
# ABSTRACT BOOK

## Summer School in Bioinformatics & NGS Data Analysis

Dolný Smokovec, Slovakia, August 14-21, 2016

**#NGSchool2016**
`https://ngschool.eu/`

**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
- Scientific committee: Broňa Brejová, Maciej Łapiński, Marina Marcet-Houben & Tomáš Vinař
- IIMCB Admin, Grant, Finance & PR Units
- Centrum Pod Lesom: arranged accommodation & boarding
- Ewa Ramotowska #NGSchool logo

**Supporters** This activity is financially supported by grant from International Visegrad Fund (Standard Grant No. 21610099) 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 two breaks at 11:00 and 16:00. Breakfast will be served from 8:00, lunch at 13:00 and dinner around 19:00.

| | | |
|---|---|---|
| | **Day 0: Sunday** | |
| 18:00 | **Welcome & Shot talks #1** | *Leszek Pryszcz* |
| | **Day 1: Monday** | |
| 9:00 | **Introduction to Linux** | *Roman Cheplyaka* |
| 14:00 | **Introduction to Bioinformatics & NGS** | *Broňa Brejová / Tomáš Vinař* |
| | **Day 2: Tuesday** | |
| 9:00 | **Genome & transcriptome assembly** | *Leszek Pryszcz / German Demidov* |
| 14:00 | **Functional genome annotation** | *Marina Marcet-Houben* |
| 20:00 | **Shot talks #2** | *Leszek Pryszcz* |
| | **Day 3: Wednesday** | |
| 9:00 | **CNV detection** | *German Demidov* |
| 14:00 | **Differential expression** | *Leszek Pryszcz* |
| 18:00 | **Library construction for NGS** | *Paulina Stachula* |
| | **Day 4: Thursday** | |
| 9:00 | **ChIP-seq** | *Maciej Łapiński* |
| 12:00 | **From ChIP-seq to Hi-C** | *Bartek Wilczyński* |
| 14:00 | **Bisulphite sequencing** | *Russell Hamilton* |
| | **Day 5: Friday** | |
| 9:00 | **Molecular data integration** | *Jacek Marzec* |
| 16:00 | **Press conference** | *Organising committee* |
| | **Day 6: Saturday** | |
| 9:00 | **NGS & Biomedicine** | *Sophia Derdak* |
| | **Day 7: Sunday** | |
| 9:00 | **Recap & farewell** | *Leszek Pryszcz* |

## Social activities

Social activities will take place during or after the dinner.
We'll go for hiking on Friday or Saturday, depending on the weather.
You can rent bikes from Galfy (`http://www.galfy.sk/sk/pozicovna-bicyklov`) in Horny Smokovec.

# Mutational signatures of DNA repair deficiencies and cytotoxin exposures in C. elegans

Nadezda Volkova

European Bioinformatics Institute (EMBL-EBI), UK

Cancer is caused by alterations in the genome. These alterations can be an effect of combination of environmental factors damaging DNA and deficiencies in DNA repair and replication leading to characteristic mutational spectra.

Mutational signatures (Alexandrov et al. 2013) became a very useful tool of cancer investigation in the last years. However, the signatures identified so far mostly represent complex conglomerates of the action of different mutational processes. For many signatures, the link with the underlying mutational processes is still unclear.

The main goal of this study is to present a systematic catalogue of mutational signatures caused by genotoxins, DNA repair deficiencies and their interactions in order to dissect the signatures acting in cancer and understand the mechanisms that generate them.

In this study we used C. elegans as a model organism to present a systematic screen with 9 types of genotoxins under 58 different genetic conditions including single and double knock-outs of DNA repair associated genes. Upon exposure over several generations we used whole-genome sequencing to study patterns of DNA damage.

We studied the mutational spectra by analysing different types of genetic lesions including point mutations, indels and structural variants using rigorous quality control procedure. This approach allows us to dissect the precise individual contributions of each factor using constrained Poisson additive models, and also identify epistatic events such as 3-fold increase in mutational burden for pms-2/pole-4 double knock-out. In summary, this analysis presents the first systematic catalogue of mutational signatures caused by genotoxins and DNA repair deficiencies.

# Decoupling of genomic and mitochondrial DNA replication and noisy partitioning during division results in reversible phenotypic heterogeneity.

Alsu Missarova

Barcelona, Universitat Pompeu Fabra, Single-cell Behaviour, Spain

While mitochondria are essential for most eukaryotes, the requirement for mitochondrial activity and function varies between cell types and across growth conditions. Mitochondria are involved in both fermentation and respiration. Budding yeast require mitochondria, but not mtDNA to survive; cells lacking mtDNA are known as petites and are unable to respire. The petite state is generally considered an irreversible binary phenotype: once a cell has lost its mtDNA it, and all of its progeny, will remain petite. We have experimental evidence suggesting that respiratory competency is actually a continuous phenotype. A single cell can be respiratory deficient, but have a non-zero probability of giving rise to respiratory proficient offspring. Therefore lack of respiration is stochastically reversible. We propose two distinct mechanisms: random segregation of mtDNA molecules during cell division, and a prolongation of the cell cycle in conditions where a fermentable carbon source is limited. Mathematical modeling suggests that the combination of these two processes leads to a stochastically reversible petite phenotype in a subpopulation of exponentially growing wild-type cells.

