Friday, July 30th	
9:30–10:00	**morning coffee**
10:00–11:30	**early morning session**
10:00	Introduction
	EVOLUTION
10:15	The world of viruses, its global organization and evolution *Eugene Koonin* *NCBI*
10:45	Heterogeneity of the GFP fitness landscape and data-driven protein design *Fedor Kondrashov* *Institute of Science and Technology Austria*
11:15	A billion-year trend of amino acid substitutions in the mitochondrial genome *Alina G. Mikhailova* *IKBFU*
11.30–12:00	**coffee break**
12:00–13:30	**late morning session**
	NON-CODING RNA
12:00	Integrative genomic discovery and drugging of long noncoding RNAs in cancer *Rory Johnson* *University College Dublin*

12:30	Translational landscape of human 3’ UTRs *Pavel V Baranov* *University College Cork*
13:00	Disruption of uORFs translation is an underestimated cause of Mendelian disorders *Alexandra Yu. Filatova* *Research Centre for Medical Genetics*
	SPONSORED TALK
13:15	The sbvIMPROVER Metagenomics Diagnostics for Inflammatory Bowel Disease Challenge: Results and Lessons Learned *Lusine Khachatryan* *PMI Science*
13:30–16:00	**lunch**
13:30–14:25	**SYSTEMS BIOLOGY SESSION**
13:30	Introduction
13:40	Interpretable model for chromatin interaction probability prediction *Alexandra Galitsyna* *Skolkovo Institute of Science and Technology*
13:55	Suppression of reverse transcriptase-driven chimeric cDNA synthesis *Alexandr Gordeev* *Institute of Protein Research RAS*

14:10	Bi-directional study chromatin organization into the eukaryotic nuclei cell *Pavel Kos* *Moscow State University*
14:30–16:00	**ONLINE POSTER SESSION**
	presenters: A–F
16:00–17:30	**early evening session**
	3D CHROMATIN
16:00	Loop extrusion on busy DNA *Leonid Mirny* *MIT*
16:30	Unraveling the structure of paired homologs and sister chromatids with Hi-C and polymer modeling *Anton Goloborodko* *IMBA*
17:00	preciseTAD: a machine-learning framework for predicting boundaries of 3D genomic elements *Mikhail G Dozmorov* *Virginia Commonwealth University*
17:15	Anopheles mosquitoes revealed new principles of 3D genome organization in insects *Igor V Sharakhov* *Virginia Tech*
17:30–19:30	**coffee break and** **OFFLINE POSTER SESSION**
19:30–…	**welcome party**

Saturday, July 31st	
9:30–10:00	**morning coffee**
10:00–11:30	**early morning session**
	MACHINE LEARNING
10:00	Federated Machine Learning *Jan Baumbach* *University of Hamburg*
10:30	Network enriched analysis of complex biological time-series data *Richard Röttger* *University of Southern Denmark*
11:00	Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis *Denis V Antonets* *SRC VB "Vector"*
11:15	Multi-platform cross-harmonization of gene expression profiles obtained using mRNA next-generation sequencing and microarray hybridization *Nicolas M. Borisov* *Moscow Institute of Physics and Technology*
11.30–12:00	**coffee break**
12:00–13:30	**late morning session**
	SINGLE CELLS — 1

12:00	TBA *Rahul Satija* *New York Genome Center*
12:30	Unraveling transcriptional intratumoral heterogeneity at single cell level *Andrei Zinovyev* *Institut Curie*
12:45	Single nucleus RNA-sequencing data reveals intra-tumoral heterogeneity in medulloblastoma brain tumors with extensive nodularity *Konstantin Okonechnikov* *German Cancer Research Center*
13:00	Variation of mutation rate between individual cells *Maria A Andrianova* *Skolkovo Institute of Science and Technology*
13:15	Joint model of RNA velocity on transcriptional manifold *Ruslan Soldatov* *Harvard Medical School*
13:30–16:00	**lunch**
13:30–14:10	**SYSTEMS BIOLOGY SESSION**
13:30	Identification and separation of sources of transcriptional variability in single-cell RNA-seq data *Konstantin Zaitsev* *ITMO University*

13:40	Origin of helper T cell diversity in human immunity: the study of TCR repertoires *Sofya Kasatskaya* *Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry*
13:50	"Investigation of T. thermophilus Type III CRISPR-Cas mediated bacterial defense mechanisms” *Daria Artamonova* *Skolkovo Institute of Science and Technology*
14:00	Direct reprogramming of human skin fibroblasts into insulin-producing cells using CRISPR/dCas9-SunTag-VP64 effector epigenome editing systems *Alexandr Artyuhov* *Pirogov Russian National Research Medical University*
14:10	Rational design of chimeric portal proteins of viruses *Evgeny Klimuk* *Institute of Molecular Genetics RAS*
14:30–16:00	**ONLINE POSTER SESSION**
	presenters: G–Ma
16:00–17:30	**early evening session**
	POST-TRANSCRIPTION
16:00	Conserved long-range base pairings are associated with pre-mRNA processing of human genes *Dmitri Pervouchine* *Skolkovo Institute of Science and Technology*

16:15	An extended catalogue of tandem alternative splice sites in human tissue transcriptomes *Aleksei Mironov* *Skolkovo Institute of Science and Technology*
16:30	Genome-wide search for pathogenic splice variants manifested in different genetic backgrounds *Mikhail Y Skoblov* *Research Centre for Medical Genetics*
16:45	Assessing proteome diversity raised by alternative splicing in Brain RNA-Seq data *Vladimir N Babenko* *Institute of Cytology and Genetics SB RAS*
17:00	Proteome recoding by A-to-I mRNA editing *Sergei Moshkovskii* *Pirogov Russian National Research Medical University*
17:15	Fast gene set enrichment analysis with multi-level Monte-Carlo approach *Alexey A Sergushichev* *ITMO University*
17:30–18:00	**coffee break**
18:00–19:15	**late evening session**
	SINGLE CELLS — 2

18:00	Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data *Gary Bader* *University of Toronto*
18:30	Cell segmentation in spatial transcriptomics *Peter Kharchenko* *Harvard Medical School*
19:00	Multimodal by default: designing the next generation of methods and tools for single-cell multi-omics *Danila Bredikhin* *EMBL*

Sunday, August 1st	
9:30–10:00	**morning coffee**
10:00–11:30	**early morning session**
	GENETICS — 1
10:00	Biological design and prediction using evolutionary information *Deborah Marks* *Harvard Medical School*
10:30	The missing link between genetic association and regulatory function *Shamil Sunyaev* *Harvard Medical School*
11:00	Peeking inside the clockworks of regulatory single-nucleotide variants with ADASTRA and ANANASTRA *Ivan V Kulakovskiy* *Institute of Protein Research*
11:15	Within-gene epistatic selection in genetically diverse populations *Anastasia V Stolyarova* *Skolkovo Institute of Science and Technology*
11.30–12:00	**coffee break**
12:00–13:30	**late morning session**
	GENETICS — 2

12:00	Estimating the timing of multiple admixture events using 3-locus Linkage Disequilibrium *Mikhail Shishkin* *HSE University*
12:15	Quantitative genetics of human protein N-glycosylation *Yurii Aulchenko* *Institute of Cytology and Genetics SD RAS*
12:30	Mapping genes involved in control of N-glycosylation of blood glycoproteins through a large genome-wide association study *Sodbo Sharapov* *Institute of Cytology and Genetics SB RAS*
12:45	Targeted sequencing of 242 clinically important genes in a sample of Russian population from Ivanovo region *Vasily Ramensky* *National Medical Research Center for Therapy and Preventive Medicine of the Ministry of Healthcare of the Russian Federation*
13:00	Mutational spectrum of the human mitochondrial genome: somatic and germline signatures differ *Dmitry Iliushchenko* *Immanuel Kant Baltic Federal University*
13:15	Precise mutation rate estimation on an individual site level *Vladimir Seplyarskiy* *Harvard Medical School*

13:30–16:00	**lunch**
13:30–14:20	**SYSTEMS BIOLOGY SESSION**
13:30	Integration of genomics and metabolomics data for discovery of small natural products *Alexey Gurevich* *Saint-Petersburg State University*
13:40	Mechanistic model of mutation accumulation based on differential kinetics of DNA repair *Maria Andrianova* *Skolkovo Institute of Science and Technology*
13:50	Investigation of the genetic architecture of human traits on the phenome scale *Yury Barbitoff* *Saint-Petersburg State University*
14:00	Transcriptional interference by RNA-guided DNA-binding proteins *Aleksei Agapov* *Institute of Molecular Genetics*
14:10	High-throughput analysis of short DNA fragments generated in vivo in *Escherichia coli* and *Saccharomyces cerevisiae* *Anna Shiriaeva* *Skolkovo Institute of Science and Technology/ St. Petersburg Branch of Vavilov Institute of General Genetics*
14:30–16:00	**ONLINE POSTER SESSION**
	presenters: Me–Shp

16:00–17:30	**early evening session**
	STRUCTURE
16:00	Intragenic compensatory variants: does the distance matter? *Nadezhda Azbukina* *Lomonosov Moscow State University*
16:15	New algorithms for finding combinatorially complete datasets in high-throughput mutagenesis experimental data *Dmitry N Ivankov* *Skolkovo Institute of Science and Technology*
16:30	HLA binding preferences of self-peptides are biased towards proteins with specific molecular functions *Vadim Karnaukhov* *Skolkovo Institute of Science and Technology*
16:45	Co-aggregation of amyloids: from structural principles to in silico prediction *Andrey V. Kajava* *CNRS*
17:00	Functional domain annotation of protein sequences with deep metric learning *Nikolay Russkikh* *AcademGene LLC*
17:15	MolDiscovery: Learning Mass Spectrometry Fragmentation of Small Molecules *Azat Tagirdzhanov* *St. Petersburg State University*
17:30–18:00	**coffee break**

18:00–19:30	**late evening session**
	BOTANICAL GARDEN AND ZOO
18:00	Models for crop species diversification within and between centers of domestication *Anna Igolkina* *Peter the Great St.Petersburg Polytechnic University*
18:15	Genomic blueprints of flax diversification and improvement *Maria Duk* *Peter the Great St.Petersburg Polytechnic University*
18:30	Numerical study of the flowering time model of wild chickpeas *Andrey Y Ageev* *Peter the Great St.Petersburg Polytechnic University*
18:45	Artificial intelligence applied to population genomics and paleogenomics *Olga Dolgova* *Center for Genomic Regulation*
19:00	A systems biology approach to understanding SARS-CoV-2 transmissibility in population *Sofija Markovic* *University of Belgrade*
19:15	A single chromosome driving rapid extremotolerant adaptation in an insect *Yuki Yoshida* *University of Tokyo*

Monday, August 2nd	
9:30–10:00	**morning coffee**
10:00–11:30	**early morning session**
	CELLS AND SIGNALS
10:00	Quantifying and Manipulating the Aging Process *Vadim N Gladyshev* *Harvard Medical School*
10:30	Identification of phenotype-specific networks from paired gene expression-cell shape imaging data *Evangelia Petsalaki* *EMBL*
11:00	Which wiring of the cell cycle opens the path to mitotic and meiotic division? *Ovidiu Radulescu* *University of Monpellier*
11:15	Learning the cell states using deep neural networks for cell type-specific interpretation of disease variants *Veniamin Fishman* *Sberbank AI Lab*
11.30–12:00	**coffee break**
12:00–13:30	**late morning session**
	COMPARATIVE AND FUNCTIONAL GENOMICS

12:00	Computational methods for genome interpretation *Emidio Capriotti* *University of Bologna*
16:30	Laying foundations for very large-scale comparative genomics *Christophe Dessimoz* *Swiss Institute for Bioinformatics*
17:00	Bringing Human Population Genetics to Protein Structure *Geoff Barton* *University of Dundee*
13:30–16:00	**lunch**
13:30–14:30	**SYSTEMS BIOLOGY SESSION**
13:30	**Awards ceremony**
13:40	The investigation of Mitochondrial Genomes of Helianthus species *Maksim Makarenko* *Institute for Information Transmission Problems*
13:50	High-Throughput Splicing Efficiency Analysis System *Sofia Mariasina* *Moscow State University/Skolkovo Institute of Science and Technology*
14:00	Computational methods for unsupervised demographic inference of multiple populations from genomic data *Ekaterina Noskova* *ITMO University*

14:10	The function and structure of pseudo-nucleus in jumbo phages *Alexei Samolygo* *Skolkovo Institute of Science and Technology*
14:20	Systematic identification and validation of novel RiPP biosynthetic gene clusters encoding YcaO-domain enzymes *Dmitrii Travin* *Skolkovo Institute of Science and Technology*
14:30–16:00	**ONLINE POSTER SESSION**
	presenters: Shu–Z
16:00–17:30	**early evening session**
	DEVELOPMENT
16:00	Pluripotency factors in zebrafish embryo *Daria Onichtchouk* *Albert-Ludwigs-University of Freiburg*
16:30	Optimal decoding of cellular identities in a genetic network *Gasper Tkacik* *Institute of Science and Technology Austria*
17:00	Early developmental asymmetries in cell lineage trees in living individuals *Alexej Abyzov* *Mayo Clinic*
17:15	The digenean complex life cycle: phylostratigraphy analysis of molecular signatures *Maksim A Nesterenko* *Saint Petersburg University*

