# NGSchool 2023

## Advances in Computational Biology

# Book of Abstracts

16-24 September 2023
Otwock, Poland

Visegrad Fund

# NGSchool2023

# Advances in Computational Biology

## Book of Abstracts

16-24 September 2023, Otwock, Poland

# NGSchool2023 Team (NGSquad)

| | |
|---|---|
| Project coordinator: | Karolina Sienkiewicz, Weill Cornell Medicine, Cornell University, US |
| Organising committee: | Gabriel Deards, Weill Cornell Medicine, Cornell University, US<br>Szymon Grabia, Medical University of Lodz, PL<br>Marcin Kaszkowiak, Institute of Hematology and Blood Transfusion, PL<br>Zuzanna Nowicka, Medical University of Lodz, PL<br>Dalma Müller, Semmelweis University, HU<br>Vishma Pratap Sur, Czech Academy of Sciences, CZ<br>Diana Sharysh, Aarhus University, DK<br>Karolina Sienkiewicz, Weill Cornell Medicine, Cornell University, US<br>Urszula Smyczyńska, Medical University of Lodz, PL<br>Wiktoria Wilman, NaturalAntibody S.A., Szczecin, PL |
| Book of abstracts editing: | Marcin Kaszkowiak<br>Zuzanna Nowicka<br>Diana Sharysh<br>Karolina Sienkiewicz |
| Conference logo designer: | Matylda Gurne |

# Preface

Dear attendees,

It is our great pleasure to welcome you all in Otwock for the 6th edition of the NGSchool's flagship event - **NGSchool 2023: Advances in Computational Biology**.

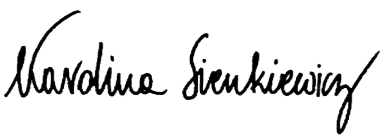
This year's Autumn/Summer School is dedicated to the latest advancements in computational biology and will cover such topics as single-cell data analysis, whole genome sequencing, spatial transcriptomics, multi-modal data integration. Population genetics and cancer evolution. In our commitment to **fostering knowledge and accessibility of scientific training**, we will be offering all course materials and recorded lectures online, free of charge.

We are proud to present an **outstanding scientific program** that includes eleven lectures, nine workshops, participant debates and four parallel hackathons. Moreover, all NGSchool participants will have a chance to **present their research** and areas of interest during a dedicated session. We trust that you will find the intensive training rewarding and have an opportunity to partake in activities planned as a part of our **extensive social program**. This is an invaluable opportunity for **early-career researchers** to interact with **leading academics and professionals**, both from within and outside the V4 and CEE region.

We would like to extend our heartfelt gratitude to **Visegrad Fund** for their generous sponsorship of this year's NGSchool. Their continued support has been invaluable. Our deepest appreciation also goes out to **all our sponsors and partners for their unwavering support.**

We wish you an enlightening and enjoyable experience at NGSchool2023!

Best regards,



Karolina Sienkiewicz

NGSchool Society President
NGSchool2023 Project Coordinator

# Table of contents

# NGSchool Society

The previous editions of NGSchool were a big success, and their popularity motivated us to organise this summer school on an annual basis. To regulate formal aspects of our activity and secure flawless organisation of future events, we established NGSchool Society in September 2018. This Autumn School is also a culmination of the 5th year of the Society's work.

The goal of the Society is to promote and support science. We do that by organising scientific events (and securing funding for such) and cooperation with scientific institutions and other scientists. While there are many great hands-on courses in bioinformatics, they tend to be expensive, especially for researchers from Central & Eastern European Countries. We decided to make a difference!

Every year we adapt the course programme accordingly to the new trends and developments in sequencing technologies. We try to address new challenges arising in computational biology and high-throughput data analysis. We will keep inviting experts in relevant fields and improve based on feedback from past editions. We are doing our best to secure funding for every course we're organising, making it accessible to everyone.

We are always open to new volunteers! Do you want to shape the upcoming editions of NGSchool? Do you want to help promote science? Do you want to organise excellent and affordable training in computational biology for young, talented scientists? If your answer is yes (or even maybe) to any of the questions - don't hesitate to contact us. We are looking forward to welcoming you to the team! :)

# Sponsors

## Visegrad Fund

The International Visegrad Fund is a donor organization, established in 2000 by the governments of the Visegrad Group countries (V4) — Czechia, Hungary, Poland and Slovakia. Its main purpose is to promote the development of closer cooperation among the Visegrad Group (V4) countries by supporting grant projects in the fields of common cultural, scientific and educational projects, youth exchanges, cross-border cooperation and tourism promotion, and by awarding scholarships and artist residency programmes. The Visegrad Fund is the main sponsor of this edition of NGSchool.

## The Company of Biologists

The Company of Biologists is a not-for-profit publishing organisation dedicated to supporting and promoting research and study across all branches of biology. The charity is run by experienced, senior scientists from a range of life science and clinical research backgrounds.
The focus of their activities is:

- publishing leading peer-reviewed journals,
- facilitating scientific meetings and communities,
- providing travel grants for young researchers,
- supporting and funding research societies.

# Partners

The International Institute of Molecular and Cell Biology in Warsaw (IIMCB), established in 1999, aims to carry out high-quality research in molecular biomedicine, to implement modern biotechnology, as well as to teach and popularize molecular biology and medicine. Research topics at IIMCB cover the wide area of structural biology, bioinformatics, computer modeling, molecular and cell biology, neurobiology, cancer biology, and developmental genomics.

Faculty of Mathematics, Physics and Informatics at Comenius University in Bratislava, founded in 1980, is consistently ranked first in the group of natural sciences in the ranking of faculties in Slovakia. The Faculty offers complex university-level education in all areas of mathematics, physics, and computer science. In collaboration with other units, the institution provides several interdisciplinary study programs (e.g. Biomedical Physics in collaboration with the Faculty of Medicine, and Bioinformatics in collaboration with the Faculty of Natural Sciences).

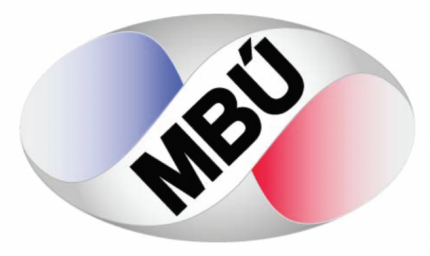
The Institute of Biotechnology CAS, v. v. i. was established in 2008 and since then it has focused on excellent basic research in biomedical molecular biology with prospective transfer of biotechnological methods and tools to human and veterinary medicine or other important areas of human activity. The BIOCEV project is realized as a part of the operational program Research and Development for

Innovations (OP RDI) by six institutes of the Czech Academy of Sciences and two faculties of the Charles University.