## Characterization of a genetic loss-of-function of FUS in zebrafish

Svetlana Lebedeva

Institute of Molecular Biology, Mainz, Germany

The RNA-binding protein FUS is a multifunctional nuclear protein implicated in transcription, alternative splicing of neuronal genes and DNA repair. Mutations in this protein, among others, has been linked to the neurodegenerative diseases ALS (amyotrophic lateral sclerosis) and FTLD (frontotemporal lobal degeneration). We genetically disrupted FUS in zebrafish (Danio rerio) using the CRISPR/Cas9 system. The FUS knockout fish appear healthy and fertile. We quantified transcriptome and proteome of the knockout adult brain. In general, deletion of FUS in zebrafish induces relatively mild changes in gene expression on the transcriptome and proteome level. Analyzing the transcriptome, however, we noticed significant influence of genetic background/SNPs on gene expression, obscuring possible minor changes that could be attributed to FUS. Unlike FUS morphants, FUS knockout embryos do not show apparent motoneuronal degeneration and perform equally well in motility test as the wild type siblings.

# Disclosing the causes underlying sterility in Zygosaccharomyces allodiploid yeast by NGS-assisted approaches

Melissa Bizzarri

University of Modena and Reggio Emilia, Department of Life Sciences, Italy

Yeast genome content and architecture can vary progressively during the adaptive response. However a stepwise turning point in genome evolution is often the outcrossing between intermating allopatric yeasts which gives rise to viable hybrids. Complete sets of orthologous genes are expected in hybrids immediately after the merging of two parental genomes but these patterns can undergo extensive homogenization processes over evolutionary time through intragenic recombination, gene conversion and differential gene loss. The result is offspring with phenotypic novelties exposed to natural selection. However we have limited knowledge of how genomes stabilize following the initial interspecific hybridization. Extending these notions will be also useful for genetic improvement of industrial yeasts which largely rely on hybrid offspring and whose sterility prevents their exploitation in inbreeding programs. Zygosaccharomyces rouxii species is a non WGD yeast complex adapted to grow in food with high solute concentrations and frequently experience variations in ploidy, resulting in different modes of reproduction. Within this group, the allodiploid ATCC 42981 strain has been studied for its ability to survive and produce glycerol under salt stress. Despite having a DNA diploid content ATCC 42981 strain is unable to undergo meiosis under different standard and stress growth conditions and displays remnants of haploid phenotype. I showed that ATCC 42981 strain had a chimeric mating system consisting of two phylogenetically divergent MATa/MAT$\alpha$ expression loci and two copies of the HO endonuclease gene. I argue that sequence divergence within the chimeric a1-$\alpha$2 heterodimer, the master mating-type transcriptional factor, could be involved in the generation of negative epistasis, contributing to the allodiploid sterility and the dysregulation of cell-type identity. To verify this hypothesis, I am currently restoring a phylogenetically congruent and putatively functional a1-$\alpha$2 heterodimer by genetic engineering of ATCC 42981 cells to ascertain whether sexuality will be restored. Subsequently three NGS approaches will be carried out on ATCC 42981: 1) de-novo genome assembling to establish genome-wide patterns of introgression; 2) and transcriptome analysis of salt stress response; 3) Chip-Seq focused on a1-$\alpha$2 heterodimer target genes. Overall this integrated NGS-assisted strategy will unveil how introgressed parental genomes interplay at different levels of complexity.

# Alternative pipelines for cross-platform microarray gene expression data integration tested with RNA-seq data

Alina Frolova

Systems Biology laboratory, Institute of Molecular Biology and Genetics, Ukraine

Background: The potential of microarrays data to bring new insights is limited by the effects of individual study-specific biases and small numbers of biological samples. Increasing sample size by microarray data integration increases the statistical power to obtain more precise estimate of a gene expression in a population of individuals resulting in lower false discovery rates. However, despite numerous recommendations for gene expression data integration, there is a lack of a systematic comparison of different data processing approaches which would account for a diversity of microarray platforms and ambiguous probesets to genes correspondence. Results: Here, we accessed several methods of the direct microarrays data merging via comparison with RNA-seq data on breast cancer samples. We aimed to evaluate different probesets annotations as well as procedures of choosing between probesets mapped to the same gene using datasets from popular Affymetrix and Illumina platforms. We show that methods which rely on actual probesets signal intensities are advantageous over the methods considering biological characteristics of the probes sequences only, and that cross-platform integration of datasets improves correlation with the RNA-seq data. Conclusion: We consider the results obtained in this paper contributive to the integrative analysis as a worthwhile alternative to the meta-analysis of multiple gene expression datasets originated from different microarray platforms.

# Identification of sister chromatid cohesion sites during interphase in mammalian cells

Azamat Gafurov

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

Over the last decade, development and application of a set of molecular genomic approaches based on the chromosome conformation capture such as Hi-C enabled high resolution and genome-wide analysis of the spatial organization of chromosomes. Hi-C is based on paired-end deep sequencing to detect and quantify pairwise chromatin interactions genome-wide. Here we have applied Hi-C data to study sister chromatid cohesion that refers to the process by which sister chromatids are paired and held together by cohesin proteins during certain phases of cell cycle, which is visually had been detected only during mitosis. We have examined both cohesin binding sites using ChIP-seq data analysis and Hi-C paired-end reads that mapped to the same restriction fragment and pointed in the same direction, reflecting interactions between paired and aligned chromatids

# The effect of Transposable Elements in the regulation of alternative splicing in human neural tissues

Quirze Rovira Castellà

Center for Genomic Regulation (CRG), Spain

The huge range of transcriptomic diversity among cells of a single organism and during different time points is an example of the importance of transcriptome regulation for cell identity and behaviour. This diversity is achieved by a tightly regulation of the gene expression, but also by many post-transcriptional mechanisms. Among them, alternative splicing is a pervasive process to regulate gene production by differentially including or excluding introns and exons during the pre-mRNA maturation. This process impacts up to 95% of the multiexonic human genes.