17:30–18:00	**coffee break**
18:00–19:30	**late evening session**
	BACTERIA
18:00	Chromosome conformation of the hyperthermophilic archaeon *Thermofilum adornatus* *Andrey Sobolev* *Institute of Gene Biology RAS*
18:15	Method for detection of parallel adaptation with genome rearrangements in bacterial populations *Nikita Alexeev* *ITMO University*
18:30	Evolutionary benefits of bacterial genomes with chromids *Olga Bochkareva* *IST Austria*
18:45	Black holes in the regulation of biofilm formation in *Escherichia coli* *Maria Tutukina* *Skolkovo Institute of Science and Technology*
19:00	A>G is a hallmark of oxidative damage in mitochondrial and bacterial genomes *Konstantin Popadin* *Ecole Polytechnique Federale de Lausanne*
19:15	Investigating microbial diversity of spontaneous fermentation beer and cider using Hi-C metagenomics *Ignat V. Sonets* *Institute of Gene Biology RAS*

19:30–…	**farewell party**

Friday, July 30, 10:15
The world of viruses, its global organization and evolution
Eugene Koonin (NCBI)

Friday, July 30, 10:45

Heterogeneity of the GFP fitness landscape and data-driven protein design

Fedor Kondrashov (Institute of Science and Technology Austria)

Understanding the relationship between genotype and phenotype, the fitness landscape, elucidates the fundamental laws of heredity and evolution may ultimately create novel methods of protein design. The fitness landscape is often conceptualised as multidimensional with one dimension representing fitness, or another phenotype, and the other dimensions each representing a genotype's locus. Absolute knowledge of the fitness landscape would reveal the phenotypes conferred by any arbitrary genotype, with immense and obvious practical implications. However, sparse experimental data, and the concomitant lack of understanding of the rules by which fitness landscapes are formed, limit the accuracy of phenotype predictions based on sequence alone. We characterized the fitness peaks of four orthologous fluorescent proteins with a broad range of sequence divergence. While two of the four studied fitness peaks were sharp, the other two were considerably flatter, being almost entirely free of epistatic interactions. Counterintuitively, mutationally robust proteins, characterized by a flat fitness peak, were not optimal templates for machine-learning-driven protein design – instead, predictions were more accurate for fragile proteins with epistatic landscapes. Our work paves the way for practical use of experimental fitness landscapes in protein engineering.

Friday, July 30, 11:15

A billion-year trend of amino acid substitutions in the mitochondrial genome

Alina G. Mikhailova (IKBFU)

It has been shown that the rates of reciprocal amino acid substitutions in prokaryotic and eukaryotic organisms are not balanced, leading to the long-term increase (i.e. ‘gainers’) or decrease (i.e. ‘losers’) in the frequency of some amino acids. However, the evolutionary driving forces establishing this trend (mutagenesis or selection) are still unknown. Here we focus on the mitochondrial genome, where we can predict the preferential direction of amino acid substitutions, based on the strongly asymmetrical mutational spectrum.

Friday, July 30, 12:00

Integrative genomic discovery and drugging of long noncoding RNAs in cancer

Rory Johnson (University College Dublin)

Friday, July 30, 12:30

Translational landscape of human 3’ UTRs

Pavel V Baranov (University College Cork)

With rare exceptions, mRNA molecules of human genes and other eukaryotic organisms are known to contain a single long Open Reading Frame (ORF) encoding the main protein product of corresponding genes. Other regions of mRNA have been termed as 5’ and 3’ Untranslated Terminal Regions (UTRs). Ribosome profiling studies supported by proteomics, and more recently CRISPR–Cas screenings, revealed productive translation of numerous short ORFs occurring in the 5’ UTRs. On the contrary, and notwithstanding a handful of documented stop codon readthrough (SCR) cases and recent reports of translated downstream ORFs (dORFs), human 3’ UTRs are believed to be the least translated. The comparatively low ribosome footprint density in 3’UTRs is believed to be largely due to technical or biological noise.

Friday, July 30, 13:00

Disruption of uORFs translation is an underestimated cause of Mendelian disorders

Alexandra Yu. Filatova (Research Centre for Medical Genetics)

In the present work, we study disruption of upstream ORFs translation in the context of Mendelian diseases. Despite the fact that a large number of human genes is thought to include functional uORFs, integration of their analysis into the DNA-diagnosis algorithms for patients is a challenge, since there is no precise annotation of uORFs. Using different bioinformatics approaches we performed manually curated analysis of uORFs in ~1’600 human Mendelian disease-causing genes from OMIM database. In addition, we performed an experimental study of uORFs in two Mendelian-disease associated genes (PAX6 and NF1) and showed that patient-derived nucleotide variants disrupting uORFs lead to development of the diseases. Our results demonstrate the importance of 5′UTR analysis during DNA diagnosis in patients with hereditary diseases.

Friday, July 30, 13:15

The sbvIMPROVER Metagenomics Diagnostics for Inflammatory Bowel Disease Challenge: Results and Lessons Learned

Lusine Khachatryan

Inflammatory bowel diseases (IBD) constitute a spectrum of chronic inflammatory disorders that recurrently affect the gastrointestinal tract. Ulcerative colitis (UC) and Crohn's disease (CD) are the two main clinically defined manifestations of IBD, each with distinctive clinical and pathological features. A growing number of reports showing the alteration of gut microbiota in subjects with IBD indicate the potential benefit of exploiting metagenomics for non-invasive IBD diagnostics.

Friday, July 30, 13:40

Interpretable model for chromatin interaction probability prediction

Alexandra Galitsyna (Skolkovo Institute of Science and Technology)

Friday, July 30, 13:55

Suppression of reverse transcriptase-driven chimeric cDNA synthesis

Alexandr Gordeev (Institute of Protein Research RAS)

Friday, July 30, 14:10

The study of Mildronate effect on mitochondrial metabolism during physical exercise

Artem Gureev (Voronezh State University)

Friday, July 30, 14:25

Bi-directional study chromatin organization into the eukaryotic nuclei cell

Pavel Kos (Moscow State University)

Friday, July 30, 16:00
Loop extrusion on busy DNA
Leonid Mirny (MIT)

Friday, July 30, 16:30

Unraveling the structure of paired homologs and sister chromatids with Hi-C and polymer modeling

Anton Goloborodko (IMBA)

Friday, July 30, 17:00

preciseTAD: a machine-learning framework for predicting boundaries of 3D genomic elements

Mikhail G Dozmorov (Virginia Commonwealth University)

Chromosome conformation capture technologies (Hi-C) revealed extensive DNA looping into Topologically Associating Domains (TADs) and chromatin loops. The relatively low resolution of Hi-C data prevents precise mapping of domain boundaries. However, the high resolution of genomic annotations associated with boundaries, such as CTCF, suggests they can inform the precise boundary location. Several methods attempted to use genome annotation data for domain boundary prediction; however, they overlooked key characteristics of the data, such as spatial associations between an annotation and a boundary and a much smaller number of boundaries than the rest of the genome (class imbalance).

We developed preciseTAD, an optimized random forest model to improve the location of domain boundaries. Trained on high-resolution genome annotation data and boundaries from low-resolution Hi-C data, the model predicts the location of boundaries at base-level resolution. We investigated several feature engineering and resampling techniques to select the most optimal data characteristics and address class imbalance. Density-based clustering (DBSCAN) was used to identify the precise location of boundary regions and summit points.

We found that the spatial relationship between boundaries and annotations and random under-sampling significantly improved model performance. Predicted boundaries were more enriched for CTCF, RAD21, SMC3, and ZNF143 signal and more conserved across cell lines, highlighting their higher biological significance. Using cell line-specific genomic annotations, the pre-trained models enable detecting domain boundaries in cells without Hi-C data.

Our study implements the R package (https://bioconductor.org/packages/preciseTAD) and the pre-trained models for precise domain boundary prediction using genome annotation data. The precise identification of domain boundaries will improve our understanding of how genomic regulators are shaping the 3D structure of the genome.

Friday, July 30, 17:15

Anopheles mosquitoes revealed new principles of 3D genome organization in insects

Igor V Sharakhov (Virginia Tech)

Chromosomes are hierarchically folded within cell nuclei into territories, domains and subdomains, but the functional importance and evolutionary dynamics of these hierarchies are poorly defined. In disease vectors, such as mosquitoes, nuclear architecture may modulate gene expression underlying epidemiologically relevant traits. Here, we comprehensively profiled genome organizations of five Anopheles mosquito species and showed how different levels of chromatin architecture influence contacts between genomic loci. Our Hi-C maps demonstrated strong centromere-centromere, telomere-telomere, and intra-chromosome interactions indicating Rabl-like configuration of chromosome territories. Patterns observed on Hi-C maps are associated with known cytological structures, epigenetic profiles, and gene expression levels. Evolutionary analysis revealed conservation of chromatin architecture within synteny blocks for dozens of millions of years and enrichment of synteny breakpoints in regions with increased genomic insulation, associated with gene-rich environments. At the level of individual loci, we identified specific, extremely long-ranged looping interactions, conserved for ~100 million years. The anchors of the largest loops are not associated with the clustering of active genes and also display low levels of H3K27me3 enrichment, which indicates that they do not correspond to Polycomb-mediated loops. Our data suggest that large loops found in Anopheles are formed by other, yet unknown, mechanisms. Overall, our results provide a new framework for understanding of how genomes are organized and function within the nuclear space in insects.

Saturday, July 31st, 10:00
Federated Machine Learning
Jan Baumbach (University of Hamburg)

Saturday, July 31st, 10:30

Network enriched analysis of complex biological time-series data

Richard Röttger (University of Southern Denmark)

Saturday, July 31st, 11:00

Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis

Denis V Antonets (SRC VB "Vector")

The transcriptomic data are being frequently used in the research of biomarker genes of different diseases and biological states. The most common tasks there are the data harmonization and treatment outcome prediction. Both of them can be addressed via the style transfer approach. Either technical factors or any biological details about the samples which we would like to control (biological state, treatment, etc.) can be used as style components. The proposed style transfer solution is based on Conditional Variational Autoencoders, Y-Autoencoders and adversarial feature decomposition. To quantitatively measure the quality of the style transfer, neural network classifiers which predict the style and semantics after training on real expression were used. Comparison with several existing style-transfer based approaches shows that proposed model has the highest style prediction accuracy on all considered datasets while having comparable or the best semantics prediction accuracy.

Saturday, July 31st, 11:15

Multi-platform cross-harmonization of gene expression profiles obtained using mRNA next-generation sequencing and microarray hybridization

Nicolas M. Borisov (Moscow Institute of Physics and Technology)

Unlike DNA sequences, transcriptomic profiles (i.e., expression levels of certain genes) are hard to compare between experiments, due to different devices and reagent kits used for such profiling. To achieve uniformity and comparability of expression profiles, they must be either normalized (i.e. made by some means comparable), or harmonized (i.e. transformed into a reference shape associated with a gold standard platform).

In 2019, we introduced Shambhala, a universal harmonizing method for gene expression data, which can simultaneously process unlimited number of profiles and experimental platforms. For that first time, Shambhala was based on a piecewise linear normalization method XPN. Very recently, a competing approach CuBlock was reported; it uses piecewise cubic, rather than linear, transformation.

That is why we performed Shambhala-assisted harmonization in piecewise cubic mode for a set of cancer and healthy transcriptome profiles. We took cancer profiles from the GDC (Genomic Data Commons) Legacy Archive, as well as from our own repository; the healthy ones – from Gene-Tissue Expression (GTEx) project, as well as from our own sample bank. The whole transcriptome profile set accounted for more than 14,000 profiles.

We evaluated the quality of cross-platform harmonization using our watermelon multisection (WM) method. The WM metric assesses the quality of class separation in cluster dendrograms. A perfect harmonization should produce better class separation according to sample type rather than according to experimental platform. The quality of Shambhala harmonization in cubic mode is not worse than for CuBlock, and essentially better than for within-platform normalization methods, such as quantile normalization (QN), or differential gene expression analysis (DESeq2). However, Shambhala harmonizes each profile independently from other profiles, which is a clear advantage over CuBlock for the task of big data conversion to a reference gold standard.