Semmelweis University in Budapest, has been a leading medical higher education institution, healthcare provider, and center of research excellence in Hungary and Central Europe for over 250 years. Its mission based on the integrity of education, research and development, and patient care has made it a regional center of excellence in the field of health sciences. The University offers academic programs that provide extensive and solid theoretical knowledge as well as competitive practical skills in medicine, dentistry, pharmaceutical and health sciences, and conductive education.

The Albert Szent-Györgyi Medical School at the University of Szeged stands as a prestigious institution dedicated to the triad of education, research, and patient care, embodying a rich tradition upheld since medical education commenced in Szeged in 1921. Named in honor of the Nobel Laureate Albert Szent-Györgyi, who conducted groundbreaking physiological research including the isolation of vitamin C at this university, the Medical School has thrived, becoming the largest among the university's twelve faculties.

The Institute of Biology at the Eötvös Loránd University has twelve departments covering wide areas in the life sciences. More than eighty faculty members work in close collaboration with over one hundred researchers who are associated with the institute through various fellowships, grants, and collaborative industrial projects.




The Institute of Microbiology of the CAS represents the largest scientific body extensively exploring life cycles, molecular mechanisms and regulatory systems of various microorganisms such as bacteria, yeast, fungi and algae as well as mammalian cell lines with respect to basic research questions as well as their prospective practical exploitation in medicine and industry. The main research interests of our Institute represent cellular and molecular microbiology, genetics and physiology of microorganisms and their resistance to antibiotics, production of microbial metabolites and their biotransformation, and grading up production strains by genetic modifications etc.

# Program

| | Meals / Breaks | Social | Organizational | Discussion Panel | | Workshops | Lectures | Small Group Project | |
|---|---|---|---|---|---|---|---|---|---|
| | **16.09** | **17.09** | **18.09** | **19.09** | **20.09** | **21.09** | **22.09** | **23.09** | **24.09** |
| | **SAT** | **SUN** | **MON** | **TUE** | **WED** | **THU** | **FRI** | **SAT** | **SUN** |
| | **DAY 0** | **DAY 1** | **DAY 2** | **DAY 3** | **DAY 4** | **DAY 5** | **DAY 6** | **DAY 7** | **DAY 8** |
| 07:30 - 09:00 | | Breakfast | Breakfast | Breakfast | Breakfast | Breakfast | Breakfast | Breakfast | Breakfast |
| 09:15 - 10:45 | | **Multi-modal single-cell analysis** *Fabiola Curion* | **Introduction to Deep Learning & Bayesian Statistics** *Viktor Petukhov* | **Whole genome sequencing** *Kerstin Haase* | **Genomic Epidemiology and GWAS analysis** *Maxim Freydin* | **Cell Identity, cell communication and spatial transcriptomics** *Vitalii Kleschevnikov* | Group Project | Group Project | Presentation of projects |
| 10:45 - 11:15 | | Coffee break | Coffee break | Coffee break | Coffee break | Coffee break | Coffee break | Coffee break | Coffee break |
| 11:15 - 12:45 | REGISTRATION | **Multi-modal single-cell analysis** *Fabiola Curion* | **Introduction to Deep Learning & Bayesian Statistics** *Viktor Petukhov* | **DNA copy number analysis** *Kerstin Haase* | **Genomic Epidemiology and GWAS analysis** *Maxim Freydin* | **Spatially mapping cell types** *Vitalii Kleschevnikov* | Group Project | Group Project | Presentation of projects & closing remarks |
| 13:00 - 14:15 | Technical set-up, preparation for workshops | Lunch | Lunch | Lunch | Lunch | Lunch | Lunch | Lunch | Lunch |
| 14:30 - 16:00 | Lunch | **Epigenetics & epigenomics** *Silvia Perez-Lluch* | **Natural Language Processing** *Tomasz Kopacz* | **Cancer evolution introduction & modeling** *Gustav Lindwall* | Networking | **Advanced Spatial Transcriptomics** *Viktor Petukhov* | Current topics in science: participant debates | Group Project | **Group project mentors:** Diana Sharysh, Vitalii Kleschevnikov, Karolina Sienkiewicz, Viktor Petukhov |
| 16:00 - 16:30 | **Introduction to single-cell analysis** *Diana Sharysh* | Coffee break | Coffee break | Coffee break | | Coffee break | Coffee break | Coffee break | |
| 16:30 - 18:00 | Coffee break | **Epigenetics & epigenomics** *Silvia Perez-Lluch* | **Natural Language Processing** *Tomasz Kopacz* | **Cancer evolution analysis** *Glada Fiandaca* | | **Advanced Spatial Transcriptomics** *Viktor Petukhov* | Current topics in science: participant debates | Group Project | |
| 18:15 - 19:45 | **Introduction to single-cell analysis** *Diana Sharysh* | Dinner | Dinner | Dinner | Dinner | NGSupper | Dinner | Dinner | |
| | Dinner | Short participants' presentations | NGSchool Society Meeting | Networking | | | | NGSchool General Assembly | |
| | Short participants' presentations | | | | | | | | |

# Selected speaker's abstracts

## Probabilistic models to resolve cell identity and tissue architecture

Vitalii Kleshchevnikov
Wellcome Sanger Institute, Cambridge, UK

Cell identity drives cell-cell communication and tissue architecture and is in return regulated by cell extrinsic cues. Cell identity is determined by the combination of intrinsic developmentally established transcription factor use (TF) and constitutive as well as cell communication-dependent TF activities. We developed two probabilistic models that advance the understanding of these processes using single-cell and spatial genomic data.

Spatial transcriptomic technologies promise to resolve cellular wiring diagrams of tissues in health and disease, but comprehensive mapping of cell types in situ remains a challenge. We present cell2location, a Bayesian model that can resolve fine-grained cell types in spatial transcriptomic data and create comprehensive cellular maps of diverse tissues. Cell2location accounts for technical sources of variation and borrows statistical strength across locations, thereby enabling the integration of single cell and spatial transcriptomics with higher sensitivity and resolution than existing tools. We assess cell2location in three different tissues and demonstrate improved mapping of fine-grained cell types. Collectively our results present cell2location as a versatile analysis tool for mapping tissue architectures in a comprehensive
manner.