Transposable elements (TEs) are mobile DNA sequences that have the ability to move around the genome and often generate new copies of themselves in the process. TEs are widespread among the evolutionary tree and represent a significant load of some genomes, being up to 70% in some plants and around 45% in primates. Although their activity can often create harmful mutation, it also represents an important source of genetic variation that has been shown to affect transcriptional regulatory networks. Furthermore, it has been observed that TE-derived sequences can impact post-transcriptional regulation, but more studies are needed to comprehend this processes and its evolutionary impact.

In this study we aim to investigate the potential role of mobile elements in the evolution and origin of alternative splicing and its regulation in vertebrates. We have analysed the scenario of TEs insertions within or around alternative splicing events annotated using VAST-TOOLS (Vertebrate Alternative Splicing and Transcription Tools) software with RNA-seq data from various human tissue-specific samples. Our preliminary results show that there is an enrichment of cassette exons containing TE-derived material among the neural-specific exons of the human genome. Moreover, some neural RNA binding protein (RBP) binding sites are present inside this TEs insertions. With the same approach we will further probe the enrichment of TE insertions for other human tissues and compare this results with mouse, cow and chicken genomes.

Our results suggest that TEs may have a role in the origin and regulation of some neural-specific exons in the human genome.

# Metagenomic analysis of Black Sea mictobial communitites

Mariia Pavlovska

Ukrainian scientific center of Ecology of Sea (UkrSCES), Odesa, Ukraine

It is widely known that about half of the annual primary production of the planet occurs in the ocean. Bacteria are involved in the chemical transformation of most elements, yet very little is known of marine microbial diversity compared to the diversity of marine fauna and flora. Such sitiuation arrises from the fact that the majority of marine microbial organisms are recalcitrant to the traditional culturing techniques. Our project adresses the problem of the lack of such data by applying the methods of metagenomic analysis to monitoring of Black Sea microbial communities. The study is novel in several aspects. First, this is the first large-scale metagenomic analysis for the Black Sea, that adds resolution to such global research project as International Census of Marine Microbes and Ocean Sampling Day. Second, it gives the insight into the pattern of functional and taxonomic diversity of deep-water microbial communities, which are found in the anaerobic (H2S) zone of the Black Sea and thus adapted to have specific metabolism. We have collected 69 sea water samples from 12 stations and 30 sediment samples from 5 stations during the Joint Open Sea Survey which was conducted under EMBLAS-II project in May-June 2016. The sea water samples were taken from the following horizons: surface, thermocline, deep chlorophyll maximum, nutrient maximum, oxygen minimum and H2S zone. Sample processing will involve DNA extraction and 16s Illumina sequencing and will be performed in July 2016. The next step will be bioinformatic analysis. First, sequence assembly and gene prediction will be performed. Second, 16S tags (Illumina mitags) will be mapped to taxonomically annotated 16S sequences from SILVA database and protein-coding marker genes suitable for metagenomic species profiling will be identified and clustered into metagenomic operational taxonomic units (mOTUs). The relative abundance of each mOTU group will be estimated giving the insight into species composition of Black Sea microbial communities. Third, the high-quality reads will be mapped to OM-RGC data base to infer gene abundace profiles for the community. Finally, we will correlate relative abundances of taxonomic and gene compositions with the data on environmental parameters, such as temperature, salinity, nutrient abundace, oxygen concentration and H2S concentration to discover specific gene and taxonomic patterns and possible adaptation of microbial communities.

# Gene expression profile of cell lines with acquired TRAIL resistance, multicellular TRAIL resistance and sensitivity.

Nadezda Dolgikh

Pushchino State Natural Science Institute, Pushchino, Russia

Since TRAIL (Tumor Necrosis Factor-Related Apoptosis-Inducing Ligand) has been described as a vital component of tumor immunosurveillance, capable to induce apoptosis exclusively in malignant cells, the question remains: how tumor progresses and survives despite the presence of TRAIL-producing cells and soluble TRAIL in the tumor environment? The sensitivity of tumor cells to TRAIL-induced apoptosis is a key factor in the effectiveness of anti-cancer therapy based on recombinant TRAIL protein. However, the molecular mechanisms of sensitivity and resistance to TRAIL-induced apoptosis in cancer cells are not fully understood. The aim of this study is to investigate molecular mechanisms of TRAIL-resistance using gene expression profile of cell lines with acquired TRAIL resistance, multicellular TRAIL resistance and sensitivity. Earlier we have observed dramatic increase in TRAIL resistance of A431 cell line and many other cancer cells in confluent culture, so called multicellular TRAIL resistance. This type of resistance occurs when initially sensitive cancer cells become highly resistant to TRAIL in multicellular confluent conditions, regardless of the amount of TRAIL protein and proliferative activity of cells. In order to investigate how cancer cells escape TRAIL-induced apoptosis, we generated A431 cells with acquired TRAIL resistance (A431-R cell line) by exposing the parental sensitive cells to subtoxic concentrations of TRAIL, and stable TRAIL-sensitive cell line in multicellular conditions (A431-S cell line). To identify changes associated specifically with TRAIL resistance/sensitivity in multicellular conditions we performed genome-wide gene expression analysis of parental A431 and generated cell lines (A431-R and A431-S). We observed that acquired TRAIL resistance was associated with complex disruption of both extrinsic and intrinsic apoptotic pathways. Moreover, one possible mechanism that might explain multicellular TRAIL resistance is a disruption of the signal transduction form ligand-receptor complex to mitochondria. The results indicate diverse protection mechanisms of cancer cells that need to be considered when developing new anticancer strategies.

# The transferability of the antimicrobial peptide resistome from human gut bactera to pathogens

Ádám Györkei

Department of Biochemistry, Biological Research Center, Hungary

The future of antimicrobial peptides (AMPs) as promosing novel antimicrobial agents depends on the extent to which resistance can evolve against them. Prior works indicate that intestinal microbial communities serve as a large reservoir of antibiotic resistance genes for horizontal gene transfer. In sharp contrast to this, the availability of AMP resistance genes in these environments remains a terra incognita, due to the shortage of relevant systematic studies. Here we functinoally characterize the transferability of the AMP resistome of the human gut microbiome by culturing the gut microbiome and screening large metagenomic libraries in the presence of different AMPs followed by applying comparative genomic techniques.