Saturday, July 31st, 12:00
TBA
Rahul Satija (New York Genome Center)

Saturday, July 31st, 12:30

Unraveling transcriptional intratumoral heterogeneity at single cell level

Andrei Zinovyev (Institut Curie)

Single cell approaches change the way we can look at the variety of cell states within a biological sample. In particular, it becomes possible to characterize the sources of intratumoral heterogeneity and the mechanisms of resistance to cancer treatment. In genomically stable Ewing sarcoma driven by a chimeric transcription factor EWS-FLI1, the intratumoral heterogeneity is almost exclusively attributed to epigenetic mechanisms, which we studied in several model systems using single cell RNASeq and other types of molecular data. In particular, we used a doxycline-regulated, shRNA-based system that enables to control for EWS-FLI1 expression in Ewing cells. We sequenced the transcriptomes single cells at different time points during EWS-FLI1 expression in vitro and in vivo and from patient derived xenograft (PDX). Using Independent Component Analysis, we managed to distinguish, in a purely unsupervised setting, the transcriptional programs of G1/S and G2/M phases of the cell cycle and deconvolute them from the proper action of an oncogene. We characterized the epigenetic heterogeneity of Ewing PDX and concluded that there exists a specific combination of biological factors inside the tumor promoting cell proliferation. We also characterized in details the properties of cell cycle trajectory in Ewing sarcoma cell lines, which is the principal source of epigenetic heterogeneity. This analysis further highlighted the interplay between the cancer driver oncogene and the mechanisms of cell cycle.

Saturday, July 31st, 12:45

Single nucleus RNA-sequencing data reveals intra-tumoral heterogeneity in medulloblastoma brain tumors with extensive nodularity

Konstantin Okonechnikov (German Cancer Research Center)

Medulloblastoma with extensive nodularity (MBEN) represents a rare entity of cerebellar tumors occurring in early childhood, but with a favorable prognosis. Even though this tumor type could be diagnosed based on the characteristic histological pattern of nodular reticulin-free and inter-nodular reticulin-rich components, various molecular biology techniques including methylation and transcriptome profiling allowed to improve MBEN characterization among other central nervous system tumor classes. Nevertheless, the biological mechanisms leading to the favorable clinical course of MBEN still remain poorly understood.

Saturday, July 31st, 13:00

Variation of mutation rate between individual cells

Maria A Andrianova (Skolkovo Institute of Science and Technology)

It is well known that somatic mutation rate varies across the genome however, information about variation of mutation rate between individual cell divisions is limited. Yeast experiments suggest that overall number of mutations could fluctuate between subsequent replications. Colonies grown from cells exposed to the pulse of the mutagen can be informative about damage to mutation conversion that happen predominantly during the first division after pulse. Here we used data from liver tumors induced with a single dose of diethylnitrosamine (DEN) in C3H/HeOuJ inbred mice to analyze and describe the variation in mutation rate between individual cell divisions for different genome regions.

Saturday, July 31st, 13:15

Joint model of RNA velocity on transcriptional manifold

Ruslan Soldatov (Harvard Medical School)

Single-cell RNA sequencing accurately quantifies RNA abundance. The approach captures a static ensemble of cell states at a point in time, but does not resolve dynamic relationship between cell states. RNA velocity – the time derivative of the transcriptional state – can be directly estimated using unspliced and spliced mRNA fragments in common single-cell RNA sequencing protocols. In the general case, the challenge of RNA velocity problem is that the time axes, or underlying physical vector field, as well as splicing and degradation rates of genes are unknown and objects of statistical inference. Here we present a general solution to velocity estimation that jointly models kinetics rates of genes and underlying physical vector field on transcriptional manifold. Simulations show that the joint RNA velocity method enables unbiased inference of linearized physical time and kinetics rates of genes under a wide variety of scenarios, successfully takes into account variable kinetics rates and provides a test for differential change in kinetics parameters. Application of joint RNA velocity method to single-cell RNA-seq data of a number of dynamic biological systems indicates its superior performance compared to other RNA velocity models and biological interpretability of inferred parameters.

Saturday, July 31st, 13:30

Identification and separation of sources of transcriptional variability in single-cell RNA-seq data

Konstantin Zaitsev (ITMO University)

Saturday, July 31st, 13:40

Origin of helper T cell diversity in human immunity: the study of TCR repertoires

Sofya Kasatskaya (Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry)

Saturday, July 31st, 13:50

Investigation of *T. thermophilus* Type III CRISPR-Cas mediated bacterial defense mechanisms

Daria Artamonova (Skolkovo Institute of Science and Technology)

Saturday, July 31st, 14:00

Direct reprogramming of human skin fibroblasts into insulin-producing cells using CRISPR/dCas9-SunTag-VP64 effector epigenome editing systems

Alexandr Artyuhov (Pirogov Russian National Research Medical University)

Saturday, July 31st, 14:10
Rational design of chimeric portal proteins of viruses
Evgeny Klimuk (Institute of Molecular Genetics RAS)

Saturday, July 31st, 16:00

Conserved long-range base pairings are associated with pre-mRNA processing of human genes

Dmitri Pervouchine (Skolkovo Institute of Science and Technology)

The ability of nucleic acids to form double-stranded structures is essential for all living systems on Earth. While DNA employs it for genome replication, RNA molecules fold into complicated secondary and tertiary structures. Current knowledge on functional RNA structures in human protein-coding genes is focused on locally-occurring base pairs. However, chemical crosslinking and proximity ligation experiments have demonstrated that long-range RNA structures are highly abundant.

Saturday, July 31st, 16:15

An extended catalogue of tandem alternative splice sites in human tissue transcriptomes

Aleksei Mironov (Skolkovo Institute of Science and Technology)

Tandem alternative splice sites (TASS) is a special class of alternative splicing events that are characterized by a close tandem arrangement of splice sites. Most TASS lack functional characterization and are believed to arise from splicing noise. Based on the RNA-seq data from the Genotype Tissue Expression project, we present an extended catalogue of TASS in healthy human tissues and analyze their tissue-specific expression. The expression of TASS is usually dominated by one major splice site (maSS), while the expression of minor splice sites (miSS) is at least an order of magnitude lower. Among 46k miSS with sufficient read support, 9k (20%) are significantly expressed above the expected noise level, and among them 2.5k are expressed tissue-specifically. We found significant correlations between tissue-specific expression of RNA-binding proteins (RBP), tissue-specific expression of miSS, and miSS response to RBP inactivation by shRNA. In combination with RBP profiling by eCLIP, this allowed prediction of novel cases of tissue-specific splicing regulation including a miSS in QKI mRNA that is likely regulated by PTBP1. The analysis of human primary cell transcriptomes suggested that both tissue-specific and cell-type-specific factors contribute to the regulation of miSS expression. More than 20% of tissue-specific miSS affect structured protein regions and may adjust protein-protein interactions or modify the stability of the protein core. The significantly expressed miSS evolve under the same selection pressure as maSS, while other miSS lack signatures of evolutionary selection and conservation. Using mixture models, we estimated that not more than 15% of maSS and not more than 54% of tissue-specific miSS are noisy, while the proportion of noisy splice sites among non-significantly expressed miSS is above 63%.

Saturday, July 31st, 16:30

Genome-wide search for pathogenic splice variants manifested in different genetic backgrounds

Mikhail Y Skoblov (Research Centre for Medical Genetics)

Hereditary diseases could be caused by pathogenic nucleotide variants that affecting pre-mRNA splicing. We assumed that if a mutation appears near to a frequent SNP their co-operative impact on splicing regulatory sequences can be very significant. At the same time, the effect of mutation could be different depending on the allelic states of SNP. For instance, the appearance of a mutation can lead to the formation of a splicing site in one allelic state, but not in another. In this work, we attempted a genome-wide search for such “complex alleles”. Using the modified SpliceAI algorithm applied on both reference and “alternative” genome versions, we predicted those mutations which could change the splicing pattern in one version of the genome but not in another. As a result, we found about 41 thousand mutations. Using the minigene system, we carried out experimental validation of some cases which could be relevant to medical genetics.

Saturday, July 31st, 16:45

Assessing proteome diversity raised by alternative splicing in Brain RNA-Seq data

Vladimir N Babenko (Institute of Cytology and Genetics SB RAS)

As it is known, alternative splicing (AS) primarily is manifested in brain cells along with muscle ones to a lesser effect. Also, it is known that at least half of the AS events manifested by exon skipping/insertion/5'-3' splice sites alteration leads to premature stop codon emergence, and is subject to Nonsense mediated decay (NMD), serving as additional means of prompt expression abrogation upon the local compartment dynamics of protein turnover. It is widely employed in the homeostasis maintenance of multicomponent complex subunits, such as ribosome, spliceosome, and chromatin remodeling machineries, and is manifested particularly distinct in brain cells (Zhang et al., 2014). Still, the most interest represents the functional protein coding splicing outcome for assessing the specific protein diversity expansion in the sample. One of the approaches filtering out NMD subjected splicing products is to target ribosome –engaged transcripts, thus guaranteeing the coding competence of mRNAs employed (Furlanis et al, 2020). However, such approaches are cost-consuming, and hence not widely represented in RNA-Seq experiments.

Saturday, July 31st, 17:00

Proteome recoding by A-to-I mRNA editing

Sergei Moshkovskii (Pirogov Russian National Research Medical University)

Influence of A-to-I RNA editing by ADAR enzymes which leaded to non-synonymous mRNA substitutions was estimated proteome-wide using publicly shotgun mass spectrometry proteomic data for fruit fly, murine and human brains. ADAR-mediated RNA editing plays a primary role in immunity regulation, and adaptive significance of protein recoding due to this type of editing is still being discussed. Using a modified database for containing recoded protein sites predicted from RNA-seq data, we have shown that recoded forms are generally depleted in proteomes in comparison to genomically encoded sequences. Thus, of thousands non-synonymous edited sites in fruit fly, mouse and human transcriptomes, we identified 1-3% sites at the proteome level in each brain proteome. In the fruit fly brain, recoded sites were shown to be enriched in proteins of SNARE presynaptic complex and other vesicle trafficking components. Selected findings were confirmed by targeted mass-spectrometry which also showed a dynamics in recoding of some protein sites during insect ontogeny. Re-analysis of deep proteomes of murine and human brains could identify as few as twenty and thirty seven recoded sites in mouse and human, respectively. Of them, eight sites in six proteins were conservative between species. Adaptive significance of these sites, as well as their possible role in animal and human pathology, should be further elucidated by functional studies.

Saturday, July 31st, 17:15

Fast gene set enrichment analysis with multi-level Monte-Carlo approach

Alexey A Sergushichev (ITMO University)

Gene set enrichment analysis (GSEA) is ubiquitously used tool for evaluating pathway enrichments in transcriptional data. However, the reference implementation of this method has a low P-value resolution, meaning that practically nominal P-values can be estimated only up to 1e-4 which precludes multiple hypothesis correction when thousands of pathways are considered.

Here we present FGSEA (Fast Gene Set Enrichment Analysis) - method for calculating GSEA P-values based on adaptive multi-level split Monte Carlo approach. FGSEA allows to estimate arbitrarily low GSEA P-values with a high accuracy in a matter of minutes or even seconds. To confirm the accuracy of the method, we also developed an exact algorithm for GSEA P-values calculation for integer gene-level statistics. Using the exact algorithm as a reference we show that FGSEA is able to routinely estimate P-values up to 1e-100 with a small and predictable estimation error.

To highlight the importance of accurate P-value calculation we perform a systematic analysis of 600 datasets from Gene Expression Omnibus. We show that for Benjamini-Hochberg procedure, about half of the datasets require resolution higher than 1e-4 to detect the presence of statistically significant hits and, ultimately, a resolution of 1e-7 can be required for large pathway collections. We also show that the ad hoc correction procedure used in the reference GSEA implementation is too conservative compared to Benjamini-Hochberg and fails to detect a huge number of statistically significant hits.

Thus, FGSEA overcomes two major problems of the pre-ranked GSEA implementations: it is fast and due to high accuracy it is compatible with standard multiple hypothesis correction procedures such as Benjamini-Hochberg procedure. Consequently, FGSEA is more sensitive and allows to detect statistically significant pathways when the reference implementation fails.

Saturday, July 31st, 18:00

Tempora: Cell trajectory inference using time-series single-cell RNA sequencing data

Gary Bader (University of Toronto)

Saturday, July 31st, 18:30
Cell segmentation in spatial transcriptomics
Peter Kharchenko (Harvard Medical School)

Saturday, July 31st, 19:00

Multimodal by default: designing the next generation of methods and tools for single-cell multi-omics

Danila Bredikhin (EMBL)

Multimodal single-cell omics is a rapidly developing field of research, and biological and biomedical single-cell multi-omics datasets being generated present a computational challenge, both due to the sheer volume of the data and to the added complexity of having multiple layers of information per cell. Here we lay the groundwork for the scalable multi-omics data storage and exchange formats as well as for multimodal integration methods. We discuss the distinctive features of such infrastructure, and we provide an implementation of these concepts in a multilanguage ecosystem for multi-omics analysis demonstrating its utility for analysing diverse multimodal datasets including the ones that combine RNA sequencing with chromatin accessibility profiling or epitope profiling as well as with spatiotemporal information. We also address the challenge of interactive data exploration by describing an omics data serialization approach and by implementing a set of software components within the scope of modern web technologies allowing to easily exchange and interact with omics data and multimodal single-cell datasets in particular.