Cell identity and plasticity is regulated by a combinatorial code mediated by transcription factors and the cell communication environment. Systematically dissecting how the regulatory code robustly defines the vast complexity of cell populations across tissues is a long-standing challenge. Measured using the assay for transposase-accessible chromatin with sequencing (ATAC-seq), DNA accessibility provides a readout of intermediate gene regulation steps at single-cell resolution. Existing methods address one or several subproblems of modelling DNA accessibility. We are missing an end-to-end mechanistic model that represents all steps of the biological process, that generalises to both new DNA sequences and TF abundance combinations and can simultaneously characterise hundreds to thousands of cell states observed in single-cell genomics atlases. To address this, we formulated cell2state, a mechanistic end-to-end probabilistic model of TF recruitment to a chromatin locus and downstream TF effect on DNA accessibility. Cell2state outperforms the state-of-the-art deep learning models and enables simulating the possible chromatin states given TF abundance of source cell types.

## Statistical inference on interacting particle systems with applications to cancer biology

Gustav Lindwall
Chalmers University of Technology, Gothenburg, Sweden

Interacting particle systems is a mathematical framework which allows for condensed modelling of complex phenomena undergoing both deterministic and random dynamics. While there are several ways to formulate an interacting particle system, this thesis focuses on modelling such dynamics using stochastic differential equations (SDE:s). The SDE framework was constructed in order to describe the in vitro population dynamics of cancer cells.

This thesis introduces the necessary mathematical and biological context, and formulates a model that is subsequently studied in the appended research papers. In the first of three papers, we introduce a novel method of inferring the diffusive properties in such systems based on a higher order numerical approximation of the underlying stochastic differential equations. In the second paper, we model the effect of cell-to-cell interactions, and conduct inference on this model using microscopy data. The third and last paper concerns modelling how the spatial distribution of the cell population affects the cell division rate, and apply our theoretical results to microscopy data.

Put together, the three papers present a cohesive package for modelling and parameter inference that can be applied to population data that is spatial and time-resolved.

## Epigenetics: From the lab to the computer

Sílvia Pérez-Lluch
Centre for Genomic Regulation

The regulation of gene expression is essential to ensure the proper cell division and differentiation. In the living organism, the differences in gene expression between cells allow for the correct specification of tissues and, in the end, to the correct development of the organism. The mechanisms by which genes undergo activation or silencing in a particular cell type comprise recruitment of specific transcription factors, methylation of the DNA and histone post-translational modifications.

Histone marks, in particular, have been assigned a central role in the determination of cell commitment during differentiation by allowing for establishment and maintenance of particular gene expression patterns. However, the mechanism by which histone marks could influence gene expression is still uncertain. Strikingly, recent studies suggest that the role that chromatin would play in regulating transcription could be not as causative as previously thought. The tri-methylation of lysine 4 of the histone H3 tail, for instance, has been canonically associated to the transcription start site (TSS) of active genes. However, very recently, we have seen that genes that are expressed transiently for a short period of time during the fruit fly development do not present this histone modification. Also the mono-methylation of lysine 4 has been widely associated to enhancer regions; however, different groups have demonstrated that it is the enzyme responsible for the methylation deposition but not the histone mark itself that it essential for the enhancer function. These observations argue against a causative role of histone marks in regulation of gene expression, and point to a more consequential relationship.

In this framework, the overarching goal of our research line is to untangle the relationship between histone post-translational modifications and the regulation of gene expression.

## Gene expression changes in plaque smooth muscle cells during LDL lowering in transgenic pigs with atherosclerosis

Diana Sharysh
Aarhus University, Aarhus, Denmark

Lowering of low-density lipoproteins (LDL) can lead to plaque regression in humans and reduce the risk of clinical events. Resolution of inflammation is involved, but the impact of LDL lowering on other cell types remains poorly understood. Smooth muscle cells (SMCs) give rise to most cells in advanced atherosclerotic plaques. In the present study, we studied gene expression changes in SMC subtypes during LDL lowering in a pig model of atherosclerosis.

Methods

PCSK9 gain-of-function mutant minipigs were fed a high-fat diet for 12 months to induce lesions and then randomized to continue on high-fat diet (n=6) or low-fat diet supplemented with microsomal triglyceride transfer protein (MTP) inhibitor to achieve efficient LDL lowering (n=5). After 3 months, abdominal aorta plaques were harvested and analyzed by single-cell RNA sequencing. Differentially expressed genes were detected by pseudo-bulk analysis to control for false positive findings. Data analysis was performed in R using Seurat toolkit.

Results

ScRNA-seq analysis revealed 6 main clusters among SMCs and related mesenchymal cell types in plaques: medial SMCs, contractile plaque SMCs, modulated SMCs, proliferative SMCs, fibroblasts, and pericytes. Contractile plaque SMCs and modulated SMCs were the most affected by the treatment with 26 and 27 differentially expressed genes, respectively. LDL lowering led to down-regulation of genes associated with glycolysis, hypoxia, extracellular matrix production, and the type 1 interferon response.

Conclusion

LDL lowering causes clear changes in gene expression in the major subtypes of plaques SMCs that may be involved in the regression of plaque and the protection against clinical events.

## Changes in circulating cell-free DNA as a biomarker of immune response to short-duration spaceflight

Karolina Sienkiewicz
Weill Cornell Medicine, Institute for Computational Biomedicine, New York, USA

The concentration of cell-free DNA (cfDNA) and its molecular profile are emerging biomarkers with great clinical and research potential, which can provide valuable insights into the dynamic response of organisms to environmental or disease-related stress factors. Here, we focus on the systemic immune system response to physiological stress related to microgravity, radiation exposure, and the other unique environmental conditions of short-duration spaceflight.
As a part of the Space Omics and Medical Atlas initiative, we profiled the cfDNA of a cohort of astronauts from SpaceX Inspiration4 mission, comparing pre-flight baseline, recovery and longitudinal responses. We present a comprehensive cfDNA analysis pipeline which includes (1) a comparison of sample-wise and chromosome-specific fragment size distribution, (2) an assessment of enrichment in cfDNA fragments originating from different tissues based on estimated nucleosomal footprinting and inferred gene expression, as well as (3) cfDNA variant calling and annotation.
Building on the results from the NASA Twins Study, we inspected the previously reported potential biomarkers of interest for long-duration spaceflights, such as the fraction of mitochondrial cfDNA relative to chromosomal cfDNA in plasma. Comparison of the deconvoluted tissue/cell type of origin profile for circulating cfDNA fragments with either reference tissue-specific expression signatures from the Human Proteome Map or individual astronaut expression profiles from peripheral blood (single-cell RNAseq of PBMC) revealed that the most represented sequences are of hematopoietic origin. We noted an increase in the cfDNA originating from immune cells after return to Earth which may reflect a delayed spaceflight-related immunological response. Moreover, our preliminary results highlight the non-invasive monitoring potential of cfDNA may be extended toward circulating epigenomic biomarkers, which could also be extremely valuable in a disease setting where the progression is associated with rapid epigenetic changes.

# Participants' abstracts

## Applying Machine Learning, Modeling, and Automated Image Processing to Enhance Single-Molecule Gene Expression Experiment Design.

Luis Aguilera de Lira
Colorado State University, Fort Collins, Colorado, United States of America.

Super-resolution fluorescence microscopy has been used to study gene expression at the single-molecule level. However, due to its resource-intensive, time-consuming, and costly nature, its application is limited to a few samples and genes. To overcome these limitations, we have developed a novel approach that combines machine learning, mechanical models, and automated image processing. Our computational pipeline classifies mRNA species based on their fluorescence intensity signals, enabling discrimination of different types of mRNA in the same cell. This approach promises to expand the use and applicability of super-resolution fluorescence microscopy and provides new avenues for investigating gene expression.

## Hepatocarcinoma individual transcriptome analysis.

Ivan Kaluzhskyi

Hepatocellular carcinoma is the most common primary liver cancer. In our study we analyze differential expression of bulk RNAseq of hepatocarcinoma, surrounding healthy tissue and cell lines and analyze single cell RNAseq to perform decomposition of bulk transcriptomes. Purpose of this study is to find genes responsible for interaction between cancer cells and surrounding healthy tissue.

## On the metabolic memory formation, plasticity, and the epigenetic consequences of type 1 diabetes in children

Jędrzej Chrzanowski
MUL, Dept. of Biostatistics and Translational Medicine, Lodz, Poland

Type 1 diabetes (T1D) is the most common metabolic disease in childhood. It is caused by the destruction of pancreatic beta cells, which are responsible for the production of insulin – an essential hormone in blood glucose levels regulation. Treatment of T1D demands life-long subcutaneous insulin supplementation to control blood glucose levels close to those observed in healthy peers and thus circumvent the development of severe, long-term complications.

However, not only treatment intensity but also its timing are critical for the development of long-term complications. Extensive clinical trials reported that better glycemic control in the first years after diabetes diagnosis was associated with a lower risk of complications development, the effect which persisted for over 16 years after the trial's completion despite the diminished difference in blood glucose levels on follow-up. This phenomenon, dubbed "metabolic memory", remains a subject of extensive resesarch.

A possible explanation of metabolic memory is connected with epigenetic modifications. These modifications affect how the genetic code (DNA) is "understood" and acted upon by the different cells in our organism. Thus, even though all cells possess the same DNA data, epigenetic changes modify the availability of its fragments and activate or suppress the production of specific proteins.

Early exposure to high glucose was associated with epigenetic changes that could persist, resulting in altered cell function and increasing the future risk of complications.

In my research, I focus on how the DNA methylation pattern (which fragments are methylated and how much) changes during the first years of T1D. By close cooperations with diabetes refernce centre, I have a chance to focus on the group of patients at risk of disrupting the proper formation of the metabolic memory due to unsatisfactory glycemic control in those crucial first years of disease. The 12 month longitudinal observation of those patients allows me to evaluate the DNA methylation changes due to their clinical improvement. We hope to elucidate the methylation plasticity in this group, and investigate the potential reversibility of methylation-mediated complications risk.

The results of this project will allow us to better understand the mechanisms and dynamics of metabolic memory formation and identify how this process can be modulated. The projects is funded under PRELUDIUM BIS 4 NCN grant no. 2022/47/O/NZ5/00683.

## Exploring biomarkers of liquid biopsy for effective early diagnosis of cancer

Peter Hunyadi
Clinomics Europe Kft., Budapest, Hungary

Graduated from Bioinformatics MSc at the University of Vienna and afterwards I was a Data Analyst Intern at Roche Diagnostics in Penzberg, Germany.
Since, I am working for Clinomics Europe in Budapest, Hungary. We are involved with liquid biopsy cancer research and molecular genetic diagnostics. With liquid biopsy, we are studying possible applications of cell-free DNA and circulating tumor cells for early cancer detection from non-invasive blood tests.
My main scientific interests are: computational oncology, multi-omics of single cells, machine learning & statistics

## Panoramic visual statistics shape retina-wide organization of receptive fields

Jan Svatoň
CEITEC MU, Brno, Czechia

Statistics of natural scenes are not uniform—their structure varies dramatically from ground to sky. It remains unknown whether these nonuniformities are reflected in the large-scale organization of the early visual system and what benefits such adaptations would confer. Here, by relying on the efficient coding hypothesis, we predict that changes in the structure of receptive fields across visual space increase the efficiency of sensory coding. Using the mouse (Mus musculus) as a model species, we show that receptive fields of retinal ganglion cells change their shape along the dorsoventral retinal axis, with a marked surround asymmetry at the visual horizon, in agreement with our predictions. Our work demonstrates that, according to principles of efficient coding, the panoramic structure of natural scenes is exploited by the retina across space and cell types.

## Mathematical and Computational Modelling of Extinction Therapy for Cancer

Srishti Patil
IISER, Pune, India

Despite the evolutionary nature of the disease, cancer therapies rarely exploit concepts from ecology and evolutionary biology
to deal with the emergence of resistance in malignancies – a big hurdle for an effective treatment response. Evolutionary therapies for cancer understand malignancies as adapting populations under Darwinian selection. Extinction Therapy (ET), a novel evolutionary therapy, aims for complete eradication of the tumour. It fights the emergence of resistance with the smart and effective use of drugs/treatments to exploit the vulnerability of a small or declining population using multiple strikes (in the form of drugs, surgery, etc). In other words, extinction therapy "kicks the tumour while it's down". Our work models ET analytically and computationally using evolutionary rescue theory and stochastic simulations to understand the behaviour of a cancer population undergoing ET. We also perform predictive mathematical modelling to aid the design and analysis of future experiments in ET. We find that the timing of subsequent strikes (after the primary therapy) is a very important determinant of the extinction probability. We calculate the optimal timing for these strikes and show how it changes with other model parameters. This work is one of the first few models of ET and sets the stage for future analytical and computational work in the field.

## Enzymes of bacteriophage origin as an alternative strategy of plant pathogens biocontrol

Yuliia Faidiuk
Hirszfeld Institute of Immunology and Experimental Therapy, PAS (Wroclaw, Poland); Taras Shevchenko National University of Kyiv and (Kyiv, Ukraine); Institute of Microbiology and Virology NAS of Ukraine (Kyiv, Ukraine)

Among the control strategies for plant diseases caused by bacterial pathogens is the use of antibiotics. This in turn contributes to the global threat of the XXI century: the emergence of multi-drug-resistant bacteria. Application of bacteriophages, viruses infecting bacteria, and phage-based products, enzybiotics, is an alternative and promising approach. One of the survival strategies of bacterial pathogens is the production of exopolysaccharide (EPS): a major component of capsule and biofilm that efficiently protects them against environmental factors, antibiotics, as well as allows bacteria to avoid host defense mechanisms. In the course of plant infection EPS can be heavily overproduced by a pathogen so eventually it obstructs plant vascular system vessels leading to wilting and plant death. Obtaining recombinant phage-coded enzymes, exopolysaccharide depolymerases (EPSDs), naturally able to hydrolyze bacterial EPS, is of great importance. Once devoid of EPS, the pathogen decreases its virulence and reduces its resistance to external factors. Recently we have sequenced the genomes of 3 broad-host-range bacteriophages E105, TT10-27, and KEY, active against plant-infecting bacteria of Erwinia and Pantoea genus, including a fire-blight disease-causing agent, Erwinia amylovora. Phage E105 represents a novel species within Autographiviridae family. Its genome of 43,856 bp features 54.46 % GC content, codes for 54 CDSs and 1 tRNA gene. Phage TT10-27 represents a novel species and novel genus within Schitoviridae family. 74,143 bp DNA with 46.7 % GC pairs codes for 86 CDSs and 4 tRNA genes. Phage Key represents a novel species within Demerecviridae family, is characterized by a genome size of 115,651 bp with 39.03% GC content, coding for 182 CDSs and 27 tRNA genes. In all 3 genomes EPSD genes were detected, coding for proteins of 93.50, 95.5, and 111.5 kDa, respectively. Relevant genes were cloned and expressed in E. coli system. When applied on the lawn of host bacteria in conditions that stimulate EPS-overproduction, all 3 EPSDs were efficiently degrading the EPS of E. billingae and P.agglomerans, however, did not reveal a substrate specificity against E.amylovora EPS. This important observation will be used to improve the strategy for the selection of promising EPSD enzymes. Stability testing, structural studies, and further substrate specificity testing will be conducted to propose obtained EPSDs as antimicrobial formulations.

## eDAVE - extension of GDC Data Analysis, Visualization, and Exploration Tools – a new platform for integrated methylomics and transcriptomics data analysis

Jan Bińkowski
Pomeranian Medical University in Szczecin, Independent Clinical Epigenetics Laboratory, Szczecin, Poland

Epigenetic changes, such as aberrations of DNA methylation, and alterations of gene expression induced by those changes are increasingly shown to play a key role in the pathology of various diseases. Moreover, it becomes clear that biomarkers based on methylation changes can significantly improve clinical management of non-communicable diseases e.g., cancer.
Consequently, with exponentially growing epigenetic research, large amounts of methylomics and transcriptomics data are generated and deposited in publicly available repositories. These repositories are a valuable research resource but deposited data records are frequently unstructured and/or poorly integrated. This makes mining of these datasets challenging, especially for the researchers without bioinformatics expertise and appropriate IT infrastructure necessary for big data processing.
To address these challenges, we have developed eDAVE, a platform that integrates over 11000 whole methylome (Infinium Methylation Assay, Illumina) and transcriptome (RNA-seq) profiles for 204 tissues and cancer types. The platform enables a user without prior bioinformatics expertise to perform both exploratory research and validation of studies results based on a large number of independent transcriptomics and methylomics datasets. Moreover, the platform is useful for fast assessment of methylation or expression biomarker candidate specificity.
Web app implemented in Python, with all major browsers supported. Freely available on the web at: https://edave.pum.edu.pl/

## Long non-coding RNA profiling in non-alcoholic fatty liver disease and cancer

Angelos Kozonakis
University of Crete - Faculty of Medicine, Heraklion, Greece

The daily life of many cultures promotes the consumption of unhealthy foods high in calories, resulting in increasing obesity phenomena leading to non-alcoholic fatty liver disease (NAFLD). Particular attention has been paid to the study of histopathological changes in the disease and the identification of potential biomarkers. At the same time, new information on long non-coding RNA (lncRNAs), highlight their importance at various levels, such as chromatin remodeling and regulation of gene transcription, in synergy with RNA-binding proteins (RBPs). Data from whole RNA sequencing (RNA-Seq) of patient subjects were published (10.1126/scitranslmed.aba4448) and their analysis focused on the role of protein-coding genes (PCGs).
To present the transcriptional profile from another perspective, the same data were analyzed with focus on lncRNAs. Differential expression analysis and automated machine learning showed several lncRNAs genes to be expressed statistically significantly differently, with two genes standing out the most, GAS6-AS1 and MIR4435-1HG. New data characterizing RBPs helped to identify 5 important genes whose expression is significantly differentiated. New data characterizing RBPs (10.1136/gutjnl-2021-325109) have also helped in the identification of important RBP genes whose expression is significantly differentiated in disease progression (ENO1, ZFP36, NOB1, PCBP2, SRSF10).
Since NAFLD may also progress to hepatocellular carcinoma (HCC), there is a lot of interest in analyzing clinical cancer data in a large scale. The purpose is to extract the splicing efficiency of lncRNAs. To do so, we will use hundreds of samples from the TCGA,Äôs RNA-seq data for both liver (HCC) and breast cancer, while integrating additional multiomics data such as expression and splice site mutations. Our goal is to build regulatory networks to infer disease progression, cancer subtype and association with drug resistance. Facilitating and managing large scale data will become easier by creating automated pipelines using Snakemake.

# Bionformatic Analysis and Comparison of Multiple Long Read Whole Genome Assembly Approaches of PacBio HiFi Human Sequence Data

Alison Diaz-Cuevas
International Laboratory for Human Genome Research

The development of next-generation sequencing (NGS) technologies has accelerated the analysis of human genomes and the characterization of human variation at scale. However, some complex variation in the genome has remained elusive to NGS technologies due to the limitations of short reads, some of which may be involved in disease. The improvement of third generation long-read sequencing technologies is enabling the sequencing and analyses of more complete human genomes. One such technology is PacBio HiFi sequencing that can currently achieve long-read lengths between 10-30 kb with accuracy greater than 99.9%. Furthermore, the recent publication of an alternative human genome assembly by the Telomere-to-Telomere (T2T) consortium provides the first gapless assembly of the human genome opening the possibility of detecting variation in previously unexplored regions of the human genome.
We carried out long-read human whole genome sequencing using the third generation PacBio HiFi technology. We performed read mapping and assembly to the GRCh38.p14 human genome reference assembly and the newly published T2T-CHM3 genome using pbmm2. We also executed a de novo assembly using Hifiasm. We compared these three assembly approaches to evaluate their quality, and performed variant calling and analysis of identified variants.

## From genomics to conservation: population genomics approach for climate change adaptation in H. spontaneum

Potapenko Evgenii
University of Haifa, Haifa, Israel

Human activities cause long-term changes in global temperature, precipitation, and weather patterns, known as climate change. Climate change significantly threatens biodiversity, natural habitats, and agriculture, potentially leading to species extinction, reduced crop yields, and food insecurity, particularly in developing countries. Crop wild relatives adapted to harsh environments have significant potential for improving crop resilience and productivity in the face of climate change. Studying ecology and adaptation in crop wild relatives can identify potential genetic sources to improve crop resilience and productivity and shed some light on the mechanisms that enable organisms to adapt to changing environmental conditions. Here we used 30 wild barley populations from different environmental gradients in Israel as a model to identify genomic regions and candidate genes that contribute to adaptation to climatic changes and associated variation in adaptive phenotypic traits to eventually build a predictive model for the adaptive potential of populations to future climate. We identified that maladaptive populations to future climate change of H.spontaneum are located on the board of desert. The methods and insights from this research can be used not only for conservation prioritization but for more sophisticated identification of adaptive haplotypes and broadly applied to understanding adaptation and micro-evolution in other species.

## Improving diagnostics of rare genetic diseases from NGS data

Anna Lewan
Department of Medical Genetics, Institute of Mother and Child, Warsaw, Poland

*Intro* Rare diseases are diseases which affect a small number of people compared to the general population, in Europe disease is considered to be rare when it affects 1 person per 2000. This rare diseases often have a genetic origin and one of the most popular diagnostic methods used to detect them is next-generation sequencing of the genome, exome, or selected panel of genes, which unfortunately has only about 30% diagnostic yield. One of the causes of difficulties in diagnosis are variants that disappear or appear incorrectly due to errors in the genome reference used. The current version of the GRCh38 reference genome, contains a number of regions that pose difficulties in the process of mapping reads from next generation sequencing (NGS) - regions falsely duplicated or collapsed. In an attempt to address these errors, a re-mapping of the reads was undertaken using the T2T-CHM13 genome-based modified GRCh38 reference genome sequence, in which some of these errors are absent.
*Methods* A re-mapping of the NGS reads was performed using the FixItFelix software. Regions were selected based on their position in T2T-CHM13, and reads were then mapped to the revised reference sequence and variant detection was performed. Two groups of patients were analysed: one group of 1,250 patients, originally analysed for selected connective tissue disorders, for variants in the CBS gene; the other group of 1,137 patients, originally analysed for epilepsies and epileptic encephalopathies, for variants in the SIK1 gene.
*Results* In the group of patients reanalysed for alterations in the CBS gene, 7 variants described as pathogenic in the ClinVar database were found, which had not previously been reported. In the second group of patients, 10 previously unreported variants were found in the SIK1 gene, which could potentially be the cause of the symptoms observed in the patients. For one patient, the presence of a variant in the SIK1 gene was confirmed using the Sanger method and was found to have arisen de novo.
*Conclusion* Re-analysis using the modified reference genome showed that errors present in the GRCh38 sequence can result in the omission of diagnostically relevant variants. The use of the FixItFelix programme allows these to be identified and considered in the analysis process, which can contribute to improved diagnostic performance.
*References*
Behera, S., LeFaive, J., Orchard, P. et al. FixItFelix: improving genomic analysis by fixing reference errors. Genome Biol 24, 31 (2023)
Nurk, S. et al. (2022). The complete sequence of a human genome. Science (New York, N.Y.), 376(6588), 44–53

## Empowering Next Generation Sequencing Data Analysis: Reproducible Pipelines, Large-Scale Joint-Genotyping, and Pharmacogenomics Advancements

Monika Krzyżanowska
Lifebit, Poland

I am a bioinformatician specialized in developing reproducible and optimized pipelines for Next Generation Sequencing data analysis, utilizing workflow management systems like WDL, CWL, and Nextflow. This approach enables scientists to conduct large-scale data analyses in a reproducible and straightforward manner. Leveraging multiple machines in the cloud accelerates the analysis process and facilitates multi-cohort studies.

I actively participated in a project focused around large-scale NGS data analysis, where I designed a pipeline for joint-genotyping. This pipeline was capable of simultaneously performing joint-genotyping on 10,000 human Whole Genome Sequencing samples in the cloud, using Arvados system. The results and the methods I developed were presented at the Arvados Conference in 2022 (https://summit.arvados.org/speakers).

Furthermore, my primary focus lies in pharmacogenomics, an emerging field closely related to personalized medicine. It enables us to predict patient-specific gene-drug interactions. In the past, I worked on a research project where I implemented tools to preprocess NGS WGS patient data and predict genotypes within pharmacogenes, which are essential for drug metabolism. However, a challenge arose in connecting genotype information to phenotypic traits and drug-related recommendations due to the complexity of various databases when dealing with fragmented data.

To address this issue, I initiated an open-source project openPGX (https://github.com/monigenomi/openpgx) that facilitates the connection between patient genotypes in star allele nomenclature and phenotype as well as drug-specific recommendations. This project draws data from reputable sources such as the FDA, DPWF, and CPIC, providing a valuable resource for clinicians and researchers in guiding personalized medicine decisions.

# Decoding Image Categorization: A Comparative Analysis of SST and VIP Neuron Responses in the Mouse Visual Cortex

Nadzeya Boyeva
Belarusian State University, Minsk, Belarus

*Introduction*

The ability to process and prioritize new information is essential for adaptive behavior. Stimulus novelty captures attention, enhances perception, and facilitates learning and memory. Recent research has proved that mice perceive and discriminate different visual objects with a high level of accuracy. Additionally, mouse visual cortex neurons play a crucial part in the categorization process and show more activity toward novel images. We aimed to derive the role of inhibitory neurons in the visual cortex in categorizing known and new images of different types (animal, plant, landscape, other, or omitted image).

*Methods and Data*

We have used Allen Institute (Garrett et. al., 2020) experimental data of real-time 2-photon-imaging of VIP (vasoactive intestinal peptide) and SST (somatostatin) inhibitory neurons in mice's visual cortex while mice performed image discrimination tasks. By analyzing fluorescence data from specific neurons with machine learning, we developed classifiers that predict novelty and the category of images for mice.

*Results*

Our inference based on model accuracies indicates that SST neurons are more suitable than VIP neurons for decoding tasks. The mean accuracy of models trained on SST cells data for novelty/familiarity discrimination was 0.76, compared to 0.68 for models trained on VIP cells. Accuracy in image categorization was 0.56 and 0.48 for classifiers trained on SST and VIP data, respectively. Furthermore, different time intervals in fluorescence data used to train classifiers showed different contributions to accuracy prediction. According to our analysis, the most defining time interval was 0.07-0.52 seconds after image exposure.

*Conclusion*

Despite the limitations of the dataset, which only recorded responses from a single type of neuron in a single mouse, we can conclude that the response of SST neurons is more informative for the categorization task compared to VIP neurons. Furthermore, in a certain period, neural responses were the most defining, which could mean that's the moment of encoding.

## Within-host diversity of avian influenza virus in different poultry species

Kamila Dziadek
Department of Poultry Diseases, National Veterinary Research Institute, Partyzantów 57 Avenue 24-100 Puławy, Poland

BACKGROUND: Avian influenza viruses (AIV) are highly variable pathogens that spread among wild aquatic birds worldwide posing a threat to domestic bird populations. Genetic variability of AIVs is determined mainly by two mechanisms: a) genetic shift, responsible for the emergence of new AIV genotypes, and b) genetic drift, being a consequence of error-prone replication. The latter is related to the lack of proofreading ability in RNA polymerases, followed by increased mutation rates during virus replication and the generation of numerous mutants forming viral populations called „quasispecies". As a result of ongoing selection and competition between quasispecies, the relative frequency of advantageous variants may change in order to adapt to alternating environmental conditions or to a new host. For AIVs, specific mutations associated with crossing the interspecies barriers, including virus adaptation to mammals, have been already identified and involve mainly genome segments encoding the polymerase complex (PB1, PB2, and PA) and surface proteins (HA, NA).
AIM: To better understand the role of different poultry species in the AIV evolution and emergence of novel variants with increased adaptive properties, experimental infection with a „wild-type" low-pathogenic (LP) avian influenza virus was performed. The study aims to investigate the within-host diversity of LPAIV in both virus replication sites i.e. respiratory and gastrointestinal tract.

## Characterising the Role of RBP STAU2 in Human Neurogenesis using hiPSCs with scRNA-Seq

Akshay J Ganesh
IDIBELL, Barcelona, Spain

Neurogenesis is the process of generating new neurons during embryonic cortical development. This crucial process is regulated at multiple levels including genetic, epigenetic and transcriptional controls. Recent studies have suggested that post-transcriptional regulation of mRNAs by RNA-Binding Proteins (RBPs) is also a crucial step in modulating neurogenesis and cortical development, but the regulatory networks that guide these mechanisms remain uncharacterised. One specific RBP, Staufen 2 (STAU2), is known to be involved in regulating neural stem cell maintenance in murine development. But the role of STAU2 in human neurogenesis remains uncharacterised. To study the transcriptomic regulatory networks guided by STAU2 activity at different stages of human neurogenesis, we differentiated control and Crispr-Cas-mediated STAU2-KO human induced pluripotent stem cells (hiPSCs) into different neurogenic populations. Single-cell RNASeq (scRNA-Seq) was performed for these cultures at different timepoints of differentiation (D0, D11, D25, D55, D70) checking for variations in neuronal subpopulations and differentially expressed genes between conditions. We observed that the expression of marker genes at subsequent timepoints recapitulates the transition from hiPSCs to Neuroepithelial cells to neuronal progenitors to fully differentiated neurons. Trajectory analysis will also be performed to study the correlation between STAU2 gene expression and other transcripts. We believe the results from this study could improve the understanding of STAU2's role in the gene regulatory networks guiding human neurogenesis.

## Machine learning models predicting response to immune checkpoint inhibitors (ICI) based on DNA and RNA biomarkers

Axel Gschwind
University of Tübingen, Tübingen, Germany

I began my PhD project last year. I aim to integrate multiple DNA- and RNASeq-based biomarkers using various machine learning methods.

Many cancer patients experience clinical benefits from ICI treatment. However, ICI is only effective for a minority of patients who qualify for this kind of treatment. Some DNA- and RNA-Seq-based biomarkers have been shown to correlate with the outcome of ICI. But, to date, only the tumor mutation burden (TMB) is an FDA-approved stand-alone biomarker for the selection of ICI, with a high TMB improving the chance for a favorable outcome. Improving ICI response prediction is an urgent need, considering that using high TMB as the sole selection criteria results in an overall response rate of only 29% of treated patients. Moreover, many patients suffer from potential life-threatening side effects of the treatment while not showing any benefits.
The goal of my study is to integrate several DNA and RNA biomarkers (somatic and germline), generated by Next Generation Sequencing (NGS) of tumor and healthy (or 'normal') tissue, into a machine learning model predicting response to ICI with high accuracy. Many different biomarkers can be derived from NGS data from tumor and normal samples. This includes, e.g., TMB, neoantigens, mutations of known ICI resistance genes, immune cell infiltration and diversity, HLA-related features, microbial infiltration, certain gene expression signatures, etc.
I am currently using a meta-cohort of ICI-treated cancer patients for which both NGS data and clinical outcome data are available. The clinical outcome used as a training label is defined as the response according to radiologic RECIST criteria, resulting in two classes: responders (R) and non-responders (NR). Much of the data can be obtained from public databases like the European Nucleotide Archive (ENA), the Gene Expression Omnibus (GEO), or dbGaP.
Besides my research, I used to work in a computational genomics group providing analyses for the medical treatment of cancer patients. I enjoy being out in the country for sports when I am not doing research. I also enjoy socializing with all kinds of people and attending NGSchools, participating now for the second time.

## Cracking the Viral Code: Outsmarting Non-Living Masterminds

Soumajit Mukherjee
King's College London, UK

How do we kill something that is not living? Viruses, one of the tiniest biological entities, are able to hijack and manipulate the sophisticated machinery of mammalian cells to ensure their own survival and replication. I am an early career researcher with strong interest to build an academic career on the interface of infection biology and bioinformatics. Throughout my research experiences, I have developed a passion for exploring new frontiers in the field of host-pathogen interactions research and have been intrigued with questions about how pathogens utilize proteins to capture host resources for survival, replication under stress and to evade immune responses. My doctoral thesis was focused on screening set of nanomaterials for potent antiviral activity and understanding the dynamics and kinetics of the inhibition by using biochemical and imaging methods. In addition, the modulatory and toxic effects of the nanomaterial were studied using transcriptome analysis focusing on immune responses and signaling pathways. In my post-doctoral research, I will be studying human secretome and released cDNA to identify natural suppressors of SARS-CoV-2. My research experience spans molecular biology, biochemistry and virology techniques. Along with this, I enjoy working with programming languages to explore host-pathogen interactome, to design CRISPR-based experiments or to analyze sequencing and multi-omics data to provide insight about changes in gene expression and pathway alterations.

## Computational Variant and Gene Prioritization in Chronic Pancreatitis

Andreas Walter Schmidt
Institute of Medical Genetics and Applied Genomics, University Hospital Tübingen, Tübingen, Germany

I started my PhD in the field of chronic pancreatitis genetics in 2020. Chronic pancreatitis is a rare oligogenic disease leading to recurrent abdominal pain and loss of exo-/endocrine activity. Heavy drinking is considered a major risk factor, however several genetically associated risk loci have been discovered in the past by targeted sequencing and GWAS studies. In my work, I process and analyze GWAS, whole-exome sequencing (WES) and whole-genome sequencing (WES) cohorts to discover disease-associated variants and genes in chronic pancreatitis.

Recently me and my group integrated (non-)alcoholic chronic pancreatitis GWAS data with pancreas specific eQTLs from GTEx in a colocalization model to prioritize disease causing variants and genes and published our results in the Journal of Pancreatology. We discovered increased CLDN2 expression and reduced CTRC expression to be associated with chronic pancreatitis under both, alcoholic and non-alcoholic conditions.

In parallel, I investigated a WES cohort of 1000 non-alcoholic chronic pancreatitis cases and discovered several rare high-impact variants in novel candidate genes by burden testing against the gnomAD database. Preliminary experimental data supports the functional relevance of the discovered mutations.

Furthermore, I recently started the WGS analysis of 450 so far unexplained non-alcoholic chronic pancreatitis cases. Based on this dataset, I aim to discover potentially rare non-coding variants associated with chronic pancreatitis.

Additionally, I have a keen interest in leveraging biological networks and databases to improve the prioritization of causal genes within my results.

## Scrutinised and Compared: HVG Identification Methods in Terms of Common Metrics

Matvii Mykhailichenko
University of Wroclaw

Highly variable gene identification plays a critical role in unravelling gene expression patterns and understanding cellular heterogeneity in single-cell RNA-sequencing data. A plethora of software packages have emerged for this purpose; however, their comparative performance remains inadequately explored. Presented study addresses this gap by independently evaluating 22 methods from 9 different packages to provide a comprehensive benchmarking of HVG identification methods. For such purpose a set of common metrics were chosen, namely overlap with highly and lowly expressed genes, runtime, and clustering indices (e.g., Calinski-Harabasz, Davies-Bouldin, and ROGUE). The results reveal substantial disparities not only between different methods but also in the performance of a single method across diverse datasets. That is to say, the dimensionality of data provided, spike-ins, and background noise are some of the key factors influencing the results. These variations underscore the significant impact of dataset characteristics on analysis outcomes. Therefore, consistent consideration of data nature is imperative. The study emphasises the urgent need for a standardised, data-driven benchmarking framework to ensure reliable and effective scRNA-seq analyses. Presented work serves as a valuable resource for both scRNA-seq software developers and experimental researchers seeking optimal methods for their investigations.

## Characterisation of regulome in human astrocytes, based on IPSC-derived astrocyte models

Eryk Duński
Nencki Institute PAS

Human astrocytes display unique features in comparison to their murine and ape counterparts, including increased size and complexity of the arborization pattern. Protein-coding genes expressed in the brain are largely conserved, therefore the changes observed in the brain during primate evolution may be mostly driven by changes in the regulome. We obtained induced pluripotent stem cells (iPSC) from human, chimpanzee, and rhesus macaque and used them to derive astrocytes in vitro (iAstrocytes). We performed RNA-seq, ATAC-seq and ChIP-seq to determine the interspecies variation in gene expression and activity of cis-regulatory elements.
We identified differentially open regions and, by intersecting ATAC-seq data with ChIP-seq for H3K27ac, we found putative enhancers that are only active in human iAstrocytes. Using massively parallel reporter assay (MPRA), we assayed 5,073 unique human sequences to determine their cis-regulatory activity. Our set included 3,026 enhancers that are located in proximity of the differentially expressed genes (DEGs), 301 enhancers surrounding genes crucial for astrocyte biology and 197 enhancers containing single nucleotide (SNP) polymorphisms linked to cognitive ability and brain-related disorders. Our results allow us draw conclusions about features of enhancers driving human-specific expression signatures of astrocytes and propose links between evolutionary changes in regulome and human disease.

## Life cycle evolution in siphonophores (Cnidaria)

Maciej Mańko
University of Gdańsk, Gdańsk, Poland

Siphonophores are exclusively marine, colonial cnidarians, characterized by complex colony organization and unparalleled zooid functional specialization. Recent genomic studies have offered an evolutionary perspective on how this morphological complexity arose, and also revealed giant sizes of their nuclear genomes (0.7-4.8Gb). Despite these advancements we are still lacking basic understanding of their early development and life cycles. Although siphonophores belong to radially symmetrical Cnidaria, both their embryos and mature colonies are bilaterally symmetrical. We thus employed immunohistochemistry to characterize symmetry breaking events, but their molecular ques have yet to be revealed. Some siphonophores also undergo a complex process of ordered colony fragmentation, that leads to a release of individual-like groups of zooids (=colony building blocks), which later disperse independently from parental colony and reproduce sexually. We have described this phenomenon with an integrative eco-evo-devo approach, focusing on developmental patterns governing the fragmentation and its control mechanisms. Now, with the first siphonophore genomes coming, we are planning to dive into siphonophore population genomics to link different life cycle patterns with population structure of these widely dispersed oceanic predators.

## The Random Forest-based approach to discover breast cancer biomarkers

Nadiia Kasianchuk
Faculty of Biology, Adam Mickiewicz University, Poznan, Poland / Bogomolets National Medical University, Kyiv, Ukraine

Employing advanced data analysis tools and bioinformatics is crucial for unraveling the complexities of breast cancer, the foremost cause of cancer-related deaths among women. This study aims to detect potential genomic biomarkers that wield substantial influence over four key prognostic factors: tumor size, lymph node involvement, metastasis, and overall survival status. The dataset from The Cancer Genome Atlas Breast Cancer, housing expression values of 19,737 genes, has been harnessed to train the Random Forest algorithm.

To attain an optimal learning model, the process was iterated 20 times for each indicator. Subsequent consideration was granted only to genes boasting a p-value < 0.05. A comprehensive assessment of the algorithm's reliability entailed the computation of diverse performance metrics, including the F1 score. This meticulous scrutiny resulted in the inclusion of 97 and 7 genes within the extended and final databases, respectively. Scientific validation confirms the pivotal roles of these selected genes within cancer-associated pathways like Toll-like receptor and NFκB, exerting influence over critical aspects such as cell proliferation, tumor development, and angiogenesis.

In summary, this study effectively underscores the potential of machine learning analyses within the biomedical arena. It not only imparts machine-generated insights into the intricacies of breast cancer development but also lays the foundation for subsequent in vitro investigations, aimed at substantiating the prognostic efficacy of the identified biomarker.