# The evaluation of dynamics of barley methylome modulation under water deficiency stress using novel MSAP-Seq method.

Karolina Chwialkowska

Department of Genetics, University of Silesia in Katowice, Poland

Plants are constantly challenged by environmental stresses and, thus, have developed a number of strategies allowing for a rapid adaptation to unfavorable conditions. One of them is the modification of multi-level epigenetic machinery, with DNA methylation ahead, resulting in gene expression modulation. This study attempts to characterize and compare the barley methylome of leaves and roots under water-deficiency treatment and the subsequent rewatering phase. Moreover, detailed analyses of the dynamics of drought stress-induced changes in DNA methylation in barley are presented. In this study we have applied developed by our group a novel Methylation Sensitive Amplification Polymorphism Sequencing (MSAP-Seq) technique. It is a simple and high-throughput but low-cost method for a direct identification of specific genomic sequences that undergo DNA methylation changes. MSAP-Seq is based on the conventional MSAP analyses, however we have greatly improved it by replacing of the conventional separation of MSAP amplicons on polyacrylamide gels with their direct high-throughput sequencing using NGS methods and automated data analysis. This technique allows for parallel identification and analysis of hundreds of thousands of sites. By using MSAP-Seq we identified dozens of thousands of differentially methylated sites, that were induced by drought stress in the barley leaves and roots. Drought stress induced mainly demethylations in leaves, but in roots new methylations dominated. Importantly, rewatering and plant recovery resulted in the reversibility of the majority of stress-induced methylation events, though this process was more efficient in the leaves than in the roots. Moreover, a detailed profiling of the barley methylome at several time points during progressive water deficiency conditions as well as rewatering phase revealed its dynamic modulation in terms of the quantity and localization of drought-induced changes. The profiles of the drought-induced changes were tracked upon rewatering phase, what revealed that changes occurring in the early drought phases are more persistent than those occurring in the late drought, which are highly reversible. Altogether our analyses suggest that the dynamic nature, as well as organ specificity of the methylome changes in response to water deficiency, might be an important regulatory system leading to the multi-level mechanisms of stress response in large-genome crops.

-

Ewa Ramotowska
University of Warsaw, Poland

-

# Genome-wide identification and characterization of carotenoid biosynthesis pathway genes in Cucurbita maxima Duchesne

Krzysztof Sobieszek

Department of Plant Genetics, Breeding and Biotechnology, Warsaw University of Life Sciences, Poland

Carotenoid biosynthesis pathway genes are well characterized only for few model plant species. So far there have not been any attempts to identify and characterize these genes in winter squash (Cucurbita maxima Duchesne). Since the fruits of C. maxima are used as carotenoid source in many baby nutritional products (due to the fruits low nitrate and heavy metals accumulation potential) it is particularly interesting and applicative to identify genes and to study carotenoid biosynthesis pathway in this species. Recently C. maxima genome has been sequenced and it is used as reference in this study. The aims of this research are: de novo identification of carotenoid biosynthesis pathway genes including regulatory regions, search for polymorphisms (SNPs, Indels) related to these genes, and mapping them on C. maxima genetic map that is under development at the Department of Plant Genetics Breeding and Biotechnology at SGGW (Warsaw, Poland). Genetic markers associated with carotenoid biosynthesis pathway genes can be powerful tool for modern C. maxima breeding programs that are focused on novel varieties creation characterized by high and custom-designed carotenoids content in the fruits.

# Candida albicans potassium uptake systems as novel targets for antifungal drug discovery

Hana Elicharová

Department of Membrane Transport Institute of Physiology of the Czech Academy of Sciences
Vídeňská 1083, 142 20 Prague 4, Czech Republic

Several tens of Candida species are opportunistic human pathogens able to cause life-threatening system infection in immunocompromised patients while C. albicans possesses among Candida species the highest incidence. Candida species as other living cells maintain relatively high intracellular concentration (200 – 300 mM) of potassium cations, which play an important role in many biological processes such as the regulation of cell volume, pH and membrane potential. Yeast cells use potassium importers (uniporters, symporters and ATPases) to provide cell with sufficient amount of K+. We identified the genes encoding putative K+- importers in genomes of nine Candida species. These differ in the number of genes for K+ importers and in their ability to grow on limiting concentration of KCl. But a higher number of genes for K+-importers does not correlate with better growth upon K+ limitation (1). C. albicans is only one species whose genome contains genes for all three known types of yeast K+-importers. Upon heterologous expression in the S. cerevisiae Δtrk1Δtrk2 strain lacking its own potassium transporters, all of them able were to provide sufficient amount of K+ to support cell growth and the division (2). In two C. albicans lab strains (SC5314 and WO-1) ACU1 gene encoding K+ -importing ATPase is divided in two ORFs. We found that the mutation resulting in STOP codon breaking this gene is present only in approx. one third of 34 C. albicans strains with known genome sequence. We plan to analyze the variability of genes for all K+-importers among C. albicans strains and follow in their functional characterization to clarify the role in potassium homeostasis of single K+-importers, which could be promising targets of new antifungals.

This work was supported by the Czech National Science Foundation GACR 16-03398S

(1) Hušeková, B., H. Elicharová and H. Sychrová (2016). Pathogenic Candida species differ in the ability to grow at limiting potassium concentrations. Can J Microbiol 62: 394-401.
(2) Elicharová, H., B. Hušeková and H. Sychrová (2016). Three Candida albicans potassium uptake systems differ in their ability to provide Saccharomyces cerevisiae trk1trk2 mutants with necessary potassium. FEMS Yeast Res doi: 10.1093/femsyr/fow039

# Comparative analysis of mitochondrial genomes of the Magnusiomyces/Saprochaete clade

Filip Brázdovič

Department of Genetics, Faculty of Natural Sciences, Comenius University, Slovakia

Mitochondria contain highly derived genomes that are variable in size, base composition, topology, and genetic organization. In this study, we have performed a comparative analysis of the mitochondrial genomes from the yeast species taxonomically classified into the Magnusiomyces/Saprochaete clade (a deeply diverging branch of the subphylum Saccharomycotina) that exhibits remarkable heterogeneity in terms of cell morphology, physiology, metabolism, and genome organization. We have determined the complete mitochondrial DNA (mtDNA) sequences of twenty-two strains from fifteen magnusiomycete species. Although the examined mitochondrial genomes comprise essentially the same set of genes (i.e. the genes for apocytochrome b, the subunits of ATP synthase, cytochrome c oxidase, and NADH:ubiquinone oxidoreductase, the ribosomal protein Rps3, a set of 25 tRNAs, and two rRNAs), these mtDNAs display a high degree of variability in the genome size (30.9 – 71.2 kbp), guanine plus cytosine composition (16.4 – 48.7 %), and intron content (0 – 23). In addition, the coding sequences of three species (M. capitatus, M. spicifer, and S. clavata) contain insertion elements dubbed byps, which remain in mature mRNAs but are ignored during mitochondrial protein synthesis via programmed translational bypassing. These genetic elements are classified into several sequence families and their distribution in the coding sequences points to mobility presumably by a transposon-like mechanism [1, 2]. Comparative analysis of byps from closely related species can be instrumental in understanding their evolutionary origin and horizontal transfer.

References 1. Lang B.F., et al. (2014). Proc. Natl. Acad. Sci. USA 111(16): 5926–5931. 2. Nosek J., et al. (2015). Trends Genet. 31(4): 187-194.

# Classical and non-classical markers for human population discrimination.

Patrycja Daca-Roszak

Department of Molecular and Clinical Genetics, Institute of Human Genetics, Polish Academy of Sciences, Poland

Markers that allow analysis of biogeographic origin of biological samples in a standard forensic laboratory are still sought for. We performed a comprehensive analysis of B-lymphocyte cell lines representing Europeans (n=75) and East Asians (n=83). Based on the results obtained using high throughput methods (microarrays, Illumina) we identified three sets of markers: classical (SNP) and non-classical (mRNA, CpG loci), potentially suitable for the discrimination of European and East Asian populations. A SNaPshot assay, developed to test 14 SNPs, was sufficient to discriminate these two populations from each other as well as from Africans; however the use of this assay in low-quality/quantity DNA samples was often hampered by drop-in and drop-out phenomena. We further identified 14 CpG loci representing statistically significant inter-population differences in the methylation status; this part of the study also indicated the importance of correcting Illumina microarray results for a confounding impact of CpG-located SNPs. Lastly, we identified 20 loci, which substantially differed in their expression level between analyzed populations. Moreover, we showed that immunofluorescent probes targeting two of the differentially expressed mRNA markers enabled separation cell sorter of a mixture of cell lines according to their ethnic origin (European versus Asian). Our findings show new perspectives in the forensic science associated with using non-classical markers for population discrimination.

This research was financed by the National Center for Research and Development AriaDNA (OR 00 0027 12)

# No

Dominik Bujna

Faculty of Mathematics, Physics and Informatics, Univerzita Komenského, Bratislava, Slovakia

No

# Molecular Parasitology affairs.

Katarzyna Basałaj

The W.Stefanski Institute of Parasitology, PAS, Molecular Parasitology lab, Warsaw, Poland

The Witold Stefanski Institute of Parasitology is a unique national establishment conducting complex research in the field of parasitology. Laboratory of Molecular Parasitology has a distinguished tradition of gastrointestinal helminth research, especially as related to vaccine development. Currently studies are focused on novel strategies for controlling liver fluke infections. One of the ongoing projects in our laboratory is "Monoclonal recombinant antibodies for the characterisation of important parasite molecules." This project will utilise phage single-chain variable fragment (scFv) antibody display libraries for use in characterisation of important parasite proteins. Numerous human parasites cause considerable suffering worldwide, while veterinary parasites have significant impacts on the economy. Our interests lie in understanding parasite biology and the interactions that occur with the host during parasitism. To understand the biology in these often complicated host-parasite interactions it is vital to characterise the molecules involved. Two phage antibody libraries will be produced. A naïve mouse library, and a rat immune library that will be generated through infection with F. hepatica. The naïve mouse library is able to be used to generate antibodies to theoretically any antigen. The rat immune library will possess antibodies against a smaller repertoire of antigens, specifically those that are generated during an immune response. Screening of phage antibody libraries will generate antibody fragments that will be used to characterise several proteins that are proposed to be important to parasite infection. The libraries will also be used to characterise the humoral immune response in rats and to identify immunogenic proteins present on the surface of the parasite at a stage when a natural immune response can occur, possibly identifying novel vaccine candidates. In summary, this project will increase our understanding of F. hepatica biology and host immunity which will have implications for fluke control, and will lead to the development of phage antibody libraries that will enable the relatively low cost generation of monoclonal antibodies to facilitate future research against a variety of parasites. An additional implication of this project will be that no additional experimental animals will be required to obtain new monoclonal antibodies from these libraries for future studies.

# Non-invasive MPS-based detection of copy number variation using internal amplification control

Bence Széplaki

University of Szeged, Biology / BRC Szeged, Institute of Genetics, Hungary

Copy number variations have important role in genetic disorders. These kind of abnormalities, if has a somatic origin or comes from the fetus, can be detected non-invasively from a mixed, normal and affected DNA circulating in the blood plasma. NGS-based techniques enable to amplify large numbers of genomic regions in the same reaction, and gain accurate sequence data from it. After defining the chromosomal origin of the sequencing reads, the relative amount of cfDNA from the aneuploid chromosome found in the plasma is counted by comparing the number of reads mapping to the chromosomes of interest with the number of reads mapping to one or more, presumably normal, reference chromosomes. Since the compared genomic regions are sequentially different, this difference results in technical bias in direct comparism, even the quantitative targeting method requires a high number of checking points in the genome. To reduce costs, but to maintain the expected reliability, our aim was to identify specific region-pairs in the genome that have same bordering sequences for priming sites, but has limited number of mismatches for the identification of the chromosomal origin. Using such internal amplification controls, the number of the chromosomes, or the amplicons we are examining can be determined.

# CpG underrepresentation and the bacterial CpG specific DNA methyltransferase M.MpeI

Honorata Czapińska

International Institute of Molecular and Cell Biology, Poland

Cytosine methylation promotes deamination. In eukaryotes, CpG methylation is thought to account for CpG underrepresentation. Whether scarcity of CpGs in prokaryotic genomes is diagnostic for methylation is not clear. We have shown that Mycoplasms tend to be CpG depleted and to harbor a family of constitutively expressed or phase variable CpG specific DNA methyltransferases. The very CpG poor M. penetrans and its constitutively active CpG specific methyltransferase M.MpeI were chosen for further characterization. Genome wide sequencing of bisulfite converted DNA indicated that M.MpeI methylates CpG target sites both in vivo and in vitro in a locus nonselective manner. A crystal structure of M.MpeI with DNA at 2.15 Å resolution showed that the substrate base was flipped and that its place in the DNA stack was taken by a glutamine residue. A phenylalanine residue was intercalated into the "weak" CpG step of the non-substrate strand, indicating mechanistic similarities in the recognition of the short CpG target sequence by prokaryotic and eukaryotic DNA methyltransferases.

# Satellite DNA and Transposable Elements in Seabuckthorn (Hippophae rhamnoides), a Dioecious Plant with Small Y and Large X Chromosomes

Janka Puterová

Institute of Biophysics, Czech Academy of Science/ Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

Seabuckthorn (Hippophae rhamnoides) is a dioecious shrub traditionally used in the pharmaceutical, cosmetic and environmental industry as a source of oil, minerals and vitamins. The first cytogenetic studies on the seabuckthorn demonstrated either a small or large Y chromosome and information on the seabuckthorn genome is lacking. We performed Illumina DNA sequencing and reconstructed the main repetitive DNA sequences. For data analysis, we have developed a new bioinformatic approach for advanced satellite DNA analysis. We showed that about 25% of genome is represented by satellite DNA and about 24% is formed of transposable elements, dominated by Ty3/Gypsy and Ty1/Copia LTR retrotransposons. Our FISH mapping revealed X chromosome-specific, Y chromosome-specific or both sex chromosomes-specific satellites but most satellites were present on autosomes. Transposable elements were present mostly in the subtelomeres of all chromosomes. The 5S rDNA and 45S rDNA were localized on one autosomal locus each. The small size of the Y chromosome and gathering of satellite DNA there, indicate that the Y chromosome of H. rhamnoides is in a later stage of evolution but an analysis of dioecious relatives like Shepherdia would shed more light on the age of these sex chromosomes.

# Genome-wide association study reveals candidate genes determining cold-tolerance of modern maize.

Marcin Grzybowski

Departament of Plant Molecular Ecophysiology, Faculty of Biology, University of Warsaw, Poland

Temperature and light are two main factors with huge impact on plant growth and development. Improving our knowledge about this influence is the way to produce crops better adapted to various climates. Nowadays, one of the most important crop is maize. This species was domesticated in Mexico, but because of it's incredible genomic diversity it's spread all around the world today. Because of its tropical origin it needs temperature above 15°C for proper development. Consequently cold stress is the main limit for maize breeding in temperate climate of Central Europe. This problem occurs usually at spring, when plants start germinating. In extreme situation cold leads to plant death, but even partial destruction has negative impact on final yield. However, progress of artificial selection allows production of lines with greater cold tolerance. The genetic mechanism of this improvement still remains unknown. We found that cold-tolerant maize lines have significantly deeper their shoot apex below the ground. Apical meristem is the source of all above-ground organs, so maintenance of proper cell division is a key for undisturbed plant development. This trait seems to be of great importance, because even small layer of soil can mitigate temperature drops and protects meristem cells. This phenotype fit well with better cold tolerance of some inbred lines. The possible explanation why some maize lines are able to avoid cold damage by hiding the shoot apex deep in a soil while others are not, is difference in blue and/or red light reception resulting in diversity of length of seedling mesocotyl and thus depth location of apical meristem. Using Next Generation Sequencing we recently re-sequenced genomes of three maize inbred lines produced in frame of one of Polish breeding program. Combining these results with data for 123 lines from HapMap3 project (www.panzea.org) we have performed genome wide association study (GWAS). Analysis of length of mesocotyl in plants grown under different light spectra shows that this trait is associated with 5 loci. The most significant result is close to gene coding transcription factor involving in response to ethylene. This plant hormone is responsible for e.g stem shortening, so it may be involved in stopping the mesocotyl elongation, as well. Small genetic variants can modulate this pathway so better understanding their influence on plant development is important step in production crops better adapted to different climates.

# Whole-exome sequencing as a tool in identifying mutations associated with T-cell lymphoblastic lymphoma development in WOM/W mice

Aleksandra Grochowska

Medical Center for Postgraduate Education, Department of Gastroenterology, Hepatology and Clinical Oncology, Roentgena 5, 02-781 Warsaw, Poland

Background: T-cell lymphoblastic lymphoma (T-LBL) is a rare, heterogeneous and clinically very aggressive neoplasm. Its molecular pathogenesis is not fully known. Since human and murine T-LBLs are histologically and phenotypically comparable, mouse models are usually used to study T-LBL pathogenesis. 90% of WOM/W mice breed in Cancer-Center-Institute commonly develop T-LBL. To date, the genetic background of this strain was not evaluated. Aim: To identify genetic variants within coding part of the genome, exome, that may lead to T-LBL development in WOM/W mice. Materials and methods: Exomes of WOM/W and BN/aW (strain resistant to develop hematopoietic neoplasms) mice were sequenced with the use of Ion Proton (Life Technologies) system. Exomes capture and DNA libraries preparation was performed with SureSelectXT Mouse All Exon Kit (Agilent). Bioinformatic analysis was conducted in R environment with the use of Bioconductor software and KEGG and MGD databases. Results: Variants specific for WOM/W strain were identified in genes related to T cells development. Mutations with potentially negative impact on encoded proteins were identified in genes including: Plcg1, Cd3d, Unc13d, H2-M3. Conclusions: We identified genetic variants that may be involved in abnormal T-cells development and its uncontrollable proliferation. Further studies are warranted to verify functional significance of identified mutations.

# Analysis of rDNA variability, evolution and transcriptional activity based on de novo sequencing of the yeast S. cerevisiae rDNA gene array.

Agnieszka Czarnocka-Cieciura

University od Warsaw, Poland

One of the most energentically demanding process in eukaryotic cells is ribosome biosynthesis. To facilitate the synthesis of ribosomal RNA(rRNA), rDNA genes are encoded in a tandem array of repeating units. In yeast Saccharomyces cerevisiae, all rDNA copies are located in a single loci on chromosome XII. The average number of rDNA units is estimated as 200 copies, but it may vary from 30 to 350 and only part of them is transcriptionally active. The repetitive nature of rDNA loci makes the assembly of this part of the genome a real challenge. Thus, despite the fact that yeast were the first eukaryotes to be sequenced, there are only a few consensus rDNA sequences available and little is known about the variability between each rDNA copy. Single rDNA unit in S. cerevisiae has approximately 9 kb and encodes three mature rRNA molecules (18S, 5.8S and 25S) that are flanked by external and internal transcribed sequences (ETS and ITS regions). In addition, within rDNA genes there are several short potential ORFs present on the opposite strand. At least one of them (TAR1) is coding a short polypeptide, but expression of the other units is questionable and their putative function remains unknown. To better understand the variability within the rDNA tandem array we are sequencing de novo this part of the yeast genome utilizing very long reads available in the PacBio sequencing approach. After obtaining reference sequence of this loci we will estimate transcriptional activity of different rDNA copies in several growth conditions by RNAseq. Finally, we plan to investigate the variability between the copies of known and putative unknown antisense ORF's encoded in rDNA sequences and in the long run to determine their transcription activity and function.

# Quality assessment of RNA 3D models optionally using restraints from high-throughput sequencing experiments

Marcin Magnus

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland

The understanding of the importance of RNA molecules has dramatically changed over the recent years. As in the case of proteins, the function of an RNA molecule is encoded in its three-dimensional structure, which in turn is determined by the molecule's sequence.

Therefore, there is a need to develop computational methods that are able to provide reliable models of RNA molecules based only on nucleic acid sequences. A standard workflow of structure prediction looks as follows: based on a nucleic acid sequence, a set of RNA 3D models is produced, the quality scores of models are computed, and based on the scores the final prediction is made.

In this report, we present a new computer program for assessing the quality of RNA 3D models, mqapRNA. We developed a machine learning based statistical model to predict the quality of analyzed structural models. mqapRNA for each structural models calculates a set of energies and measurements are taken as: radius of gyration, clash and secondary structure agreement between secondary structure of a model and secondary structure predicted for a model's sequence. With the use of above-mentioned values and the RMSD between a given model and the native structure, a deep learning model was developed. Later, this model was benchmarked using testing decoys yielding predicted RMSDs that can be then used as a measurement of the quality of the model (large predicted RMSD equals low quality). The highest correlation with RMSDs was observed for predictions generated by our model that makes the method a useful tool for assessing the quality of models. Our method was able to correctly identify, in the set of alternative models, models similar to the native structure. mqapRNA can take also optimally spatial restraints obtained from high-throughput sequencing experimental technique such as MOHCA-Seq [1] and/or Mutate-and-Map [2].

To the best of our knowledge, this is the first time where such analysis has been performed and based on the results a computer program has been developed. At this stage of testing, it appears to be a promising tool for RNA 3D structure prediction.

[1] Cheng, C.Y., Chou, F.-C., Kladwang, W., Tian, S., Cordero, P., and Das, R. (2015) "Consistent global structures of complex RNA states through multidimensional chemical mapping." eLife

[2] Cordero, P., and Das, R. (2015) "Rich structure landscapes in both natural and artificial RNAs revealed by mutate-and-map analysis" PLOS Computational Biology

# Ser181 phosphorylation – a possible factor changing the composition of ZBP1 RNA granules

Anna Urbańska

Laboratory of Molecular and Cell Neurobiology, International Institute of Molecular and Cell biology in Warsaw, Poland

Neurons can properly process information thanks to their characteristic and well defined shape. Aberrations in neuronal morphology occur in various neuropathologies, like Autism Spectrum Disorders. One of mechanisms contributing to neuronal development is local translation, which enables protein synthesis distantly from cell soma – in distal parts of developing dendrites or axons. mRNAs to be locally translated, are bound in nucleus by several mRNA binding proteins and distributed throughout dendrites in their dormant form as RNA granules. After extracellular stimuli, mRNAs are released from RNPs and translated. Zipcode Binding Protein 1 (ZBP1) is one of mRNA binding proteins. We have previously shown that it is crucial for proper development of dendrites. Although it binds a wide variety of mRNAs, it was mostly investigated as a protein that represses ß-actin mRNA. The only known, so far, regulator of ZBP1 is Src kinase, which phosphorylates ZBP1 at Tyr396, resulting in ß-actin mRNA release. Here we show that ZBP1 is also phosphorylated at Ser181, possibly by mTORC2. This modification is important for proper development of neurons and distribution of ZBP1 in dendrites. We have also discovered the correlation between phosphorylation of Ser181 and mRNA binding by ZBP1. It becomes dephosphorylated at Ser181 after stimulations resulting in mRNA release. We speculate that this modification occurs not after complete disassembly of ZBP1 RNA granule, but it can rather change its composition. This hypothesis is supported by data showing that ZBP1 that cannot bind mRNA does not form granules and ZBP1, both phosphorylated and dephosphorylated at Ser181, form granules. Therefore we would like to describe differences between RNA compositions of ZBP1 phosphorylated and non phosphorylated at Ser181. Obtained data will be the first to describe changes in RNA granule composition due to post translational modification of RNA binding protein.

# Estimating Sequence Similarity from Read Sets for Clustering Sequencing Data

Petr Ryšavý

Dept. of Computer Science, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

Clustering biological sequences is one of the most popular tasks in bioinformatics. Input of a clustering algorithm usually contains a set of sequences. Those sequences are product of reading a DNA. The sequencing machines usually do not read the whole sequence at once, but they read short substrings (named reads) that contain tens to hundreds of symbols.

Common approach to cluster those NGS read-bags is to assemble the reads to estimates of the original sequences first and then run a clustering algorithm. The sequence assembly problem is NP-hard while hierarchical clustering algorithms usually require only a polynomial number of steps. Our goal is to avoid the assembly step and cluster the set of read-bags directly.

If we look in detail to the hierarchical clustering algorithms (for example UPGMA or neighbor joining), we can notice that they rely only on distance matrix between the clustered objects. Therefore our approach is based on approximating the distance matrix using the read-bag representation of the sequences. We approximate the Levenshtein distance between the sequences. We also assume that the reads are distributed uniformly on the set of all possible substrings of given length.

Our approach is based on Monge-Elkan distance known from the field of databases. For each read in the first read-bag we find the least distant read in the second read-bag. The reads are compared by a modified Levenshtein distance that matches the sequencing motivation behind. Then we average distances between the paired reads and scale.

In experiments, our approach gives better estimates of the distances between the sequences than the classical first-assemble-then-cluster approach. As a result it produces better clusterings. Moreover our approach avoids the NP-hard step and is coverage**2 times slower than clustering of the true sequences.

For more details, please check paper Ryšavý P., Železný F.: Estimating Sequence Similarity from Read Sets for Clustering Sequencing Data. Draft is available on http://ida.felk.cvut.cz/zelezny/pubs/ida20 The paper was accepted to IDA 2016 conference.

# Integration platform for instantaneous processing and analysis of microscopy images software

Agnieszka Gromadka

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw (Poland); Institute of Fundamental Technological Research PAN, Poland

Fluorescent microscopy constitutes one of the key components of modern biology. It allows not only to observe biological processes at the cellular and molecular scale but also provides quantitative measurements of studied processes. Progressed to the high throughput capability available microscopes allow to scan behaviour of studied systems across numerous experimental conditions at relatively little time and labour cost. Such high-throughput microscopy provides gigabytes of images within minutes. Efficient extraction of relevant information requires therefore advanced numerical techniques and extensive computational resources. Fortunately, a plethora of tools has been already developed. Nevertheless, a majority of available software packages is designed to solve very specific problems and exhibit a rather limited flexibility. As a result sophisticated image analysis processes are difficult to perform automatically at the high-throughput scale as tools performing different elements of the image analysis pipeline cannot be easily integrated. Therefore we have developed IPIQA, a computational platform that enables integration of most commonly used image analysis packages in one image processing pipeline to enable flexible, instantaneous image analysis on the high-throughput scale. The major advantage of our package is that it facilitates most relevant tools already used for microscopy image processing, including CellProfiller and ImageJ. Moreover, IPIQA provides a user-friendly programming interface with R-environment, python and ImageJ Macro Language to enable adaptation of individual scripts. As a result all essential functionalities needed for extraction of relevant information from microscopy data can be flexibly accommodated. Given the increasing role of high-throughput microscopy our package contributes to filling an important niche in the spectrum of available image analysis tools. In our laboratory it already led to substantial savings in terms of labour and computation time. Therefore we believe that it will lead to similar improvements in other laboratories performing quantitative high-throughput microscopy experiments.

# Identification RNA Viruses in Sequence Data

Dávid Smoľak

Biology (Molecular Biology), Faculty of Natural Sciences, Comenius University Bratislava, Slovakia

Aproaches to sequencing of RNA viruses -Sequencing of isolated viruses -Strategies sequencing straight from host´s tissue -Mapping to known viruses (describing algorithms) -Identification of new viruses by de novo sequencing (describing de novo algorithms, their comparations and using in virological works) -Strategies of overrepressented viruses sequences (in my case: removing rRNA, mapping host´s sequences)

P.S. This is the scheme of my bioinformatics part of my next year bachelors work (last version can be little different). I like to start and finish my bioinformatics part this summer and then I would like to use this information to create the poster. The poster as well as general NGS School could be great experience for me before my bachelor habilitation.