Sunday, August 1st, 10:00

Biological design and prediction using evolutionary information

Deborah Marks (Harvard Medical School)

Sunday, August 1st, 10:30

The missing link between genetic association and regulatory function

Shamil Sunyaev (Harvard Medical School)

Sunday, August 1st, 11:00

Peeking inside the clockworks of regulatory single-nucleotide variants with ADASTRA and ANANASTRA

Ivan V Kulakovskiy (Institute of Protein Research)

We present a comprehensive meta-analysis of more than 15 thousand ChIP-Seq experiments that allowed us to identify nearly a million allele-specific binding events of hundreds of human transcription factors. The respective single-nucleotide variants are often associated with various hereditary traits and serve as eQTLs. The resulting ADASTRA database and ANANASTRA webserver are accessible at https://adastra.autosome.ru and https://ananastra.autosome.ru.

Sunday, August 1st, 11:15

Within-gene epistatic selection in genetically diverse populations

Anastasia V Stolyarova (Skolkovo Institute of Science and Technology)

As proposed by Fisher and Dobzhansky, polymorphism within populations can be shaped by coadaptation between chromosomal regions. However, since recombination disrupts the coadapted gene complexes, coadaptation between distant loci requires some form of recombination suppression. Within a single gene, physical proximity alone may suffice to limit recombination, so sets of coadapted variants may evolve. At population level, this should result in increased linkage disequilibrium (LD) between interacting positions. Most species, however, lack variation to make this increase detectable. The basidiomycete fungus Schizophyllum commune possesses the highest genetic diversity among the studied eukaryotic species, with ~20% of neutral sites differing between any two individuals.

Here, we study the LD patterns in 54 complete genomes of S. commune from North America and Europe. In both populations, the LD between nonsynonymous mutations is higher than that between synonymous ones at the same nucleotide distance; furthermore, the LD between nonsynonymous mutations within a gene is higher than between those in different genes at the same nucleotide distance. Simulations show that the elevated nonsynonymous LD cannot result from differences in negative selection, selective sweeps or background selection. Instead, it implies abundant epistasis between nonsynonymous sites. The LD is increased between sites for which interactions are more likely a priori, e.g. those encoding amino acids adjacent in the protein structure. As expected, the elevated intragenic LD between nonsynonymous mutations is undetectable in humans and barely detectable in fruit flies. Our results are the first observation of a genome-wide pattern in within gene polymorphism caused by epistasis.

Sunday, August 1st, 12:00

Estimating the timing of multiple admixture events using 3-locus Linkage Disequilibrium

Mikhail Shishkin (HSE University)

Estimating admixture histories is crucial for understanding the genetic diversity observed in present-day populations. Existing allele frequency or phylogeny-based methods are excellent for inferring the existence of admixture or its proportions, but have less power for estimating admixture times. Approaches for estimating these times use spatial information from admixed chromosomes, such as the local ancestry or the decay of admixture linkage disequilibrium (ALD). One popular method, implemented in the programs ALDER and ROLLOFF, uses two-locus ALD to infer the time of a single admixture event, but is only able to estimate the time of the most recent admixture event based on this summary statistic. We derive analytical expressions for the expected ALD in a three-locus case taking into account migration, genetic drift and recombinations as linear operators acting on covariance function. We provide a new statistical method based on these results that is able to resolve more complicated admixture histories. The fast Fourier transform is used to speed up the calculations of ALD from real data. Using simulations, we show how this new statistic behaves on a range of admixture histories. As an example, we also apply our method to real data. The method is implemented in Python and the code is available on GitHub https://github.com/Genomics-HSE/ThLd

Sunday, August 1st, 12:15

Quantitative genetics of human protein N-glycosylation

Yurii Aulchenko (Institute of Cytology and Genetics SD RAS)

Addition of carbohydrates (glycosylation) is a common co- and post-translational modification of proteins. Different glycosylation of the same protein changes its physical properties as well as its biological function. Although changes in protein glycosylation are observed in a wide range of diseases and pathological states, the examples of use of glycans as biomarkers and therapeutic targets is limited. This is not in small part because the understanding of human glycome regulation in vivo is incomplete. Combination of human glycomics and genomics offers a powerful “data-driven hypotheses” approach to dissect the complex human glycobiology in vivo in an agnostic manner.

In this talk, I will summarise our efforts to discover networks driving human protein glycosylation using methods of quantitative genetics and computational functional genomics. Together with our collaborators, we implemented large multi-center genome-wide association studies (GWAS) that discovered a large number of new loci robustly associated with variation in N-glycome. While many loci contain genes encoding enzymes directly involved in glycosylation, we show that with increased sample sizes GWAS of human N-glycome start revealing regulators of the biochemical network of N-glycosylation. Some of these regulators demonstrate pleiotropic effects on human disease, especially autoimmune and inflammatory. We emphasize the use of in silico functional methods and multi-omics approaches to prioritise functional mechanisms to be further validated in laboratory experiments. This combined approach will lead to better understanding of mechanisms of regulation of human protein glycosylation and will hopefully provide a source of etiologic insight, therapeutic interventions and biomarkers.

This work was supported by a grant from the Russian Science Foundation (RSF) No. 19-15-00115.

Sunday, August 1st, 12:30

Mapping genes involved in control of N-glycosylation of blood glycoproteins through a large genome-wide association study

Sodbo Sharapov (Institute of Cytology and Genetics SB RAS)

Glycosylation is a common and diverse post-translational modification of proteins that influences their physical properties and biological functions. Although changes in protein glycosylation are observed in many diseases (1–3), the examples of the use of glycans as biomarkers and therapeutic targets are limited (4). This is in small part because the understanding of human glycome regulation in vivo is incomplete and fragmented.

To bridge this gap, we performed the largest genome-wide association study of the human blood plasma protein N-glycosylation measured by ultra-high performance liquid chromatography (5) in 10,765 people. We studied the association between 8.8 million genetic polymorphisms on human autosomes and 117 relative abundances of N-glycan structures. We discovered and replicated 31 associated loci, 16 of which are novel. The SBayesR prediction models (6) that included on average 1,090,196 genetic polymorphisms explained up to 21% of glycan variance (36% of SNP-based heritability).

To prioritize potentially causal genes in the established loci, we performed an in silico functional study. Eight loci contained genes coding enzymes with a known role in N-glycan biosynthesis, while 23 loci, including 16 novels, may contain regulators of protein glycosylation, including transcription factors, transporters, blood pQTLs for glycosylated proteins. Using a network-based approach, we explored a functional network formed by the glycome-associated loci. Our results set the scene, provide data and hypotheses for future studies that will establish genes and gene networks involved in the regulation of global, cell-, tissue-, and protein-specific pathways of protein glycosylation.

Acknowledgments: This work was supported by a grant from the Russian Science Foundation (RSF) No. 19-15-00115.

Sunday, August 1st, 12:45

Targeted sequencing of 242 clinically important genes in a sample of Russian population from Ivanovo region

Vasily Ramensky (National Medical Research Center for Therapy and Preventive Medicine of the Ministry of Healthcare of the Russian Federation)

In spite of ongoing efforts to aggregate genome sequencing data, a significant part of population-specific variation remains uncharted. Russian population, being the largest among the European countries, is under-represented in large-scale sequencing projects. We performed targeted sequencing of 242 clinically important genes mostly associated with cardiovascular diseases in a representative population sample of 1,658 individuals from the Ivanovo region. To the best of our knowledge, this is the first analysis of incidental genetic findings in the Russian population. Our results also emphasize the importance of sequencing large (>1,000 individuals) cohorts to uncover the population-specific genetic variation of clinical relevance.

Sunday, August 1st, 13:00

Mutational spectrum of the human mitochondrial genome: somatic and germline signatures differ

Dmitry Iliushchenko (Immanuel Kant Baltic Federal University)

To reveal key mutagens, acting on the human mitochondrial genome (mtDNA) in somatic and germline tissues we reconstructed two mtDNA mutational spectra using collections of mtDNA substitutions from human cancers (7611 substitutions from PCAWG Consortium) and human phylogenetic tree (300’000 synonymous substitutions, derived from a tree based on 40000 complete human mtDNAs). Two reconstructed spectra were very similar in terms of strong excess of C>T and A>G transitions (mtDNA heavy chain notation) and rareness of all transversions. However, deep analyses of the nucleotide context revealed one remarkable difference in the pattern of the most common C>T transition: in somatic tissues it occurs mainly in the CpG context, while in germline tissues - in CpC. An excess of C>T in the CpG context of mtDNA in cancer tissues has been shown before and is very similar to the clock-like age-related cosmic mutational signature SBS1, suggesting the spontaneous deamination of 5-methylcytosine in mtDNA. C>T in the CpC context has recently been associated with oxidative damage of single-stranded DNA (ssDNA). These results suggests that (i) process of cytosine methylation, although still questionable for mtDNA, can have a place in somatic but not germline tissues; (ii) oxidative damage is stronger in mitochondria of oocytes versus mitochondria of somatic tissues, emphasising high enough aerobic metabolism of oocytes. Altogether we conclude that mtDNA mutations of somatic tissues demonstrate an “aging” signature (C>T in CpG context) while mtDNA mutations of germline show a signature of increased redox stress (C>T in CpC context).

Sunday, August 1st, 13:15
Precise mutation rate estimation on an individual site level
Vladimir Seplyarskiy (Harvard Medical School)

Sunday, August 1st, 13:30

Integration of genomics and metabolomics data for discovery of small natural products

Alexey Gurevich (Saint-Petersburg State University)

Sunday, August 1st, 13:40

Mechanistic model of mutation accumulation based on differential kinetics of DNA repair

Maria Andrianova (Skolkovo Institute of Science and Technology)

Sunday, August 1st, 13:50

Investigation of the genetic architecture of human traits on the phenome scale

Yury Barbitoff (Saint-Petersburg State University)

Sunday, August 1st, 14:00

Transcriptional interference by RNA-guided DNA-binding proteins

Aleksei Agapov (Institute of Molecular Genetics)

Sunday, August 1st, 14:10

High-throughput analysis of short DNA fragments generated in vivo in Escherichia coli and Saccharomyces cerevisiae

Anna Shiriaeva (Skolkovo Institute of Science and Technology/ St. Petersburg Branch of Vavilov Institute of General Genetics)

Sunday, August 1st, 16:00

Intragenic compensatory variants: does the distance matter?

Nadezhda Azbukina (Lomonosov Moscow State University)

Many tools for computational prediction of mutation's phenotypic effect rely on the presence of analogous substitutions in homologous sequences in other species: if the mutant amino acid is observed in a related protein, it is likely to be benign. However, up to 10% of variants present as wild-type in other species are pathogenic in humans. One possible explanation of that phenomenon is compensation, a particular case of positive epistasis. In this study, we focus on intra-genic compensations and leverage the largest to date dataset of single and double mutations in over 1,000 proteins with described phenotypes to investigate the mechanistic details of compensations. We set out to explore two questions: are compensations local, i.e. do they happen at protein sites spatially adjacent to the mutation site; and how do they alter the thermodynamic stability of the proteins. Using our own and third-party methods of structural bioinformatics, we demonstrate that compensations are non-local in the protein core and that they compensate for destabilization of the protein structure caused by deleterious amino acid changes. In other protein regions, in particular on protein-protein interaction interfaces, compensations are local. We detect a set of universal compensators -- mutations capable of compensating for multiple deleterious changes, -- and show that they act at even larger distances and, when introduced as single mutations, lead to a more stable structure than the wild type.

Sunday, August 1st, 16:15

New algorithms for finding combinatorially complete datasets in high-throughput mutagenesis experimental data

Dmitry N Ivankov (Skolkovo Institute of Science and Technology)

Epistasis, the dependence of the mutational effect on the genetic background, is the principal obstacle for prediction phenotype and fitness from genotype. This is the reason to explore epistasis using currently available experimental data. The most straightforward approach to identify epistasis is to analyze the so-called combinatorially complete datasets, consisting in the simplest case of a reference genotype, two different single mutants and a double mutant with both of the single mutations. Higher-order epistasis is the result of generalization of this approach for mutations in three or more positions.

A combinatorially complete dataset built from K single mutations at K positions contains 2^K genotypes and represents K-dimensional hypercube in the genetic space. Using random mutagenesis, researchers measured phenotypes of thousands-to-millions genotypes, which provides invaluable source of the information about epistasis. However, identification of all hypercubes from such data is a non-trivial task since the measured genotypes populate genetic space randomly.

Previously, we presented the first algorithm of that kind called “HypercubeME”. The central concept in the algorithm is diagonal of a hypercube: if two hypercubes have the same diagonal, they are parallel to each other; if, in addition, they are located at the distance of one mutation, they form next-dimension hypercube. The algorithm has complexity O(N^2×K), where N is the number of genotypes in the dataset and K is the number of mutated positions. Now, we present three more algorithms, which use the same concept of hypercube diagonal but are much more efficient in hypercube identification. Two of them have complexity O(N×L×A) while the third one has complexity O(N×L×[logL+A]), where L is the length of the sequence and A is the size of the alphabet.

Sunday, August 1st, 16:30

HLA binding preferences of self-peptides are biased towards proteins with specific molecular functions

Vadim Karnaukhov (Skolkovo Institute of Science and Technology)

Human leukocyte antigen (HLA) is highly polymorphic and plays a key role in guiding adaptive immune responses by presenting foreign and self peptides to T cells. Each HLA variant selects a minor fraction of peptides that match a certain motif required for optimal interaction with the peptide-binding groove. These restriction rules define the landscape of peptides presented to T cells. Given these limitations, one might suggest that the choice of peptides presented by HLA is non-random and there is preferential presentation of an array of peptides that is optimal for distinguishing self and foreign proteins. In this study we explore these preferences with a comparative analysis of self peptides enriched and depleted in HLA ligands. We show that HLAs exhibit preferences towards presenting peptides from certain proteins while disfavoring others with specific functions, and highlight differences between various HLA genes and alleles in those preferences. We link those differences to HLA anchor residue propensities and amino acid composition of preferentially presented proteins. The set of proteins that peptides presented by a given HLA are most likely to be derived from can be used to distinguish between class I and class II HLAs and HLA alleles. Our observations can be extrapolated to explain the protective effect of certain HLA alleles in infectious diseases, and we hypothesize that they can also explain susceptibility to certain autoimmune diseases and cancers. We demonstrate that these differences lead to differential presentation of HIV, influenza virus, SARS-CoV-1 and SARS-CoV2 proteins by various HLA alleles. Finally, we show that the reported self peptidome preferences of distinct HLA variants can be compensated by combinations of HLA-A/HLA-B and HLA-A/HLAC alleles in frequent haplotypes.

Sunday, August 1st, 16:45

Co-aggregation of amyloids: from structural principles to in silico prediction

Andrey V. Kajava (CNRS)

Typically, naturally occurring amyloid fibrils consist of multiple copies of the same protein. In these amyloid structures, each polypeptide chain is folded into the same β-arc-containing block and they are stacked in a parallel and in-register manner. In the last few years, however, a considerable body of data has been accumulated about co-aggregation of different amyloid-forming proteins. Among the most studied examples are co-aggregation of PrPSc prion from human and animals, different yeast prions, human proteins Rip1 and Rip3 and heteroaggregates of bacterial CsgA and CsgB proteins. Since the co-aggregation is linked to such important phenomena as infectivity of amyloids and molecular mechanisms of functional amyloids, we decided to analyze its structural aspects in more details.

Sunday, August 1st, 17:00

Functional domain annotation of protein sequences with deep metric learning

Nikolay Russkikh (SRC VB "Vector")

Functional domain annotation of amino acid sequences is a very important and long-standing task. Identification of functional domains within the protein can provide us insights not only into its functions but also into its origin and evolution. Current methods of choice are based on homology detection. However, this task still remains challenging in certain cases, especially when dealing with sequences with very little or no homology towards annotated proteins. We propose that the issues of remote homology detection can be addressed with deep language modeling and deep metric learning approach. Our model was built on top of ProtBert transformer-based language model and was additionally finetuned with ArcFace loss on a training set of proteins extracted from Pfam 33.0 database. The resulting deep metric learning model maps each aminoacid within the protein sequences to a spherical embedding space, such that amino acids belonging to the same functional domains are close to each other. To classify the amino acid embeddings, derived from query proteins, to functional domains, we used a weighted kNN algorithm variation named NED (Normalized sum of Exponential of the Distances) which, in contrast to conventional kNN, takes into account the actual distances between the items of interest and their neighbors. The testing results obtained for token-wise functional domain prediction on the proteins from hold out dataset demonstrated very good performance. The token-wise accuracy was 0.9637, macro averaged recall was 0.8663 and macro averaged precision was equal to 0.8480. When testing the model we have also observed a number of examples where it was able to correctly identify domains absent from Pfam annotation of query sequences. Our model was implemented as a web-service and it could be freely accessed at http://217.79.62.70:3000/

Sunday, August 1st, 17:15

MolDiscovery: Learning Mass Spectrometry Fragmentation of Small Molecules

Azat Tagirdzhanov (St. Petersburg State University)

Identification of small molecules is a critical task in various areas of life science. Recent advances in mass spectrometry have enabled the collection of tandem mass spectra of small molecules from hundreds of thousands of environments. To identify which molecules are present in a sample, one can search mass spectra collected from the sample against millions of molecular structures in small-molecule databases. This is a complicated task as currently, it is not clear how small molecules are fragmented in mass spectrometry. The existing approaches use chemistry domain knowledge or quantum simulation to predict the fragmentation of molecules. However, these rule-based methods fail to explain many of the peaks in mass spectra of small molecules. Recently, spectral libraries with tens of thousands of annotated mass spectra of small molecules have emerged, paving the path for learning more accurate fragmentation models for mass spectral database search. We present molDiscovery, a mass spectral database search method that improves both efficiency and accuracy of small molecule identification by (i) utilizing an efficient algorithm to generate mass spectrometry fragmentations, and (ii) learning a probabilistic model to match small molecules with their mass spectra. We show our database search is an order of magnitude more efficient than the state-of-the-art methods, which enables searching against databases with millions of molecules.

Sunday, August 1st, 18:00

Models for crop species diversification within and between centers of domestication

Anna Igolkina (Peter the Great St.Petersburg Polytechnic University)

The geographical spread of crops from centers of domestication is clouded in mysteries, stemming from incomplete linguistics and archeology. The history of species domestication goes hand in hand with human history and is blueprinted in the genome of crops. Chickpea (Cicer arietinum) is a crop that according to archeological records was first domesticated in the Fertile Crescent 10 thousand years ago, but its following diversification in antiquity in South Asia, Ethiopia, and the Western Mediterranean is still obscure and controversial if based on only archeological and historical evidence. Chickpea has two market classes, one, the ‘desi’ cultivated type, which has similar in flower and seed coat color to chickpea’s wild relatives, and a derived light seeded ‘kabuli’ type linguistically tied to Central Asia but with unknown geographic origin.

Based on the genetic data of 421 chickpea landraces from six geographic regions, we tested complex historical hypotheses of chickpea migrations and admixtures on two layers: within and between regions. For the former layer, we developed popdisp, the Bayesian model of population dispersals from the region center towards sample locations of landraces, and confirmed that chickpea spreads within each region along trade routes rather than by simple diffusion.

To resolve uncertainty on the ways by which chickpea reached different regions, we have developed another model, migadmi, that evaluates multiple and nested admixture events. Applying this model to desi populations, we found both the Indian and Middle Eastern traces in Ethiopian chickpea, which are in line with the presence of the seaway from South Asia to Ethiopia and the cultural legacy of the Queen of Sheba. Comparing two contrast hypotheses of kabuli’s first domestication site, we found significant evidence of a Turkish origin rather than a Central Asia one.

The research was supported by RFBR grant 18-29-13033

Sunday, August 1st, 18:15

Genomic blueprints of flax diversification and improvement

Maria Duk (Peter the Great St.Petersburg Polytechnic University)

Flax, an economically crop was first domesticated in Fertile Crescent about 12000 years ago and after that spread into Indian subcontinent, Central Asia, Mediteranean and East Africa, thus forming centers of secondary diversity. The demographic history of flax diversification has many bank pages. Here we applied population genetic methods to a dataset of 306 flax accessions from collection of the Federal Research Center for Bast Fiber Crops to unravel fiber flax origins and breeding history.

Sunday, August 1st, 18:30

Numerical study of the flowering time model of wild chickpeas

Andrey Y Ageev (Peter the Great St.Petersburg Polytechnic University)

Accurate prediction of flowering time helps breeders to develop new varieties that can achieve maximal efficiency in a changing climate. We have developed a methodology for construction a simulation model for flowering time in which a function for daily progression of the plant from one to the next phenological phase is obtained in analytic form by stochastic minimization. For demonstration, we build a simulation model for time to flowering in recently assembled dataset of wild chickpeas. The resulting model demonstrated high accuracy on the dataset. The inclusion of genotype-by-climatic factors interactions accounted to 77% of accuracy in terms of root mean square error. MarkSim weather generator was used to produce 30 repetitions of daily weather forecasts for Ankara (Turkey) from 2020 to 2080 and socioeconomic scenarios rcp26, rcp45, rcp60 and rcp85 described by Representative Concentration Pathways (RCPs) of carbon dioxide. Time intervals of 20 years from 2020 to 2080 were considered, and the changes in average time to flowering were found for more than 50% of different combinations of scenarios and collection sites.

Sunday, August 1st, 18:45

Artificial intelligence applied to population genomics and paleogenomics

Olga Dolgova (Center for Genomic Regulation)

Detecting different ways of archaic introgression into Hominin genome is an extremely active field within the community of human population genomics. Nevertheless, inference of admixture and differentiation between ancient populations represent a serious treat when conducting methods based on model simulations, properly identifying its signature is essential for interpreting the different evolutionary processes (both demographic and selective) that shaped the genome of the species. Several algorithms have been proposed for unraveling these evolutionary events. However, it has been suggested that the proposed methods show a list of limitations, both in biological and technical terms. First of all, the algorithms do not model the relationship between the ancestral populations. As a consequence, several demographic scenarios can produce the same output and interpretation of the outcome is complex. This situation is even more complex when considering both ancient and modern samples at the same time. The algorithms do not correct for the temporal difference among the samples, thus producing a systematic bias on the estimated proportions of ancestry in the ancient sample. Besides, so far the proposed algorithms do not consider polymorphisms at low frequency, despite these variants represent a considerable proportion of the genetic variation of the species and they can be more informative for detecting population substructure. Furthermore, including these markers can produce artifacts in the results from the different algorithms.

Our group has developed a novel model-based approach based on coupling of Deep Learning with Approximate Bayesian Computation for alleviation of these reported problems and performing multiple hypotheses tests. We recently applied it to detection of demographic processes along evolutionary history of particular ancient and modern human populations as well as of great ape species, detecting multiple ancient introgression and admixture events.

Sunday, August 1st, 19:00

A systems biology approach to understanding SARS-CoV-2 transmissibility in population

Marko Djordjevic, Sofija Markovic (University of Belgrade)

We studied the influence of intervention measures and environmental factors on COVID-19 transmission. The research combines computational systems biology [1,2], bioinformatics [2], and scaling relation analysis [3]. Our results show that several demographic and meteorological factors significantly affect the inherent transmission (basic reproductive number R0) of the virus in the population. We also analyzed the disproportion between the spread of the infection in Wuhan (Hubei) and the much smaller numbers in other Chinese provinces. We showed that this puzzle can be explained by a combination of significantly higher inherent virus transmission in Wuhan (influenced by environmental factors) and higher effectiveness of epidemic control measures in other provinces. Overall, our results indicate that the dynamics of epidemic spread may significantly depend on potentially highly heterogeneous and seemingly random factors, such as variations in demographic and meteorological conditions, as well as their complex interaction with introduced control measures. Understanding these factors is crucial, not only for risk analysis during a pandemic but also for long-term prediction of virus behavior in a population if the disease becomes endemic.

Sunday, August 1st, 19:15

A single chromosome driving rapid extremotolerant adaptation in an insect

Yuki Yoshida (University of Tokyo)

Midges (Chironomidae) are known to inhabit a wide range of environments, and certain species can tolerate extreme conditions. In particular, the sleeping chironomid Polypedilum vanderplanki is known for the remarkable ability of its larvae to withstand almost complete desiccation by entering a state called anhydrobiosis. Chromosome numbers in chironomids are higher than in other dipterans and this extra genomic resource might facilitate rapid adaptation to novel environments. A chromosomal-level genome assembly would allow this hypothesis to be tested, but previous genome sequencing projects have resulted in fragmented assemblies. Here, we used improved sequencing strategies to assemble a chromosome-level genome sequence for P. vanderplanki. We provide evidence for the specialization of chromosome 4 through extensive acquirement of novel genes. A high degree of subfunctionalization in paralogous anhydrobiosis-related genes occurs in this chromosome, as well as pseudogenization in a highly duplicated gene family. These findings suggest that the fourth chromosome in chironomids is a site of high genetic turnover, allowing it to act as a ‘sandpit’ for evolutionary experiments, thus facilitating the rapid adaptation of midges to harsh environments.

Monday, August 2nd, 10:00

Quantifying and Manipulating the Aging Process

Vadim N Gladyshev (Harvard Medical School)

DNA methylation of a defined set of CpG dinucleotides emerged as a critical and precise biomarker of the aging process. Multi-variate machine learning models, known as epigenetic clocks, can exploit quantitative changes in the methylome to predict the age of bulk tissue with remarkable accuracy. We developed various types of aging clocks and applied them broadly to assess the effect of interventions that extend lifespan as well as the effect of treatments that support rejuvenation. Several recent studies will be presented at the meeting to show how these clocks achieve broad insights into the aging process.

Monday, August 2nd, 10:30
Identification of phenotype-specific networks from paired gene expression-cell shape imaging data
Evangelia Petsalaki (EMBL)

Monday, August 2nd, 11:00

Which wiring of the cell cycle opens the path to mitotic and meiotic division?

Ovidiu Radulescu (University of Monpellier)

Using mathematical modeling and experiments on frog egg extracts we challenge a traditional picture of cell cycle dynamics during mitosis and meiosis. According to this picture, cyclin B can activate the maturation promoting factor (MPF) by pushing its inactive state over a tipping point resulting from positive feedback loops. We show that both mitotic and meiotic entries need two activators, one playing the role of a trigger, and the second being the MPF, i.e. the cyclin B – cdk1 complex that plays only the role of a driver. Using experimentation and a machine learning pipeline we identify the triggers of mitosis and meiosis and validate their property to lower the barrier between the active and inactive states of cyclin B – cdk1 complex. This dual regulation scheme leads to robustness of the cell cycle progression with respect to fluctuations. Our results improve the realism of models of circuitry and spatio-temporal dynamics of the eukaryote cell cycle. This is particularly important for medicine where components of cell cycle machinery are targets for cancer therapy.

Monday, August 2nd, 11:15

Learning the cell states using deep neural networks for cell type-specific interpretation of disease variants

Veniamin Fishman (Sberbank AI Lab)

Scoring the effects of individual genetic variations is one of the most challenging problems of modern genetics. Several machine-learning-based approaches were recently developed allowing the prediction of gene expression changes caused by single nucleotide variations. Although valuable, these methods require a comprehensive dataset of epigenetic information to train the models, which is not available for many cell types.

Here we introduce the concept of cell state learning, which allows the prediction of gene expression changes caused by sequence variations. Biologically, we interpret cell states as summarized representations of all trans-factors and other regulatory units active in specific cell types. Mathematically, we define the cell state as a latent representation of cell type, or cell type embedding, which can be used to transform DNA sequence into the cell type-specific epigenetic profile.

We show that cell state embeddings can be learned simultaneously with solving the problem of cell type-specific prediction of experimentally measured epigenetic profiles from DNA sequences. This is achieved by training a deep neural network to build two embeddings: the cell state embedding (a vector invariant across genomic loci and epigenetic marks) and DNA embedding (a vector invariant across cell types). Importantly, since the cell state embedding is invariant across epigenetic marks, it could be learned using incomplete datasets with missing epigenetic data.

We demonstrate that clustering of the inferred cell state vectors reflects biologically meaningful relations between cell types. Moreover, using DNA sequence and cell state inputs we achieved accuracy for DNAseI accessible regions prediction comparable with state-of-the-art methods. These results can be extended in the future to predict cell type-specific alterations of gene expression.

Monday, August 2nd, 12:00
Computational methods for genome interpretation
Emidio Capriotti (University of Bologna)

Monday, August 2nd, 12:30

Laying foundations for very large scale comparative genomics

Christophe Dessimoz (Swiss Institute for Bioinformatics)

Monday, August 2nd, 13:00
Bringing Human Population Genetics to Protein Structure
Geoff Barton (University of Dundee)

Monday, August 2nd, 13:40

The investigation of Mitochondrial Genomes of Helianthus species

Maksim Makarenko (Institute for Information Transmission Problems)

Monday, August 2nd, 13:50

High-Throughput Splicing Efficiency Analysis System

Sofia Mariasina (Moscow State University/Skolkovo Institute of Science and Technology)

Monday, August 2nd, 14:00

Computational methods for unsupervised demographic inference of multiple populations from genomic data

Ekaterina Noskova (ITMO University)

Monday, August 2nd, 14:10

The function and structure of pseudo-nucleus in jumbo phages

Alexei Samolygo (Skolkovo Institute of Science and Technology)

Monday, August 2nd, 14:20

Systematic identification and validation of novel RiPP biosynthetic gene clusters encoding YcaO-domain enzymes

Dmitrii Travin (Skolkovo Institute of Science and Technology)

Monday, August 2nd, 16:00
Pluripotency factors in zebrafish embryo
Daria Onichtchouk (Albert-Ludwigs-University of Freiburg)

Monday, August 2nd, 16:30
Optimal decoding of cellular identities in a genetic network
Gasper Tkacik (Institute of Science and Technology Austria)

Monday, August 2nd, 17:00

Early developmental asymmetries in cell lineage trees in living individuals

Alexej Abyzov (Mayo Clinic)

Post-zygotic mosaic mutations can be used to track cell lineages in humans. By using cell cloning and induced pluripotent cell lines, we analyzed early cell lineages in two living individuals (a patient and a control), and a postmortem human specimen. Of ten reconstructed post-zygotic divisions, none resulted in balanced contributions of daughter lineages to tissues. In both living individuals one of two lineages from the first cleavage was dominant across tissues, with 90% frequency in blood. We propose that the efficiency of DNA repair contributes to lineage imbalance. Allocation of lineages in postmortem brain correlated with anterior-posterior axis, associating lineage history with cell fate choices in embryos. Recurrence of germline variants as mosaic suggested that certain loci may be particularly susceptible to mutagenesis. We establish a minimally invasive framework for defining cell lineages in any living individual, which paves the way for studying their relevance in health and disease.

Monday, August 2nd, 17:15

The digenean complex life cycle: phylostratigraphy analysis of molecular signatures

Maksim A Nesterenko (Saint Petersburg University)

A complex life cycle with distinct stages occurs in various clades of the animal kingdom. One of the most remarkable examples of a complex life cycle belongs to the parasitic flatworms from the Digenea group. Given all life cycle stages use one common genome, the questions about the evolutionary origin and transformation of stage phenotypes are raised. We applied the phylostratigraphy approach to transcriptomes of digenean species in order to understand how the multiple phenotypes of contrast stages were evolved on the single genome. According to phylostatigraphy content analysis results, the formation of stage-specific molecular signatures is based on the changes in the activity of genes that have arisen at different phylogenetic levels from unicellular organisms to the species. The analysis revealed that the “youngest” signatures have Fasciola gigantica metacercaria, 21-days-old juvenile of F. hepatica, adult worm of Psilotrema simillimum, and schistosomules of Schistosomatidae representatives. We can suggest the “youth” of stage transcriptome may reflect the response of the stage to the environment by activation of groups of genes with relatively new phylogenetic origin. Given the stage-specific molecular signatures plasticity, it is difficult to define the sequence of evolutionary events leading to digenean life cycle complexity increase.

Monday, August 2nd, 18:00

Chromosome conformation of the hyperthermophilic archaeon Thermofilum adornatus

Andrey Sobolev (Institute of Gene Biology RAS)

Three-dimensional structure of chromosomes displays diverse patterns across the tree of life, with various levels of organization being quite universally observed. The Archaea remains understudied to this extent so far, despite being an intriguing area from the evolutionary perspective. Pioneering studies of chromatin structure of the archaeal kingdom based on high-throughput sequencing revealed considerable variability across the species and phyla. Particularly, the chromosome architecture of Crenarchaeota, one of the hyperthermophilic phyla of Archaea, manifests diverse patterns across different orders.

Here, we describe the spatial chromosomal organization of a hyperthermophilic crenarchaeon Thermofilum adornatus strain 1910b based on 3C-seq approach. The chromosome contact map showed a curved secondary diagonal almost orthogonal to the main one. No evidence of chromosome loops was present. We were able to identify boundaries of different strengths between chromosome interaction domains (CIDs) albeit moderate. The plaid-like patterns previously reported for Sulfolobus archaea were not observed. However, the calculation of A/B compartments divided the genome into 2 domains that were different by the density of predicted highly expressed genes and location of origins.

Further comparison of these domains with whole-genome gene expression profiles will allow to test whether these domains represent expression-associated compartments. If so, it is possible that they represent primitive compartments evolutionarily older than the plaid patterns of Sulfolobus and higher eukaryotes. Further exploration of 3D chromatin in all branches of archaeal diversity will elucidate the evolution of the links between structural and functional organization in live organisms.

This work was supported by the Russian Science Foundation [19-74-10092].

Monday, August 2nd, 18:15

Method for detection of parallel adaptation with genome rearrangements in bacterial populations

Nikita Alexeev (ITMO University)

Motivation:

High plasticity of bacterial genomes is provided by numerous mechanisms including horizontal gene transfer and recombination via numerous flanking repeats. Genome rearrangements such as inversions, deletions, insertions, and duplications may independently occur in different strains, providing parallel adaptation. Specifically such rearrangements might be responsible for multi-virulence, antibiotic resistance, and antigenic variation. However, identification of such events requires laborious manual inspection and verification of phyletic pattern consistency.

Results:

Here we define the term ``parallel rearrangements'' as the events that occur independently in phylogenetically distant bacterial strains and present a formalization of the problem of parallel rearrangements calling. We implement an algorithmic solution for identification of parallel rearrangements in bacterial population as a tool PaReBrick. The tool takes synteny blocks and a phylogenetic tree as input and outputs rearrangement events. The tool tests each rearrangement for consistency with a tree and sorts the events by their parallelism score and provides diagrams of the neighbors for each block of interest allowing the detection of horizontally transferred blocks or its extra copies and the inversions in which copied blocks are involved. We proved PaReBrick's efficiency and accuracy and showed its potential to detect genome rearrangements responsible for pathogenicity and adaptation in bacterial genomes.

Monday, August 2nd, 18:30

Evolutionary benefits of bacterial genomes with chromids

Olga Bochkareva (IST Austria)

The term ‘chromid’ was introduced by Harrison et al. to describe large secondary replicons that have plasmid-type replication but possess essential genes found on chromosomes in other species. The chromid is more similar in all genomic signatures to the true chromosome rather than to other replicons, such as megaplasmids or plasmids. The higher similarity of the chromid with the chromosome likely indicate their convergent evolution. The acquisition of new genes by chromid may be through chromosomal gene duplication or intergation of environmental DNA. At the time of gene gain they may not have been essential and even have functional homologs on the chromosome, while subsequent loss of their homologs from the genome may render chromids essential. However, the advantages of multi-chromosomal genome organization remain unclear.

We performed comparative and evolutionary analysis of 186 genomes from Vibrio genus to reveal the adaptive advantage of such genome organization. Firstly we investigated the pan-genome structure of the genus focusing on gene distribution across the replicons. We found strong correlation of replicons' sizes and explained the results with 'pan-genome' evolutionary model. Secondly we investigated replicons’ architecture revealing structured conserved regions in the primary chromosomes and high variability of the secondary replicons. Then we reconstructed scenarios of genes evolution and estimated gene flows in multipartite genomes. Finally we focused on evolution of gene paralogs revealing over-representation of gene paralogs in the genomes with chromids in comparison to the genomes with single chromosome. Summarised, these results explain different evolutionary trajectories of primary and secondary replicons.

Monday, August 2nd, 18:45

Black holes in the regulation of biofilm formation in Escherichia coli

Maria Tutukina (Skolkovo Institute of Science and Technology)

We will discuss the candidate master-regulators of biofilm formation in pathogenic and non-pathogenic E. coli strains, and low molecular weight ligands capable of changing their DNA-binding capacity to switch the biofilm formation off.

Monday, August 2nd, 19:00

A>G is a hallmark of oxidative damage in mitochondrial and bacterial genomes

Konstantin Popadin (Ecole Polytechnique Federale de Lausanne)

G>T transversion is expected to be one of the strongest markers of oxidative damage in the nuclear genome, however, in mitochondrial and bacterial genomes this substitution is very rare and is not associated with oxidative damage. To discover a new signature of oxidative damage, specific for mitochondrial and bacterial genomes, we performed several analyses.
We hypothesized that since mitochondria are tightly involved in aerobic energy production, it is expected that mtDNA mutational spectra may be affected by the oxidative damage which is increasing with age and also depends on the level of metabolism. Thus, from comparative species data, we can uncover the mtDNA specific signature of the oxidative damage. First, correlating species-specific mtDNA mutational spectra with the generation length of mammalian species we observed only one substitution: A>G which is positively associated with generation length. Second, we compared mtDNA mutational spectra of short-lived cockroaches with long-lived termites and observed results in line with our mammalian findings: an excess of A>G in termites. Third, we analyzed mtDNA mutational spectra of ray-finned fishes and observed that the same A>G substitution positively correlates with the temperature of the environment. Thus, we propose that A>G can be a new hallmark of oxidative damage in mitochondrial genomes.
To extend the logic of our mitochondrial findings we performed the first pilot analyses comparing aerobic and anaerobic bacteria. Our first findings are in line with the mitochondrial observations, suggesting that A>G substitution can shape the high G content of aerobic bacterial species and low G content of anaerobic bacterial species.
Altogether, we propose, that A>G is a marker of oxidative damage in mitochondrial and bacterial genomes. Potential mechanisms of this mutation are under discussion.

Monday, August 2nd, 19:15

Investigating microbial diversity of spontaneous fermentation beer and cider using Hi-C metagenomics

Ignat V. Sonets (Institute of Gene Biology RAS)

Complex bacterial and yeast communities of spontaneously fermented beer and cider produce, besides the alcohol, the metabolites with potential health effects. Augmentation of metagenomics with chromosome conformation capture methods (like Hi-C/3-C) provides improved profiles of microbiome structure and functions. We developed a Hi-C metagenomics-based pipeline for investigating food microbiome with a particular focus on yeast. This pipeline includes the experimental stage, Hi-C-aided reconstruction of metagenome-assembled-genomes [MAGs] and secondary analysis (with yeast comparative genomics analysis).

We analyzed 3 spontaneous fermentation beverages (2 beers and 1 cider) selected from a larger collection based on the 16S rRNA and ITS sequencing (PMID:33279083). Along with the multiple bacterial MAGs, an abundant yeast MAG of Brettanomyces bruxellensis was detected. The Hi-C signal was used to perform the chromosome-wide scaffolding of the yeast MAG. We compared it to most publicly available genomes of the species isolated from beer (n=3) and wine (n=3). Phylogenetic analysis showed that all beer strains (including our MAG) were closely related, while the wine strains were distant and did not form a distinct cluster. Functional enrichment analysis of wine- and beer-specific accessory genes revealed several functions likely linked to the adaptation to distinct niches.

Our results show that the Hi-C metagenomics is a promising technique for in-depth analysis of food microbiome. It allows improved reconstruction of bacterial and yeast genomes compared to conventional metagenomics. A combined investigation of food community structure along with the functional potential of each of its members will allow better understanding of microbial ecology, identification of new probiotics and contribute to the improved quality control and development of products with health-modulating effects.

This work was supported by the Russian Science Foundation [grant 19-74-10092].

Poster session

Modelling Segmental Duplications in the Human Genome ***Eldar Abdullaev*** *(Max Planck Institute for Molecular Genetics)*
Transformer-based model for recognition of quadruplexes using information on physical and chemical DNA properties ***Ivan Agafonov*** *(Higher School of Econimics); Maria Poptsova*
SARS-CoV-2 escapes cytotoxic T cell immune response during long-term infection of immunocompromised patient with non-Hodgkin's lymphoma ***Evgeniia Alekseeva*** *(Skolkovo Institute of Science and Technology); K. Safina; E. Nabieva; S. Garushyants; G. Klink; G. Bazykin*
Two Cobalt Chelatase Subunits Can Be Generated from a Single chlD Gene via Programed Frameshifting ***Ivan Antonov***
Single cell RNAseq-based transcriptome profiling of mesenchymal stromal cells reveals subpopulations with different responses to profibrotic stimuli ***M.S. Arbatsky*** *(Lomonosov Moscow State University), N.A. Basalova, O.A. Grigorieva, N.I. Kalinina, A.Yu. Efimenko*
Evolution of Transcriptional Regulation of Histidine Metabolism in Gram-positive Bacteria ***German Ashniev*** *(IITP RAS); Natalia Sernova; Alexey Shevkoplias; Ivan Rodionov; Irina Rodionova; Alexey Vitreschak; Mikhail Gelfand; Dmitry A Rodionov*

Population specific enhancer affecting optic disc development timespan underlies Glaucoma predisposition. ***Roman Babenko*** *(Institute of Cytology and Genetics, DB RAS), Vladimir Babenko*
Enhancing eukaryotic gene structures by implementing statistical changepoint analysis of expression data using FINDER - a completely automated gene annotator ***Sagnik Banerjee*** *(Iowa State University); Priyanka Bhandary; Margaret Woodhouse; Taner Sen; Roger Wise; Carson Andorf*
Population genomics and population multi-omics of adaptation ***Antonio Barbadilla*** *(Universitat Autònoma Barcelona); Sònia Casillas; Jesús Murga*
Origin and splicing of mutually excluding exons and pre-mRNA secondary structure in human genes of voltage-gated calcium channels ***Ilya S. Belalov*** *(Skolkovo Institute of Science and Technology); Timofei M. Ivanov; Marina Kalinina; Dmitri D. Pervouchine*
The power law of CRISPR-Cas systems ***Ilya S. Belalov*** *(Skolkovo Institute of Science and Technology); Yekaterina S. Pavlova; David Paez-Espino; Andrew Yu. Morozov; Ilya S. Belalov*
Dynamics of dN/dS at short evolutionary distances ***Evgenia A. Belousova*** *(Lomonosov Moscow State University); Anastasia V. Stolyarova; Alexey S. Kondrashov; Georgii A. Bazykin*

Multiscale investigation into the active site composition of AmiN kinase ***Julia Belyaeva*** *(Lomonosov Moscow State University); Alexander Zlobin; Andrey Golovin*
The dependence of homologous recombination rate on the level of heterozygosity in hypervariable fungus *Schizophyllum commune* ***Aleksandra V Bezmenova*** *(Skolkovo Institute of Science and Technology); Elena Zvyagina; Tatiana Neretina; Anna Fedotova; Georgii Bazykin; Alexey Kondrashov*
The expansion of the range of microbial rhodopsins by their artificial sequences ***Elizaveta Bogdanova*** *(Lomonosov Moscow State University); Shaitan K.V.; Novoseletsky V.N.*
Application of the intron sequence of the NXF1 gene in mammalian phylogeny ***Dmitrii Bondaruk*** *(St. Petersburg State University); E.V. Golubkova; L.A. Mamon*
Structure-Based Identification of Small Molecule Inhibitors for Selective Targeting of SARS-CoV-2 Main Protease: An Integrative Computational Approach ***Ivan P Bosko*** *(UIIP NASB); A.M. Andrianov; Yu.V. Kornoushenko; A.D. Karpenko; A.M. Yushkevich; K.V. Furs; A.V. Tuzikov*
Convergent adaptation in mitochondria of phylogenetically distant birds: does it exist? ***Valentina Burskaia*** *(Kharkevich Institute for Information Transmission Problems); Ilja Artyushin; Nadezhda Potapova; Kirill Konovalov; Georgii A. Bazykin*

Calcium coordination in enzymes: the intertwining of structural and functional features ***Michelle F Buyanova*** *(Lomonosov Moscow State University); Arthur O. Zalevsky; Andrey V. Golovin*
HiChew: a tool update for TAD boundaries clustering in development ***Nikolai S Bykov*** *(Skolkovo Institute of Science and Technology); Aleksandra A Galitsyna*
The Spatial Organization of the genome of sea sponge *Halisarca dujardini* ***Alexander V Cherkasov*** *(Skolkovo Institute of Science and Technology); Alina Ryabova; Olga Kozlova; Alexander Finoshin, Oksana Kravchuk; Ekaterina Khrameeva*
Modular assembly of immune-event-labeled synthetic AIRR-datasets for the development and benchmarking of AIRR-based machine learning ***Maria Chernigovskaya****; Victor Greiff*
Trajectory inference methods applied to clinical and bulk transcriptomomic data ***Alexander Chervov*** *(Institute Curie); Andrei Zinovyev*
MSAtoGFA: a Graph Representation of Multiple Sequence Alignments ***Fawaz Dabbaghie*** *(University Hospital Düsseldorf); Tobias Marschall; Olga Kalinina*
DNA-Methylation for the Detection and Distinction of 19 Human Malignancies ***Ludmila Danilova*** *(Johns Hopkins University); John Wrangle; James G. Herman; Leslie Cope*

Tick-borne encephalitis virus phylodynamics ***Andrei A Deviatkin*** *(Sechenov First Moscow State Medical University); Galina Karganova; Yulia Vakulenko; Ivan Kholodilov; Alexander Lukashev*
Gene expression pattern in Edward syndrome: A bioinformatic analysis on what creates significant low life expectancy ***Supantha Dey*** *(University of Dhaka)*
A neural network approach to the QM / MM metadynamics' quantum mechanical description level ***Igor D Diankin*** *(Lomonosov Moscow State University); A. V. Golovin;*
The Classification of *ipaH* Genes in *Shigella* and Enteroinvasive *Escherichia* ***Natalia O Dranenko*** *(IITP RAS); Maria Tutukina; Olga Bochkareva*
The analysis of *Drosophila melanogaster* Hi-C maps with autocorrelation function and Fourier transform ***Alexey I Drozhdev*** *(Lomonosov Moscow State University); A.A.Galitsyna; M.S.Gelfand*
Secondary structure of the SARS-CoV-2 genome affects molecular evolution ***Bogdan E. Efimenko*** *(IKBFU); Alexandr Voronka; Sergey Oreshkov; Konstantin Popadin; Konstantin V Gunbin*
Protein-coding potential of the human repetitive elements ***Artyom A. Egorov*** *(Lomonosov Moscow State University); Sergey Dmitriev*

Insights into the genetic components of chronic back pain ***Elizaveta E Elgaeva*** *(Institute of Cytology and Genetics SB RAS); Maxim B. Freidin; Frances M. K. Williams; Pradeep Suri; Yurii S. Aulchenko; Yakov A. Tsepilov*
Catalytic mechanism of MnmE GTPase as a member of ion-dependent GTPases class ***Evgenia Elizarova*** *(Lomonosov Moscow State University); A.S. Zlobin; A.V. Golovin; A.Y. Mulkidjanian*
Two sequence variants of *yjjM* gene in the *Escherichia coli* genomes ***Vera Emelianenko*** *(IST Austria); Olga Bochkareva; Maria Tutukina; Anna Kaznadzey*
Telling the story of best friends: marker rank statistics ***Alexander Favorov*** *(JHMI); Vasiliy Ramensky; Andrey Mironov*
Discovery of non-AUG PANTs: Proteoforms with Alternative N Termini ***Alla Fedorova*** *(University College Cork); Stephen Kiniry; Pavel Baranov*
Predicted Spike-ORF8 Genomic RNA-RNA Interaction Unique to SARS-CoV-2 May Allosterically Impact the Rate of Nucleocapsid Sub-genomic RNA Synthesis ***Mario A Flores*** *(UTSA); Karla Paniagua; Yufang Jin*
Allele specific transcription factor binding sites mark positive selection loci in the human genome ***Marina Fridman****; P. Bykadorov*

Asymmetrical mutagenesis drives aminoacid composition of the human mitochondrial genome ***Alima Galieva*** *(Immanuel Kant Baltic Federal University); Alina A. Mikhailova; Alina G. Mikhailova; Victor Shamanskiy; Valeria Lobanova; Kristina Ushakova; Konstantin Gunbin; Konstantin Popadin*
Tandem segmentation-classification approach for localization of morphological predictors of *C. elegans* lifespan and movement ***Evgeniy Galimov*** *(AILS); Yakimovich A*
RedC-nf: an automatized pipeline to map RNA-DNA interactions ***Aleksandra A Galitsyna*** *(Skolkovo Institute of Science and Technology); Mikhail D. Magnitov; Aleksey A. Gavrilov; Andrey A. Mironov*
Creation of new antibacterial peptides ***Oxana V Galzitskaya*** *(Institute of Protein Research RAS); S.R. Kurpe; S.Y. Grishin; A.V. Glyakina; M.V. Slizen; A.V. Panfilov; A.P. Kochetov; E.I. Deruysheva; A.V. Machulin; A.K. Surin*
Pentads: a novel pile-up analysis tool for assessing the alterations of chromatin compartments using Hi-C data ***Azat Garaev*** *(Lomonosov Moscow State University); M.D. Magnitov; S.V. Ulianov; S.V. Razin*
Calculation of Sublimation Entropy and Dissociation Constants from a Simple Evaluation of Movement Range of Molecules in Crystals ***Sergiy O Garbuzynskiy*** *(Institute of Protein Research RAS); A.V. Finkelstein*

Whole-genome atlas of transcribed regulatory elements expression during time-course disuse and recovery in skeletal muscles of rats ***Guzel Gazizova*** *(Kazan Federal University); Sergey Pintus; Ilya Akberdin; Ivan Yevshin; Fedor Kolpakov; Oksana Tyapkina; Leniz Nurullin; Ruslan Devyatiyarov; Elena Shagimardanova; Pavel Makhnovskii; Daniil Popov; Oleg Gusev*
Machine-learning model for prediction of new antimicrobial peptides: *Hirudo medicinalis* metagenome case study ***Ekaterina Nikolaevna Grafskaia*** *(FRCC PCM); Maja Malakhova; Victoria Lavrenova; Ivan Latsis; Vladislav Babenko; Vassili Lazarev*
Feature extraction by neural network trained to predict RNA structure ***Lisa Grigorashvili*** *(Skolkovo Institute of Science and Technology); Mikhail Gelfand; Zoe Chervontseva*
Dissection of Loci Underlying Flowering Time of Guar (*Cyamopsis tetragonoloba (L.) Taub.*) via Integrated Transcriptome-Metabolome Analysis ***Elizaveta Grigoreva*** *(ITMO University); Alexander Tkachenko; Aleksandar Beatovic; Serafima Arkhimandritova; Dmitry Karzhaev; Vladimir Volkov; Cecile Ben; Laurent Gentzbittel; Elena Potokina*
QTL mapping of oleic acid content and tocopherol composition in sunflower from VNIIMK collection ***Rim Gubaev*** *(Skolkovo Institute of Science and Technology); S.Boldyrev; E. Martynova; A. Chernova; T. Kovalenko; S. Goryunova; D. Goryunov; T. Peretyagina; Y. Demurin; P. Khaitovich*

Automatic reconstruction of species-specific mitochondrial mutational spectra based on DNA polymorphisms ***Konstantin V Gunbin*** *(Immanuel Kant Baltic Federal University); Konstantin Popadin*
Mitochondrial mutational spectrum in birds: evidence of increased oxidative damage in species with high level of metabolism ***Yury Gusarov*** *(Immanuel Kant Baltic Federal University); Alina G. Mikhailova; Sergey Oreshkov; Alina A. Mikhailova; Dmitry Knorre; Leonard V. Polishchuk; Alexander Kuptsov; Konstantin V. Gunbin; Valentina Burskaya; Konstantin Popadin*
Binary metabolic phenotypes and phenotype diversity metrics for functional characterization of microbial communities ***Stanislav Iablokov*** *(IITP); Dmitry A Rodionov*
mtDNA mammalian evolution: mice evolve by many little steps while elephants with a few big ones ***Dmitrii Iliushchenko*** *(IKBFU); Konstantin Popadin; Konstantin V Gunbin*
Evolutionary repertoire of eukaryotic chromatin-associated proteins ***Ivan Ilnitskiy*** *(FBB MSU); A.A. Mironov; A.A. Zharikova*
Current advances and future perspectives on single-cell Hi-C bioinformatics ***Alexey Ivanov*** *(Higher School of Economics); Aleksandra Galitsyna; Mikhail Gelfand*
Tandem exon duplications expanding splicing repertoire ***Tim Ivanov*** *(Skolkovo Institute of Science and Technology); Dmitri Pervouchine*

Long-range RNA-RNA pairings in human pre-mRNAs and their implications in alternative splicing
Marina Kalinina *(Skolkovo Institute of Science and Technology); Olga Babadei; Dmitry Skvortsov; Svetlana Kalmykova; Olga Dontsova; Dmitri Pervouchine*

Viral metagenomic analysis of publicly available genomic and transcriptomic samples from Simuliidae and Ceratopogonidae insects
Ekaterina Kamanova *(State Research Center of Virology and Biotechnology "Vector"); M.E. Starchevskaya; T.S. Nepomnyashchikh; Denis V. Antonets*

Inversions and translocations in a highly polymorphic fungus *Schizophyllum commune*
Asia Kamyshnikova *(Skolkovo Institute of Science and Technology); Anastasia V. Stolyarova; Georgii A. Bazykin; Alexey S. Kondrashov*

Indexing Sequencing Data at Petabase Scale
Mikhail Karasikov *(ETH Zurich); Harun Mustafa; Daniel Danciu; Gunnar Rätsch; André Kahles*

Design and Bioinformatics Analysis of Novel 2-Arylaminopyrimidine-Based Compounds as Potential Bcr-Abl Tyrosine Kinase Inhibitors
Alexander Andrianov; Yury V Karnaushanka; ***Anna D Karpenka*** *(UIIP NASB); Ivan P Bosko; Zhanna Ignatovich; Elena Koroleva; Julia Siniutsich*

Evolution of different genotypes of hepatitis B and hepatitis delta viruses in endemic region ***Anastasia Karlsen*** *(I. Mechnikov NIIVS); O.V. Kuzmin; F.A. Asadi Mobarhan; O.V. Isaeva; S.S. Sleptsova; Victor A. Manuylov; Andrei A. Pochtovyi; Vladimir A. Gushchin; K.K. Kyuregyan; M.I. Mikhailov*
Search for the genetic determinants of yeast mtDNA suppressivity ***Nataliia Kashko*** *(Lomonosov Moscow State University); Dmitry Knorre*
Computational inference of metabolic phenotypes for new bacterial isolate and metagenome-assembled genomes *Marat Kazanov; Semen Leyn;* ***Dmitry A Rodionov*** *(IITP)*
mtDNA codon usage of Chordata is primarily shaped by mutational spectra ***Eldar Khaibulin*** *(IKBFU); Alina G. Mikhailova; Victor Shamanskiy; Konstantin Popadin*
In search of molecular traces of adaptations to life in high mountains in mammals ***Darya Khaleneva*** *(Saint Petersburg State University); O.V. Bondareva; S.Yu. Bodrov; T.V. Petrova; E.A. Genelt-Yanovskiy; N.I. Abramson*
Composition of metabolic loci in bacterial genomes ***Evgenia Khodzhaeva*** *(Institute for information Transmission Problems RAS); Zoe Chervontseva*
Multiomic view of gene regulation in Plasmacytoid Dendritic Cells ***Maksim Kholmatov*** *(EMBL); Judith B. Zaugg*

Deterministic DNA embeddings with no predefined k-mer length ***Bogdan Kirillov*** *(Skolkovo Institute of Science and Technology); Maxim Panov*
Sensitive mapping of transposon insertions using Cas9-Targeted Nanopore sequencing and a novel pipeline ***Ilya Kirov*** *(VNIISB); Pavel Merkulov; Sofya Gvaramiya; Murad Omarov; Maxim Dudnikov; Elizaveta Kolganova; Roman Komakhin; Mikhail Divashuk; Gennady Karlov; Alexander Soloviev*
A comparative genomic study of attenuated and virulent strains of the African swine fever virus genotype I and serogroup 2 ***Galina Koltsova*** *(Federal Research Center for Virology and Microbiology); S. Namsrayn; A. Koltsov*
Generative-Adversarial Neural Networks for Quadruplex and Z-DNA prediction ***Nikita Konstantinovsky*** *(Higher School of Economics); Nikita Kazakov; Maria Poptsova*
Precipitant ions influence on lysozyme mono- and oligomers stability investigated by molecular dynamics simulation ***Yuliya Kordonskaya*** *(Shubnikov Institute of Crystallography of Federal Scientific Research Centre "Crystallography and Photonics" Russian Academy of Sciences); Vladimir Timofeev; Yulya Dyakova; Margarita Marchenkova; Yuriy Pisarevskiy; Michael Kovalchuk*
A comparison of computational tools for average nucleotide identity estimation ***Aleksei Korzhenkov*** *(NRC KI)*

Evolution of mitochondrial DNA inheritance patterns *Pavel Kravchenko;* ***Dmitry Knorre*** *(Belozersky Institute of Physico-Chemical Biology)*
Evolutionary insights into microbial oxylipin signalling ***Georgy F Kurakin*** *(Tver State Medical University); Anna Samoukina; Nadezhda Potapova*
Genotype-by-environmental interactions for chronic back pain ***Ivan A Kuznetsov*** *(Skolkovo Institute of Science and Technology); Yakov A. Tsepilov; Yurii S. Aulchenko*
Transfer learning for whole-genome G-quadruplex annotation ***Pavel Latyshev*** *(Higher School of Economics); Maria Poptsova*
Dating gene cluster acquisition: The case of Form III RubisCO-mediated transaldolase variant of the Calvin cycle ***Alexander V Lebedinsky*** *(Federal Research Centre «Fundamentals of Biotechnology of the Russian Academy of Sciences); Ilya Kublanov; Evgenii Frolov; Elizaveta Bonch-Osmolovskaya; Nikolay Chernyh*
Bioinformatic analysis of prokaryotic rotary ATPases: revealing the correlations between coupling ion type and microorganism physiology ***Anna Valeryevna Litvin*** *(Lomonosov Moscow State University); Maria S. Shoshinova; Alexander S. Yakushev; Anna S. Lapashina; Boris A. Feniouk*

A billion-year trend of amino acid substitutions in the mitochondrial genome: comparative analysis of different taxa ***Valeriia V Lobanova*** *(Immanuel Kant Baltic Federal University); Konstantin Popadin; Alina G. Mikhailova; Kristina Ushakova; Victor Shamanskiy; Alina Mikhailova; Alima Galieva*
Complex evolution of Fbxl21 gene in mammals ***Vassily A Lyubetsky*** *(Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute)); G.A. Shilovsky; O.A. Zverkov; L.I. Rubanov; A.V. Seliverstov*
On Selective Imputation Using Beagle v.5. ***Sergey V. Malov*** *(Saint-Petersburg State University); Alexander Kanapin; Anastasia Samsonova; Gaik Tamazian*
The stochastic nature of the interaction between CRISPR interference and replication events provides the coexistence of virus and bacteria populations ***Viktor Mamontov*** *(Skolkovo Institute of Science and Technology); Yaroslav Ispolatov; Ekaterina Semenova*
Sequence determinants of cross-regulatory feedback loops in RNA-binding protein interaction network ***Sergey Margasyuk*** *(FBB MSU); Dmitri Pervouchine*
mtmlSEM with random effects uncovers genotype-phenotype associations through pleitropic SNPs *Georgy A Meshcheryakov (SPbPU); Anna Igolkina*
Using protein-protein interaction and gene networks to improve Connectivity Map ***Mariia Minaeva*** *(Moscow Institute of Physics and Technology); Khalimat Murtazalieva; Yulia Medvedeva*

The atlas of expression of cis-regulatory elements of human craniofacial embryonic tissues
Elena Minkina *(Kazan Federal University); Akberova Natlia; Shagimardanova Elena; Adameyko Igor; Gusev Oleg*

Efficient detection of structural variations using capture-Hi-C data
Evgeniy A Mozheiko *(Institute of Cytology and Genetics SB RAS); Maria Gridina; Emil Valeev; Veniamin Fishman; Ludmila P. Nazarenko; Maria E. Lopatkina; Igor N. Lebdev; Maria I. Yablonskaya; Viktoria Yu Voinova; Zhanna G. Markova; Nadezhda V. Shilova*

Optimization of the directed evolution experimental design by the genetic algorithm theory
Ekaterina Myasnikova *(Peter the Great Saint-Petersburg Polytechnical University); Alexander Spirov*

A sensitive method for search for orthologs of de novo annotated lncRNAs
Dmitry E Mylarshchikov *(FBB MSU); Andrey Mironov*

Bcbio – a set of best practices NGS data processing pipelines for bioinformatics cores and diagnostic labs
Sergey Naumenko *(Harvard Chan School of Public Health)*

Analysis of Poxviridae and Coronaviridae genomes using k-mer frequency profiles and comparison with alignment-based approach
Tatyana S Nepomnyashchikh *(SRC VB "Vector"); D.V. Antonets; M.E. Starchevskaya; T.V. Tregubchak; A.N. Shvalov; E.V. Gavrilova; S.N. Shchelkunov; Rinat A. Maksyutov*

Evolution scheme of the CRISPR-Cas locus in *Xanthomonas citri* can be reconstructed based on its spacers content, but the result disagrees with the strain phylogeny for the species
M.S. Nikitin *(MIPT); N.V. Zakharevich; I.I. Artamonova*

Comparative genomics of placozoans: fast evolution of neuronal genes in nerveless animals
Mikhail A. Nikitin *(Lomonosov Moscow State University); Simha I. Borman; Daria Yu. Romanova; Maria D. Logacheva; Leonid L. Moroz*

A Deep Learning-Based Approach for Prediction of Potential HIV-1 Entry Inhibitors Targeting CD4-Binding Site of GP120
Alexander Andrianov; ***Grigory I. Nikolaev*** *(United Institute of Informatics Problems of the National Academy of Sciences of Belarus); Nikita Shuldov; Ivan P Bosko; Alexander Tuzikov*

Is bacterial species pangenome structure associated with habitat diversity?
Daria Nikolaeva *(Skolkovo Institute of Science and Technology); Sofya K. Garushyants*

vsFilt: a new tool for structural filtration and virtual screening
Dmitry K Nilov *(Lomonosov Moscow State University); I.V. Gushchina; D.A. Suplatov; V.K. Švedas*

New features in GADMA for efficient and flexible inference of demographic history
Ekaterina Noskova *(ITMO University); Vladimir Ulyantsev*

A potential of polygenic risk scores as a tool to inform population screening for breast cancer
Arina Nostaeva *(Institute of Cytology and Genetics); Tatiana Shashkova; Lennart Karssen; Yurii Aulchenko*

Chromatin association of BIRC5/survivin protein genome binding in human CD4+ T cells ***Nina Oparina*** *(GU); Malin Erlandsson; Karin M.E. Adersson; Venkatagavan Chandrasekaran; Anastasios Damdimopoulos; