises, at least in part, from an extensive migratory capability of GBM cells that disseminate throughout surrounding neural tissue. Our proposal focuses on mechanisms responsible for this migration. One of the genes that participate in this process is doublecortin (Dcx), primarily found in migrating neuroblasts in developing CNS. Several studies addressed its role in GBM cells' migration albeit their results as well as conclusions are contradictory. It is likely that these discrepancies relate to highly heterogeneous nature of GBM with a number of clones present within a single tumor that might significantly differ in their phenotypes. The majority of aforementioned studies were based on existing glioma cell lines that implies clonal selection and thus hampers the evaluation of Dcx role in GBM biology in vivo. Therefore in our project we propose a novel migration assay based on chicken embryos. It ensures more versatile microenvironment when compared to classical cell cultures and allows tumor cells extraction at a given time with high reproducibility of the results. Implementation of this model should also facilitate an evaluation of probable differences in migratory capabilities between genomic subtypes of GBM defined in The Cancer Genome Atlas (TCGA) study. This important initiative defined four different subtypes of GBMs: proneural, neural, mesenchymal and classic based on genome profiling. To our best knowledge our project is the first to date that addresses possible differences in migratory patterns of different glioma subtypes. Results of our proposal should, first of all, allow a full characterization of a novel migratory assay that we propose. The implementation of the assay should result in detailed description of the relationship between migratory capabilities of GBM cells and Dcx expression and prove the foreseeable correlation between GBM genomic subtypes and migratory patterns of their cells. Moreover, the experiments put forward in our proposal might provide a theoretical base for development of a novel biomarker of tumor's cells migration into surrounding brain tissue. It should in turn facilitate stratification of brain tumor patients into therapeutic subgroups with more targeted therapeutic protocols (personalized medicine).

Structural variants discovery from whole genome sequencing data

Agnieszka Kraft

Centre of New Technologies, Warsaw University, Poland

The structural variants (SVs) discovery is a key element in understanding the impact of genome rearrangement on the 3D chromatin structure. SVs are prone to arise in repetitive regions and can form complex internal structures, therefore variants discovery remain challenging and limited to sequencing data. Next-generation sequencing (NGS) technologies generate reads ranging from dozens to hundreds of base pairs (bp) in length and with relatively low per-base error rates. Although NGS sequencing error rates are relatively low and their effects can often be mitigated with increased genomic coverage, repetitive sequence creates mapping ambiguity. Moreover, NGS methods do not always completely characterize large structural variants. Such limitations can be overcome by taking advantage of continuous long-reads. The best approach for calling SV is to use both short-read and long-read data as it can dramatically expand the ability to call structural variants. Popular approach to overcome performance limitations of any SV identification method is to use an ensemble of in silico calling methods and merge SV callsets under consensus, but such approach restrict the number of total reported possibly true events. We present new SV discovery framework that merges and optimizes performance across groups of callers for both short-read and long-read data using available knowledge as prior input.

For testing our framework we use 1000 Genomes Project Phase 3 data from both Illumina short-read and PacBio long-read sequencing. In addition, we have data from sequencing families of four (800 people), in which one child has type one diabetes (T1D). Our framework enable to determine variants common in both children and those unique for child with T1D. The unique variants together with Hi-C data from 3D genomic experiments allow us to find structural differences in chromatin three-dimensional conformation between T1D-child and healthy sibling leading to link the phenotype with differences at the sequence level.

Single-cell Hi-C data analysis

Aleksandra Galitsyna

MSU, Moscow, RU

Application of the next generation sequencing for single-cell studies is one of the molecular biology cutting edges providing insights into mechanisms and variability of fundamental cellular processes. One of the most recent examples is the single cell analysis of chromatin structure with chromosome conformation capture (Hi-C), which was introduced in 2013 and further developed four years later by three independent groups. The single-cell Hi-C data processing is not a trivial task. First of all, no gold standard exists even for the processing of ensemble Hi-C data. Second, high levels of missing data and noise require the invention of new robust approaches.

In this workshop I reproduce the solution from one of the recent seminal studies [Flyamer et al. Nature 2017]. It utilizes Python package hiclib (https://bitbucket.org/mirnylab/hiclib/), which is currently one of the most popular software for Hi-C data analysis. In user-friendly Jupyter notebook environment, I show the traditional ensemble Hi-C data processing from raw sequencing output, including steps of reads mapping, filtering, data binning and construction of chromatin interaction map. An important lesson for students is the detection of chromatin features, such as chromatin compartments and topologically associating domains (TADs). I demonstrate how the same processing protocol can be applied and modified for single-cell analysis, following Flyamer and colleagues example.

As a result, I would like to demonstrate the specifics of the field of chromatin research, its unsettled state and variability of approaches.

Expansion of the antimicrobial peptide repertoire in the invasive ladybird Harmonia axyridis

Alexander Lüttringhaus

Free University of Berlin Department of Mathematics and Computer Science, Germany

The harlequin ladybird beetle Harmonia axyridis has emerged as a model species in invasion biology because of its strong resistance against pathogens and remarkable capacity to outcompete native ladybirds. The invasive success of the species may reflect its well-adapted immune system, a hypothesis we tested by analysing the transcriptome and characterizing the immune gene repertoire of untreated beetles and those challenged with bacteria and fungi. We found that most H. axyridis immunity-related genes were similar in diver- sity to their counterparts in the reference beetle Tribolium castaneum, but there was an unprecedented expansion among genes encoding antimicrobial pep- tides and proteins (AMPs). We identified more than 50 putative AMPs belonging to seven different gene families, and many of the corresponding genes were shown by quantitative real-time RT–PCR to be induced in the immune-stimulated beetles. AMPs with the highest induction ratio in the chal- lenged beetles were shown to demonstrate broad and potent activity against Gram-negative bacteria and entomopathogenic fungi. The invasive success of H. axyridis can therefore be attributed at least in part to the greater efficiency of its immune system, particularly the expansion of AMP gene families and their induction in response to pathogens.

Workflows & pipelines

Alina Frolova

IMBG, Kyiv, UA

Reproducible and efficient scientific pipelines are the core element of the successful and solid research. And the bioinformatics field is not an exception here, on the contrary, the seeming easiness of re-doing the "experiment" and therefore less thorough testing and intermediate results check-up may lead to mistakes in the initial steps and ruin your final conclusions. Putting aside errors in the calculations or results interpretation, bioinformaticians usually need to run the same tool multiple times with different parameters or need a series of consequent code testing while developing new software. And doing it manually is not an option. To address mentioned issues number of tools have been developed, many of which originate from pure Computer Science like version control systems, package, dependency and environment management systems or integrated development environments. There are also bioinformatics workflow management systems — specialized form of workflow management systems designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, that relate to bioinformatics. Among the well-known examples are Galaxy, GenePattern, KNIME. However, there are even more flexible tools for precise pipelines constructions and fine-grained parameters tuning such as SnakeMake — a scalable bioinformatics workflow engine, or NextFlow — a domain specific language for parallel and scalable computational pipelines. During the workshop we will try to review key components of efficient workflows management to make the everyday life of bioinformatician easier and more productive.

Population structure analysis using the NGS data

Anastasiia Hryhorzhevska

Warsaw University Of Technology, Faculty Of Electronics and Information Technologies, Poland

Genomic variant data obtained from the next generation sequencing can be used to study the population structure of the genotyped individuals. Typical approaches to ethnicity classification/clustering consist of several time consuming pre-processing steps, such as variant filtering, LD-pruning, and principal component analysis or multi-dimensional scaling transformations of genotype matrix. We have developed a framework using R programming language to analyze the influence of various pre-processing methods and their parameters on the final results of the classification/clustering algorithms. The results indicated how to fine-tune the pre-processing steps in order to maximize the supervised and unsupervised classification performance, and choose the most accurate predictive model. Since the whole genome sequencing data generated in large-scale sequencing projects, increases the feature space by orders of magnitude, we have developed another, distributed computing, framework using Apache Spark to process the large data sets in a manageable time. The scalability of machine learning methods was examined, defining scalability by the effect that an increase in the size of the dataset has on the computational performance of the algorithm from the particular library. Tests performed on 1000 Genomes data set confirmed the efficiency and scalability of the presented approach. Finally, the dockerized version of the implemented frameworks can be easily applied to any other variant data set, including data from large scale sequencing projects or custom data sets from clinical laboratories.

Interaction between phages and bacteria

Andrej Baláž

Comenius University, Faculty of Mathematics, Physics and Informatics Geneton s. r. o., Bioinformatics Department, Slovakia

In the last years new strains of bacteria resistant to many known antibiotics appeared. The development of new antibiotics requires a lot of funds and many experts. The process of approving new drugs is highly time consuming and expensive. Altogether with a fast evolution of bacteria, which are able to quickly develop resistance to antibiotics, this creates pressure for new approaches, how to cure bacterial infections. Phage therapy appears as one possibility how to address this problem.

In our work we are searching for genes responsible for phage effectiveness in killing pathogenic bacteria. We search for marker genes that are vital for the phage to invade bacterial cell, replicate and kill specific bacteria.

We hope our further research will lead to better understanding of interactions between phages and bacteria. Furthermore, it can help us to predict effectiveness of specific phage used on a microbiome with a potential for curing of resistant bacterial strains infections.

De novo genome and transcriptome assembly

Andrey Prjibelski

SPSU, St. Petersburg, RU

While the majority of projects related to human health use reference-based analysis of sequencing data, de novo analysis is essential when studying previously unsequenced organisms. De novo genome assembly from short reads is a challenging algorithmic problem, complexity of which highly depends on the genome size and structure. Although bacterial genomes are small and typically do not have complex repeats, there are multiple ways of sequencing bacteria: conventional (isolate) sequencing, metagenomics (sequencing whole bacterial community at once) and single-cell sequencing (implies whole genome amplification). During the first part of this class we will cover common NGS processing tools for QC and filtering, learn about basic assembly algorithms and assemble conventional and single-cell bacterial dataset. In addition, we will talk about differences between these ways of sequencing bacteria in terms of quality of raw data and assembly. The second part will be devoted to RNA-Seq data processing: QC, filtering, de novo assembly and it evaluation. Although size de novo transcriptome assembly seems to be a less challenging problem than a genome assembly, it is amplified by highly uneven coverage depth (due to different expression levels) and presence of alternative splicing in eukaryotic genomes. During this class we will also discuss similarities and differences between genomic and transcriptomic data processing, de novo assembly and assembly evaluation.

Machine learning and third generation sequencing.

Bogumil Konopka

Department of Biomedical Engineering, Wroclaw University of Science and Technology, Poland

Machine learning techniques have a broad range of applications in different fields of bioinformatics. For instance they have been successfully used to support protein structure prediction at different stages of prediction process. Recently we have developed a machine-learning based tool that allows accurate estimation of the quality of inter-residue contact predictions. Based on those estimates it is possible to state the usefulness of contact predictions in the structure prediction process. Advanced data analysis methods can also be applied to next –generation sequencing data. The Oxford Nanopore Technologies's smart phone sized sequencer – MinION, thanks to its characteristics, has the potential to revolutionize many sequencing applications, such as viral and bacterial identification, structural variant calling, genome assembly. However, before that, many issues need to be solved. For instance the quality of sequencing needs to be improved. Currently it is still significantly lower than the quality of other sequencing techniques. Also most methods for processing sequencing data, such as alignment or variant calling, were designed for short read sequencing. New methods need to be developed or old ones tuned for processing ONT long reads data. Our interest is to use various machine learning and data exploration techniques to improve the applicability of ONT MinION sequencing.

Introduction to Linux, Bioinformatics & NGS

Broňa Brejová & Tomáš Vinař

Univerzita Komenského, Bratislava, SK

In this workshop, we will go through several exercises which will allow participants to learn or improve basic skills in working with Linux command line and processing NGS data using bioinformatics software. Exercises will include genome assembly and read mapping, gene finding, and detection of positive selection. Participants can choose exercises according to their proficiency levels, advanced bioinformaticians are welcome to help beginners. In case of interest, exercises will be complemented by short lectures on basics of bioinformatics methods used. During this workshop, we will also help participants to install software needed later during the summer school.

Introduction to the analysis of single-cell RNA-seq data

Davis McCarthy

EBI, Hinxton, UK

Today it is possible to obtain genome-wide transcriptome data from single cells using high-throughput sequencing (scRNA-seq). The main advantage of scRNA-seq is that the cellular resolution and the genome wide scope makes it possible to address issues that are intractable using other methods, e.g. bulk RNA-seq or single-cell RT-qPCR. However, to analyze scRNA-seq data, novel methods are required and some of the underlying assumptions for the methods developed for bulk RNA-seq experiments are no longer valid.

In a short space of time, many methods have been developed to address important questions that can be address with scRNA-seq and key aspects of scRNA-seq analysis: pre-processing and quality control, visualisation, clustering, differentiation trajectories and differential expression. This workshop provides an introduction to these key topics and demonstrates the use of a set of open-source R packages to solve frequently-encountered problems in scRNA-seq analysis.

The workshop will use material developed with Martin Hemberg, Tallulah Andrews and Vlad Kiselev, which is available here: https://hemberg-lab.github.io/scRNA.seq.course/index.html. The course is taught through the University of Cambridge Bioinformatics training unit, but the material found on these pages is meant to be used for anyone interested in learning about computational analysis of scRNA-seq data. The course is taught twice per year and the material here is updated prior to each event.

Genetics in animal husbandry (cows)

Ekaterina Orlova

Wroclaw University of Life and Environmental Science, Poland

I can briefly tell about application of bioinformatics in animal husbandry, health and performance

Long terminal repeats power evolution of genes and gene expression programs in mammalian oocytes and zygotes

Filip Horvat

Institute of Molecular Genetics, Prague, Croatia/Czech Republic

Retrotransposons are "copy-and-paste" insertional mutagens that substantially contribute to mammalian genome content. Their amplification threatens genome integrity through insertional mutations and chromosomal aberrations so defensive mechanisms evolved to suppress their activity in genome. However, retrotransposons can also provide functional gene parts, such as promoters, enhancers, exons, terminators or splice junctions. Retrotransposons often carry long terminal repeats (LTRs) for retrovirus-like reverse transcription and integration into the genome. In mouse, we identified >800 LTRs from the mammalian endogenous retrovirus-related ERVL retrotransposon class which compose mobile gene-remodelling platforms by providing promoters and first exons. The LTR-mediated gene remodelling also extends to hamster, human, and bovine oocytes. We show several examples of LTRs role in a stage-specific manner during the oocyte-to-embryo transition - activating transcription, altering protein-coding sequences and producing non-coding RNAs. We also report a novel protein-coding gene evolution where a LTR provided a promoter and the 5' exon with a functional start codon while the bulk of the protein-coding sequence evolved through a CAG repeat expansion. Altogether, ERVL LTRs provide molecular mechanisms for stochastically scanning, rewiring, and recycling genetic information and are tightly connected with gene expression and evolution in the germline.

EIS-DB: Exon-Intron Structure Database

Irina Poverennaya

Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Russia

We present a new exon-intron structure database (EIS-DB) containing comprehensive data of well-annotated genes in more than 100 eukaryotic genomes from different taxonomic groups (vertebrates, invertebrates, plants, and fungi). It allows extracting data related to a special gene or isoform of a special organism or obtain statistical data related to the given set of genes and/or organisms. Although the similar databases exist, they are mainly out of date, or taxonomic- or isoform specific. EIS-DB is a relational database managed by PostgreSQL. Structurally, it contains 15 tables. The main ones are 'Organisms', 'Genes', 'Isoforms', 'Orthologous groups', 'Exons' and 'Introns'. The others contain auxiliary data, e.g., taxonomy; EIS-DB also contains fasta-files with related sequences. The detailed intron section – there are data about intron length, phase, sequence, splice sites, etc. – makes EIS DB especially appealing for studying various intron features in different organisms such as phase vs length correlations, non-canonical splice sites distribution and etc. It makes also possible to study evolution of introns as a part of gene (intron gain, loss, sliding) and as a separate sequence. As the main source of gene sequences and annotations, 112 RefSeq genome assemblies, (current to March 2017) were used along with additional input data on gene orthology obtained from NCBI. To ascertain orthology between exons (and introns) we have developed a special tool. It first builds multiple protein alignment using modification of MUSCLE program taking into account data on exon borders. Then it realigns the alignment regions where exon borders are not well aligned. The orthologous groups of exons and introns are determined based on the refined alignment. The preliminary version of web interface of EIS-DB is available at http://212.47.226.240:3000/; the database could be downloaded to user's PC for more advanced requests.

Molecular data integration

Jacek Marzec

BCI, London, UK

The open access to huge volumes of genomic data stored in public data repositories, such as Gene Expression Omnibus (GEO), ArrayExpress or Sequence Read Archive (SRA), enables novel large-scale discovery studies. Indeed, integration of data produced across many studies is now well-recognised as a powerful approach that provides novel biological insights while allowing for the identification of numerous alterations contributing to a given phenotype not evident from single experiments. Multi-study data integration increases the statistical power to capture consistent molecular alterations that might be hampered by a limited sample size and experimental artefacts associated with individual datasets, and thus offers more accurate signatures. Moreover, cross-study data integration gives the opportunity for broader data overview and the potential to ask novel biological questions.

During this workshop you will integrate publicly available molecular data from independent microarray experiments on prostate cancer.

Studying variability of organellar genomes of bryophytes

Kamil Myszczyński

Department of Botany and Nature Protection, University of Warmia and Mazury, Poland

I am a fourth year PhD student at the University of Warmia and Mazury in Olsztyn, Poland. My interest in next generation sequencing data began four years ago, when I was about to finish my master degree project which was based on microarray data analysis. Right now as a member of bryologists team I am participating in research project which is focused on comparative genomics of mosses and liverworts. These small plants used to be considered as highly conservative in their genomes structure and sequences. However our studies revealed organellar genomes of bryophytes to be more variable than expected. Using next generation sequencing methods we were able to compare diversity between species, genra and families of bryophytes on mitochondrial and chloroplast genome level. To make those analyses efficient and reliable I am doing my best to adapt bioinformatics methods in daily work. Right now I have some experience in analysis of next generation sequencing data - especially in assembling small organellar genomes. However I am fully aware that it is just the top of the iceberg in terms of NGS data analyses and I would like to learn new bioinformatics methods. I am convinced that participation in the course will help me to plan ahead RNAseq experiment in a proper way and draw relevant biological conclusions. I would be delighted to have a chance to exchange my experience with participants and speakers of the course, sharing their great bioinformatics knowledge and good analytical practice.

Integrative Galaxy tool for genomic data visualization

Karolina Sienkiewicz

University of Warsaw, Institute of Informatics, Poland

Nowadays with the advance of high-throughput genomics and common application of machine learning to biological data, most of the current scientific projects are based on big data analysis. Dealing with such abundance of data can require researchers to keep track of additional metadata, which usually has to be afterward published or otherwise released. Supporting results with data visualization is desirable - especially when conducting whole genome experiments.Plenty of solutions were proposed to make a complex analysis of biological data more accessible, however, visualization of a large amount of data and sharing it remain an ongoing problem.

We are creating the integrated tool which allows exporting data from Galaxy web platform and visualizing it automatically. Galaxy provides an accessible web interface for reproducible analysis of biological data. It can be easily used by scientists without backgrounds in bioinformatics and big groups of researchers working together. Our main goal is to automatize data flow between computational biomedical research and data visualization for projects which conduct a genome analysis.

We are working on the incorporation of data visualization in a variety of ways. The most important of them are visualization in individual instances of JBrowse Genome Browser (already implemented) and automatic creation of Track Hubs for data visualization in UCSC Genome Browser. We also want to provide a universal interface for the individual project's websites (opportunity to easily browse and download files) and a simple way to manage access settings to keep data secure while simultaneously sharing and collaborating on it.

ATAC-seq

Katarzyna Kedzierska

University of Virginia, Charlottesville, USA

Assay for Transposase Accessible Chromatin followed by Next Generation Sequencing (ATAC-seq) is a rapid, sensitive and efficient method for mapping chromatin accessibility genome-wide. The method requires no more than 50k or as little as 500 cells as input and the straight-forward protocol comes to isolating the nuclei and in vitro transposition of sequencing adaptors into native chromatin. Samples can be ready for sequencing in less than 3 hours following cell harvest.

Some of the many advantages of the method include: ATAC-seq doesn't require sonication or the phenol-chloroform extraction (FAIRE-seq), no antibodies needed (ChIP-seq), no sensitive enzymatic digestion (MNase-seq or DNA-seq), and significant reduction of the required input material and time needed to process the samples.

The workshop will focus on hands-on analysis of the already published data. We would work through experimental design, pre-processing the data and the analysis. Topics covered by the workshop include motif search, nucleosome positioning and TFs footprinting.

Systems Biology Analysis of Kinase Inhibitors in Cancer Cells Using Next Generation Sequencing Data

Kübra Narcı

Health Informatics, Middle East Technical University, Turkey

Hepatocellular Carcinoma (HCC) is the fifth most prevalent type of cancer although it is the second leading cause of cancer-related mortality (GLOBOCAN 2012). Due to rising global obesity rates (a risk factor for HCC), this cancer is not only one of the important health issues facing Turkey but also in Western nations. The mechanism responsible for the development of HCC is highly complex due to tissue heterogeneity. Although the traditional approaches focus on single gene or locus, understanding the variations in the signalling pathways/networks of diseased cells during hepatocarcinogenesis may help to develop new strategies or drugs to prevent cancer progression in the patients. This thesis proposal aims to focus on enlightening the transcriptome sequencing of dysregulated genes in HCC mainly concentrating on known disease signalling pathways. For this purpose, RNA sequencing (RNA seq) data of two HCC cell lines Mahlavu and Huh7 targeted by three kinase inhibitors and two of their combinations or DMSO as control will be analysed for differential gene expression. We acquired a list of genes potentially responding these kinase inhibitors and their combination for each cell line. We investigated the functional pathways enriched with these important genes by solving a graph problem called as Prize Collecting Steiner Tree on human interactome. Significantly enriched HCC networks are analysed inside and between kinases perspective. In this wise, the investigation provides a global view of responses to the kinase inhibitors, information about characterization of novel markers and give us molecular clues to develop new HCC treatments. Furthermore, as a result of this study, a pipeline exampling analysis of cross studies of transcriptome data will be available. Finally, we hope that novel HCC drug targets will be acquired as a result of this study. We anticipate that as a result of our efforts, a measurable socio-economic impact will be the potential improvement in the treatment of HCC by discovering novel drug targets which may lead to more cost-effective and diverse treatment options available for the treatment liver cancer.

AGEL laboratories

Květoslava Faltýnková

Biomedical engineering, VŠB-Technical University of Ostrava /Medical genetics, AGEL laboratories, Czech Republic

I work as bioinformatics in genetic laboratory in Czech Republic. I analyze data rising from the next generation sequencer and I create a final report for molecular genetics. Our lab uses targeted sequencing. We design our own sequential panels that are used to diagnose hereditary syndromes and cancer. We have Nephrology panel, Cardiomyopathy panel, Ophthalmology panel, Oncology panel, Child-syndromology panel and Connective tissue panel. Our research interests are: - Design and optimization of the targeted sequencing panel for the identification of cancer susceptibility in high-risk individuals from the Czech Republic. In these patients, the identification of pathogenic mutations in cancer susceptibility genes has an important predictive and in some cases, prognostic value. Some mutations are very rare, occurring with substantial population variability, and their clinical interpretation is very complicated. The key objective of this project is to create a database which would help improve the interpretation of rare or population-specific variants in cancer susceptibility genes. - Identification of genetic damage EGFR signaling pathway in the cell to determine prognosis and treatment for colorectal cancer. The aim of this work is to design and implement an EGFR signaling pathway model in a cell using Artificial Intelligence methods - neural networks and fuzzy expert systems.

New genetic mechanisms of primary immunodeficiency diseases

Laura Batlle Masó

Experimental and Health Sciences Department, Pompeu Fabra University (Barcelona, Spain), Spain

Primary Immunodeficiency Diseases (PIDs) affect nearly 6 million individuals worldwide, more than 680.000 persons in Europe. Genetic defects in cells and molecules of the immune system are underlying the disease, causing a heterogeneous group of disorders characterized by poor function or lack of one or more components of the immune system. Clinical manifestations of the disease are highly variable as well as the severity of it. Although enormous advances have been achieved, especially in the innate immunity field, genetic bases of many PIDs still remain unknown. The research project I propose is to use Next Generation Sequencing (NGS) and Molecular Inversion Probe (MIP) technologies to identify novel genetic mechanisms in the pathogenesis of PIDs. The study will be carried out with patients without clear diagnosis; we aim to detect new genetic variants both germline and somatic. Once identified the candidate genes, validation studies (gene expression analysis, segregation studies and cytometry assays) will be performed to establish unequivocally the causative relationship with the observed phenotype. The results of the study will create new genetics data which will lead to an important step towards PIDs pathogenesis comprehension, as well as to a better understanding of the contribution of somatic mutations to human disease.

Investigation of the epidermal transcriptome in psoriasis

Lorenzo Pasquali

Dermatology and Venereology Unit, Department of Medicine, Solna (Karolinska Institutet), Sweden

Background: Psoriasis is a multifactorial immune-mediated skin disease, characterized by epidermal hyperproliferation and infiltration of immune cells the skin lesions. Both keratinocytes and immune cells contribute to the development and maintenance of chronic skin inflammation. Previous studies characterizing the transcriptomic landscape in psoriasis used full-depth skin biopsies, in which cell-specific expression changes may have been missed. Objective: Here we aimed to identify transcriptomic changes in the epidermis in lesional and non-lesional psoriasis skin. Methods: CD45neg epidermal cells (epidermal non-immune cells, mainly keratinocytes) were sorted from skin biopsies collected from healthy skin as well as from non-lesional and lesional skin of psoriasis patients. Affymetrix Human Transcriptome Array 2.0 platform was used to identify differentially expressed genes (DEGs). Enrichment of genes belonging to functional categories (GO categories) was assessed using the Enrichr web-tool. Differential expression of selected DEGs has been validated by RT-qPCR. Results: In CD45neg epidermal cells, 525 genes were significantly up-regulated and 198 genes were down-regulated in psoriasis lesional skin compared to healthy skin. Network analysis of functional alterations revealed transcripts related to innate immunity, type I interferon response, cell cycle, keratinization and response to cytokines genes to be the hubs of molecular alterations in the epidermal tissue in psoriasis lesions. qPCR analyses confirmed the up-regulation of the interferon-response genes IFI44, IFI44L, and DDX60, as well as of TNIP3, a regulator of NF-kB signalling. Moreover, down-regulation of SPRR4, a gene involved in skin barrier formation, and NR4A3, a transcription factor implicated in inflammation, were confirmed by qPCR. Conclusion: Detailed characterization of the epidermal transcriptome in psoriasis, excluding the immune cell signature, has revealed dramatic gene expression changes in the psoriatic lesional epidermis, with deregulated genes related to innate immunity, type I interferon response, proliferation and differentiation. Our results highlight the role of keratinocytes in psoriasis and identify transcripts previously uncharacterized in psoriasis.

Massive Parallel Sequencing-based RNA Structure Probing

Łukasz Kiełpiński

Hoffmann-La Roche, Frederiksberg, DK

RNA molecules are central to conveying and regulating gene expression. They exist as three-dimensional entities with the structure largely determined by their base-pairing pattern (secondary structure). Function of many non-coding RNAs, such as ribosome, tRNAs or riboswitches, among many others, directly depends on their structure. Messenger RNA molecules, apart from coding for proteins, also carry a structural layer of information, which can influence interactions with other molecules leading to multitude of downstream effects, such as modulating splicing, polyadenylation, translation efficiency, RNA modifications or interactions with microRNAs and RNA binding proteins. Moreover, understanding mRNA structure is important for the design of RNA drugs as it can affect the siRNA and antisense oligonucleotides knockdown efficiencies. The structure of an RNA molecule can be to a certain extent predicted using many different computational approaches, whose accuracy can be largely increased using evolutionary or experimentally defined constraints. Strong evolutionary signal is considered a gold-standard for proving a certain functional structure, but the data is often not available. Traditional experimental methods are low throughput and labor intensive. The advent of the massive-parallel sequencing allowed to simultaneously probe RNA structures of large sets of molecules in a single experiment. Published methods for high-throughput RNA structure probing will be discussed. During the workshop we will use DMS-Seq as an example (Rouskin et al. 2014). We will start from processing raw sequencing reads, and we will calculate the normalized structure scores, all using publicly available tools such as bowtie2 (Langmead et al. 2012) and RNAprobR (Kielpinski et al. 2015). We will compare those scores to the annotated structure of the mitochondrial ribosome to evaluate the performance of the DMS-Seq method and visualize the data.

Whole genome and transcriptome studies of two non-model species of euglenids: Euglena longa and E. hiemalis

Magdalena Płecha

Department of Molecular Phylogenetics and Evolution, Institute of Botany, Faculty of Biology, University of Warsaw, Poland

Euglenids are a diverse group of the Euglenozoa and comprises of primary heterotrophs, phototrophs, and secondary heterotrophs. Although, the euglenids phylogeny has been studiem extensively, the sequences of whole nuclear genomes have not yet been introduced. While, the studies of this group's genetic composition shall particularly bring new insights into the studies of atypical introns. This project regards the E. longa and E. hiemalis de novo genomes sequencing, assembly and annotation. We plan to obtain and analyze the E. hiemalis and E. longa transcriptome, followed by the analysis and comparison to their genomes. Obtained data would be a source of information about the novel genetic regions and their regulation mechanisms. We are also aiming to seek for general patterns of introns distribution, their origin and types. Total DNA and RNA will be isolated from E. longa (SAG 1204-17a) and E. hiemalis (CCAP 1224/35) cultures. An Illumina RNA sequencing pair-end library with the polyA selection will be constructed, whereas for DNA - PacBio and Illumina pairend and mate-pair libraries will be obtained. Data quality will be assessed using the FastQC and Trimmomatic. For transcriptome assembly Trinity will be used, whereas for genome assembly - SPAdes, SOAP, AbySS and MIRA. Coverage would be evaluated with bowtie2, while possibile contamination with Blobology. Further steps will include ab initio gene prediction using Augustus, supported by transcriptomic data mapping and functional annotation with BLASTP and HMMER3 (against NCBI nonredundant and Pfam protein motif databases respectively). Recently, we have obtained the preliminary sequencing data (MiSeq) for E. longa and E. hiemalis genomes. The quality of reads was high, although as expected in regards to the type of sequencing method used, the estimated coverage was quite low. The first SPAdes assembly of E. longa resulted in 212517, whereas E. hiemalis 169905 contigs. The total lengths of the E. longa and E. hiemalis assemblies are close to 0.23 Gb and 0.14 Gbp, with N50 equal to 1058 and 809 bp, respectively. Roughly around 70% of raw reads have been mapped to both obtained assemblies using the bowtie2. The longest contig of the E. hiemalis assembly corresponds to yet unpublished sequence of this species chloroplast genome and its annotation is now in progress.

Predicting disease from gut microbiota codon usage profiles

Maja Kuzman

University of Zagreb, Faculty of Science, Department of biology, Croatia

Metagenomics projects use next-generation sequencing to unravel genetic potential in microbial communities from a wealth of environmental niches, including those associated with human body and relevant to human health. Analysis of metagenomic data often includes ranking importance of gene functions according to their respective abundance. Such analyses might fail to identify functions, or even entire pathways that are differentially regulated rather than dependent on presence or absence of a particular gene. While metatranscriptomic, and subsequently metaproteomic approaches provide more biologically relevant information, experimental methods to obtain the data are less robust, more complex, and expensive. By applying concepts of translational optimization through codon usage adaptation on entire metagenomic datasets, we demonstrate that a bias in codon usage present throughout the entire microbial community can be used as a powerful analytical tool to predict for community lifestyle-specifc metabolism. Here we combined this approach with machine learning, to classify human gut microbiome samples according to the pathological condition diagnosed in the human host. By exploring synonymous codon usage selection and their adaptation across the community, we determined levels of translational optimization and predicted genes optimized for high levels of expression in intestinal metagenomes of cirrhotic patients and healthy individuals. Based on their translational optimization and predicted expressivity, we used Random Forests machine learning method to classify genes and metagenome samples into groups associated with healthy and diseased phenotype. We also classified gene functions according to their annotations available through the orthology database KEGG, sorted them in corresponding metabolic pathways, and analyzed in terms of abundance of translationally optimized genes. Unequal abundance of translationally optimized genes in different metabolic pathways of intestinal microbial communities of healthy and sick individuals provides a diagnostically relevant signal and opens up a possibility for mechanistic insight into the interaction between microbial and human metabolism in development of this disease.

Big data methods in NGS data analysis

Marek Wiewiórka

Warsaw University of Technology, Poland

Genomic population studies incorporates storing, analyzing and interpretation of various kinds of

genomic variants as its central issue. When thousands of patients sequenced exomes and genomes are being sequenced, there is a growing need for efficient database storage systems, querying engines and powerful tools for statistical analyses. Scalable big data solutions such as Apache Spark, ADAM,Apache Impala, Apache Kudu, Apache Phoenix or Apache Kylin can address many of the challenges in large scale genomic analyses.

Applying cancer subpopulations analysis methods to metagenomic series

Margarita Akseshina

Saint Petersburg Academic University, Russia

In many metagenomic studies, multiple similar metagenomic samples are available forming time- or spatial- series. In theory, such series provide unprecedented opportunities for decomposition of the mixtures of closely-related bacterial strains.

Arguably the basic problem one can formulate is: using a metagenomic series data and reference genome for particular species, identify the number of related strains for and their relative abundances across the samples. Surprisingly, very few options exist to perform this kind of analysis. In 2015 Luo et al. developed ConStrains tool, which has a lot of shortcomings, in particular it uses questionable computational model and we did not succeed in reproducing results from the paper.

While the problem of strains detection has only recently been introduced in metagenomic series analysis, a closely related computational problem of detecting cancer subpopulations in a series of tumor samples has been extensively studied in the past five years. Multiple software tools have been developed around advanced statistical and/or algorithmic approaches (e.g. Clomial [Zare et al, 2014], PhyloWGS [Deshwar et al, 2015], Pyclone [Roth et al, 2013], etc.).

In this study we tried to apply those tools to the analysis of related strains in metagenomic series. We adjusted Clomial and PhyloWGS tools to metagenomic series data and tested them on simulated and real datasets. While we failed to achieve good results with PhyloWGS so far, we observed that Clomial could be successfully used in metagenomics setting, producing more accurate results compared to ConStrains.

Analysis of population specific genetic risk factor distribution using machine learning techniques

Maria Nikoghosyan

Bioinformatics and Bioengineering , Russian-Armenian University, Armenia

The genome wide association study (GWAS) approach is an important tool of clinical genomics and allows estimation of genetic risk factors for diseases for an individual, as well as for the entire population. While the most common setting of GWAS is the case control studies, it has been rarely used to evaluate the distribution of disease associated polymorphisms on the population level and comparison of complex disease susceptibility between ethnic groups. In this study we were interested in the description of an entire landscape for disease associated single nucleotide polymorphisms (SNPs) in different geographical regions. For this purpose, we used array-based genome-wide data of SNPs, taken from Human Genome Diversity Project and from Ethnogenomics Laboratory of Institute of Molecular Biology NAS RA. Disease-SNP association information was obtained from DisGenNet and GWASdb databases. Overall our data contains 40037 disease associated SNPs for 1094 samples from 33 population in 8 geographical regions (Armenia, Africa, East Asia, Middle Ease, Central South Asia, Europe, Oceania, America). To construct the map of disease associated SNPs, we used self-organizing map (SOM) analysis. A SOM clusters the SNPs, which have a similar distribution across the samples and represents those clusters as spots. SNPs in particular spots have different distributions across populations and each spot can be descriptive for one single population. It is possible for SNPs in a spot to be overrepresented in more than one population which could indicate some form of similarity between these populations. Our results show strong similarity in disease associated SNP distribution between Armenia, Middle East, Europe and East Asia. All populations were characterized by presence of SNP associated with by coronary artery disease, amyotrophic lateral sclerosis and carcinoma of lung. However, in different populations different SNPs were overrepresented, which suggest about development of population/geographical area-specific patters for gene-environment interaction and susceptibility to the complex diseases. Analysis of population specific genetic risk factor distribution using machine learning techniques.

Functional genome annotation

Marina Marcet-Houben

CRG, Barcelona, ES

One of the natural steps that needs to be done after assembling a new genome is to predict which genes are encoded in it and have an idea about their functionality. Discovering genes in prokaryotic genomes is a relatively simple matter due to their lack of introns and clear promoter-sequences. For eukaryotic genomes the process is more complex due in large part to the presence of long introns. The first part of the class will show different kinds of programs to use to predict gene in a genome and how to decide which program is the best one for our kind of data. Once genes are predicted we are also interested in knowing their function. Experimental analysis are expensive and time-consuming so it is a good idea to have a general idea of the function of an unknown protein before we start working with it. Additionally, while blast searches are the universal tools used to assign function to a protein, there are times that this transference is incorrect or that it does not provide us with any information. We will explore and understand the limitations of blast, how to go around them and which other tools we can use when blast searches fail.

How can be rumen microbiome influenced by air intake and type of feeding?

Martina Zapletalová

Department of Biochemistry, Faculty of Science, Masaryk University, Czech Republic

Rumen is an open anaerobic microbial ecosystem inhabited by a variety of bacteria, protozoa, archaea, fungi, and viruses. These microorganisms are responsible for fermentation and degradation of feed and influence health and immune responses of the cattle. With the number of more than 1 000 cells per ml, bacteria are the predominant microorganisms in the rumen fluid. Bacteria are highly sensitive to fluctuations in the rumen milieu when besides the other parameters they are highly influenced by the feeding regime and by redox potential. In vitro techniques realize more controlled and reproducible conditions compared to in vivo experiments so they are widely used to study the various processes in the rumen. A rumen simulation should mimic physical conditions of the natural rumen (temperature of 39-40 °C and pH values range between 6.3 and 7.3) and maintain the diversity and concentrations of the natural microbial populations. Regarding the available data, redox potential values vary between -180 to -220 mV. Previously, we highlighted the major influence of the low redox potential on rumen microbiome during in vitro cultivation. Low redox potential is caused by strict anaerobic conditions in the bioreactor, so we decided to increase it to the physiological value by using air supply. In this study we focused on monitoring of the rumen microbiome during cultivations with and without air inflow by using various types of feeding. To ensure stable conditions we monitored pH values, redox-potential values and VFA content throughout the ten-day cultivation run. The changes in bacterial community of rumen fluid were monitored by 16S rRNA sequencing. Under conditions of micro aeration, we found much more stable microbiome during the ten days of cultivation, compared with cultivation without air intake. Moreover, we found that the stability and richness of microbiome are largely influenced by the type of feeding. Ascertain differences may partly explain the problems with acidosis in cattle under diet changes conditions.

Metapopulation analysis of the soil mycobiome after application of biopreparations in lettuce cultivation.

Michal Oskiera

Microbiology Laboratory, Research Institute of Horticulture, Poland

Development of Next Generation Sequencing (NGS) technology, increasing the quantity and quality of obtained data, offers many new opportunities for environmental research. High-throughput sequencing allows simultaneous recognition of hundreds of taxa and determination of their proportions in the sample. In addition, the identification of microorganisms without the necessity of their cultivation on microbial media is not limited by the selective growth to particular microbial groups. Availability of the taxonomic information about microorganisms in the environment, their proportions, and the ability to simultaneously analyze multiple samples, allows monitoring of the changes that occur in the environment under the influence of various factors. In the Research Institute of Horticulture in cooperation with WULS, soil fungi were studied after biopreparations application in the field cultivation of lettuce. Mixtures of the T. atroviride and T. harzianum sensu stricto overgrown on organic carrier was added to the soil and planted with lettuce. Representative soil samples were taken from experimental plots prior to application of Trichoderma biopreparations, during lettuce cultivation and after harvest. DNA was isolated from the samples and the fungal ITS1 was amplified. Illumina Miseq platform was used for sequencing, and sequences were analyzed bioinformatically. The study provides new knowledge about the fungi present in the analyzed soil environment. This study was financed as part of the project No UDA-POIG.01.03.01-00-129/09-08.

A systems biology approach to associate lipid metabolism to the immune response after muscle injury.

Nikolaos Giannakis

Department of Biochemistry and Molecular Biology, Medical School, University of Debrecen, Hungary

Muscle injury and regeneration of the damaged tissue is a process orchestrated by polymorhonuclear leukocytes and macrophages. The interaction of these cells relies on lipids that play a key role in the physiology of the inflammatory response either acting as anti—inflammatory compounds or by mediating the resolution of inflammation. To address, how lipid mediators affect inflammation and its resolution in a cell-type specific manner , cardiotoxin injection (CTX) at the tibia anterior skeletal muscle in mice was used as an in vivo model of tissue injury. Moreover, liquid chromatography combined with mass spectrometry was used to obtain the lipid mediator profiles at day 0, 1, 2, 4, or 8 after the cardiotoxin injection. In addition, transcriptomic analysis revealed the expression profiles of genes involved in lipid expression, including COX and LOX biochemical pathways. This systems biology approach showed that arachidonic acid metabolism derived lipids had an elevated profile during the inflammatory response and the tissue returned to homeostasis stage 8 days after the injection. Our data show that Prostaglandins and Leukotrienes, which act at the first stages of inflammation were up-regulated during the 4 first days after the tissue damage. Hitherto, at the beginning of resolution there was a transition in the expression level of lipids with high expression of Resolvins and Lipoxins. Additionally, the transcriptomic analyses, which contribute to alterations in lipid profiles, were mainly derived from leukocytes, and notably from neutrophils. Due to the fact that our data suggest a correlation between transcriptional changes in key lipid metabolism enzymes with the dynamics of lipid profiles, we currently integrate to our study a time-course chromatin accessibility dataset (ATAC-seq) to identify potential genomic regulatory elements and transcriptional regulators of the lipid changes during inflammatory response.

Team-working: Moving forward and enjoying Sciences together

Panagiotis Theodorakis

Institute of Physics - Polish Academy of Sciences

Every step forward in science and our society is a result of team-work, which includes sharing the work-load, emotions, information, etc. This summer school is an excellent example of team-work. In this lecture, we have discussed the benefits of team-work and what constitutes a good team. Furthermore, we discussed the features of a good project leader. However, there are certain qualities that every person of a team should master. For this, we have considered as an example Imperial College London and we discussed what a high-calibre institution expects from its members. Finally, we have underlined the importance of team-work and its broad implications.

Human SNP associations with genetically determined disorders

Polina Avdiunina

Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Russia

Creating genome database and decoding individual genomes of people have been actively carried out all over the world since the decoding of the first human genome in 2000 and up to the present time. In particular, one of the key research areas is the search for genomic variations and the identification of their clinical significance. Most genetic differences are represented by point mutations in the genome single nucleotide polymorphisms(SNP), with some of them being able to determine susceptibility to specific diseases, individual immune responses to pathogens, therapeutic agents, metabolites, toxins, etc. Here we analyzed the effects of such polymorphisms arising in the sites of proteins-compounds binding, and also combine our predictions with ClinVar annotation records. We obtained data on human SNPs from the 1000 genome project. After we selected all small molecular compound-protein structural complexes from PDB (ProteinDataBank) in which the compound was identical or highly similar to FDA-registered drug and drug-like compounds and identified the binding sites of the protein in direct interaction with the ligand. We then combine data about 1000g SNPs and about protein-ligand binding sites to select polymorphisms with the potential impact on the binding properties. Using a docking approach we calculate the difference in the free energy of the drug and wild-type protein interaction and that of the drug with the protein sequencing incorporating the amino acid polymorphism. Finally, we combine our data with clinical annotation using public archives of reports of the relationships among human variations and phenotypes. The search for SNP was based on the database of the project '1000genomes' and 'The Cancer Genome Atlas'. Among the 2504 available human genomes, our approach identified 53 SNPs that probably affect a drug-protein interaction with 30 FDA-approved drugs and 192 SNPs that possibly have this effect on 75 FDA-approved drugs, 390 SNPs probably affecting 165 FDA experimental drugs and 835 SNPs possibly having this effect on 357 FDA experimental drugs. For instance, among the SNPs from the binding sites with serious $\Delta\Delta G$ effect, mutations rs114468011, rs369382075 and rs558267822 were found in the protein Glycogen phosphorylase b (1GPB), which plays a key role in the Glycogen storage disease, type V formation. The results of the annotation require further analysis.

Multi-omics integration reveals new insights in cell biology

Rodrigo García-Valiente

Proteomics Unit, Cancer Research Center, Spain

High-throughtput data offers an unique opportunity to improve our global understanding of biological systems and predict its outcome in pathological situations. Despite of the immense progress in the design and development of -omics technologies, there is still a need for a systematic and comprehensive interpretation in a particular cell situation. In fact, a single -omics does not globally explain all the cell biology and it can offer a biased view. Hence, a comprehensive integration of all the -omics will be a dramatically improvement because it could provide an overview of the biological complexity in a particular situation. Then, here it is presented the design and validation of an algorithm for direct integration of Transcriptomic and Proteomic datasets; in order to generate a proteogenomics characterization and obtain relevant information -functional, spatial and quantitative- of a biological system, understanding its development and behaviour. Concretely, we have focused on the study of B-cell differentiation, focusing on its main five cell populations (centroblasts, centrocytes, naive B cells, memory B cells and plasma cells). For this purpose, using Bioinformatic and Biostatistic techniques to integrate and understand multi-omic data, we achieve a new characterization of its dynamic differentiation and proliferation, giving an insight of how the different B-cells switch between populations during the antigen-dependent differentiation. Giving our successful experience, we are currently developing similar approaches for the identification and validation of diagnostic and prognostic biomarkers for leptomeningeal disease and altered cell signalling pathways in B-cell chronic lymphocytic leukemia.

Keywords: Proteogenomics, Multi-omics integration, B-cell differentiation, Proteomics, Bioinformatics

Identifying High impact Mutations

Rupika Wijesinghe

University of Colombo School of Computing, University of Colombo, SriLanka

Sequencing of human genome/ exome facilitates identifying genetic mutations. Since, majority of disease causing mutations (i.e., SNPs) are in exome, more fo- cus has been given towards sequencing the exome alone. Next Generation Se- quencing (NGS) technology is the currently using sequencing technology due to its high throughput sequencing capability. But, NGS-based mutation detection is also prone to erroneous calls due to sequencing and read mapping errors. There are about 15,000 – 20,000 genetic mutations per individual exome. Hence, it is hard for geneticists to analyze all of them manually.

Considering the limitations in existing literature, this study proposed a computa- tional approach, utilizing supervised machine learning techniques to identify high quality SNPs in exome sequencing, which helps to reduce the large volume of data in to a human manageable amount. Series of machine learning algorithms such as Naïve Bayes, SVM and ANN have been experimented. Data is obtained from the Human Genetics Unit, Faculty of Medicine, Colombo, Sri Lanka and applied a systematic feature engineering process to transform the initial data set in to a model compatible format. The study utilized range of data level, algorithmic and hybrid techniques to over- come the class imbalance problem. For the evaluation, we used wide range of evaluation measures to analyze the performance of each learning algorithm before and after applying class imbalance mitigation techniques. Experimented results indicated that ANN model trained, applying over-sampling and boosting techniques is the best model to identify high quality SNPs in a given sequenced exome of an individual.

Entering the world of science

Tereza Faitova

Applied Informatics, Science Faculty, University of South Bohemia, Czech Republic

My name is Tereza Faitova and I am 20 years old student of Bioinformatics. My study program is based on cooperation of two universities, namely University of South Bohemia and Johannes Kepler University in Linz. I have successfully finished my first year in Czech and now I am finishing second year of my studies, currently in Austria. I am highly interested in life sciences and genetics. At the very beggining, when I met just basics of genetics on highschool, I got absolutely fascinated by all the ongoing processes inside our bodies. At that time, I did not have idea about all the existing methods to analyze them, which is even more exciting. In my future I would like to turn more into medical computational biology and further in my job work together with medical companies. I got to know about the NGS Summer School few minutes ago. I have immediately realized that this is a great opportunity to meet lot of inspirative, motivating and bright people from Bioinformatics field. I am full of enthusiasm for this kind of projects and really appreciate all ideas that bring people together. Looking forward to Warsaw!!

Applied molecular plant physiology to improve food security

Tinashe Chabikwa

School of Biological Sciences, University of Queensland, Australia

My name is Tinashe Chabikwa, I am currently a PhD student at the University of Queensland, Australia studying molecular plant physiology. I am in the final year of my study getting started on writing my thesis as well as preparing manuscripts for publication, the first of which I am planning to submit by the end of July 2017. I am currently studying the role of hormones and sugar signalling in shaping plant architecture. Plant architecture, which is determined by shoot branching, has major implications on crop and biomass productivity, which directly impacts food security. The major part of my project involved conducting a comparative transcriptome profiling of dormant pea axillary buds on intact plants and axillary buds activated by removing the shoot tip. In this experiment, I identified important transcription factors that trigger axillary bud outgrowth. Further experiments through Chip-Seq (Chromatin Immunoprecipitation Sequencing) and TARGET (Transient Assay Reporting Genome- wide Effects of Transcription factors) will enable us to discover potential targets of these transcription factors for further characterization. My thesis has provided an opportunity to develop wet-lab and bioinformatics skills that, if given an opportunity, I am willing to further develop during your prestigious workshop.

My research interests include the applied molecular biology to contribute towards solving critical challenges faced by farmers, such as developing cultivars tolerant to pests and diseases and managing post harvest physiological deterioration in field crops, vegetables and fruits. I am particularly interested in applied RNA-Seq based transcription profiling to identify candidate genes that influence important trait. This allows for the manipulation of these genes by approaches such as selecting hypomorphic alleles, gene silencing or introgression of these alleles through traditional breeding or developing cisgenic plants. I am also interested in chromatin remodelling, histone modification and their effect on plant adaptation to adverse biotic conditions such as drought and salt stress.

TMicroRNA sponges: High-throughput approach for in silico design and testing.

Tomas Barta

Department of Histology and Embryology, Faculty of Medicine, Masaryk University, Czech Republic

MicroRNA (miRNA) sponges are RNA transcripts containing multiple tandem high affinity binding sites that bind and sequester specific miRNAs to prevent them from interaction with their target mRNAs. Due to their high specificity and strong inhibition of target miRNAs, these molecules have become increasingly applied in miRNA loss-of-function studies and also tested in miRNA-based therapies. However, improperly designed miRNA sponge constructs may sequester off-target miRNAs, which may lead to false-positive results and/or off-target effects, therefore it has become increasingly important to develop a tool for in silico miRNA sponge construct design and testing. Here we introduce a novel, user-friendly, and freely available tool for in silico design and testing of miRNA sponges - miRNAsong: microRNA SpONge Generator, located at http://www.med.muni.cz/histology/miRNAsong/. This tool allows to generate miRNA sponge sequences specific to a target miRNA, miRNA family and/or cluster and test generated sponge sequence for potential off-targets in 219 species covering 35,828 mature miRNA sequences. Furthermore, we experimentally verified the functionality of our tool. Using miRNAsong we generated and optimised miRNA sponge for miR-145 inhibition. We cloned the generated sequence into an inducible expression vector and established cell lines that strongly inhibit miR-145.

Employing differential gene co-expression network analysis to identify pathways impaired in ageing

Veronika Kedlian

studying at Taras Shevchenko National University of Kyiv/ doing internship at European Bioinformatics Institute (EMBL-EBI), studying in Ukraine / doing an internship in UK

Ageing is defined as a progressive deterioration and loss of integrity of all the systems within an organism that occur with the passage of time. The process of ageing has multiple manifestations in cell types, tissues and organs of the human body. Consequently, it is associated with a very diverse range of changes on the molecular level. As a result, employing conventional RNA expression analysis on old versus young samples provide us with a long list of genes that change expression during ageing but doesn't provide an idea of general mechanisms involved. This motivates our study – the differential co-expression analysis between young and old human tissues. Co-expressed genes are defined as a group of genes, that have a high correlation in the expression pattern, as a result, these genes are likely to participate in one biological function and represent the members of the same pathway. We plan to perform a differential co-expression gene network analysis using Weighted Gene Correlation Network Analysis (WGCNA) on young vs old human RNA-sequencing datasets. Identified gene modules will be used to perform pathway and transcription factor binding sites enrichment that will potentially allow us to distinguish coordinated cellular responses to ageing from random molecular changes that are associated with ageing phenotypes. Identification of pathways with decreased co-expression in ageing could guide our search for the anti-ageing and pro-longevity drugs. We plan to use published drug perturbed gene expression profiles to select drugs, which target pathways with decreased co-expression in age.

Microbiome composition change influenced by rotavirus infection in humans

Vladyslav Dembrovskyi

ESC 'Institute of Biology and Medicine', Taras Shevchenko National University of Kyiv, Ukraine, Ukraine

In the spotlight - human microbiome. Bacteria influence human health, especially having impact on immune system and metabolism. It appeared that the relationships between bacteria in microbiome and human organism are difficult to trace. You need to sequence the whole metagenome from human intestine, gain an array of data and process them with state-of-the-art software tools in order to get significant results. Knowledge derives from tons of data – this is what bioinformatics is used for. There is simply no other way to carry out this kind of research without bioinformatics tools, and I am excited about that. If we understand the key connections between human microbiota and human health, we can learn how to change bacterial composition in a proper way. That will bring us to the era of personalized medicine. However, for these purposes we need great computational power and methods because everything depends on the data analysis and solving data riddles. I have already worked a bit on research of rotavirus in human intestine. Rotavirus causes diarrhea and depresses the microbiome, arousing even more complications. Serial analysis has demonstrated that curing with probiotics helps to rehabilitate microbiota and alleviates the clinical course this way. For this analysis, NGS and High Throughput Sequencing Data Analysis were needed. It has taken me several months to learn these methods. It was a challenging and a fascinating task. As a third-year undergraduate student, I have learnt a bit, so I want to learn and be able to conduct much more. I want to apply appropriate methods, understand design of the experiment analysis clearly and take part in its development, work not only at the level of taxonomy but also at the functional level because different bacteria can have similar functions. For these purposes I need to know how to analyze whole sequence metagenomic data, not only by 16s RNA, as I've done before. Indeed, I am also interested in the field of bioinformatics which investigates the early stages of animal evolution. Required data has been sunk in the ocean of genomes of diverse organisms from all over the world. These data has already been acquired, we only need to get it from the bottom.

Novel molecular disease monitoring tools for clear cell renal cell carcinoma(ccRCC)

Weronika Majer

Laboratory of High Throughput Technologies
Institute of Molecular Biology and Biotechnology
Faculty of Biology
Adam Mickiewicz University , Poland

Clear cell renal cell carcinoma (ccRCC) is the most common kidney malignancy worldwide. Late detection and diagnosis is major reason for low recovery rate of ccRCC patients. Nucleic acids such as cell-free DNA (short DNA fragments of released to the bloodstream by tumor cells) might potentially serve as new biomarkers. It has been shown that cell-free DNA concentrations are higher in cancer as compared to healthy controls. The aim of the project is to detect chromosomal aberrations present in cfDNA characteristic for ccRCC tumors using new generation sequencing techniques (NGS). CfDNA were isolated from 3ml of plasma and 4ml of urine from ccRCC patients and healthy individuals using commercial kits. Libraries were prepared with NEBNext system and sequenced was performed with HiScanSQ. Results were generated with bioinformatics tool Wisecondor, and next compared to previously obtained results of cytogenic analysis of tumor samples using GWAS. Concentration of cfDNA were higher in plasma of ccRCC patients as compared to controls, the quantity of cfDNA was higher in plasma than in urine. The mean concentration of patients cfDNA from plasma was 7.09 ng/µl and 0.885 ng/µl from urine with average fragments size 150bp. Comparison of cfDNA and tumor revealed discrepancies in number and quality aberrations. Results showed cytogenic analysis of cfDNA might enable more in depth investigation of tumor aberration. Therefore there is a necessity to study relations between cytogenic changes and clinical outcome in larger cohort.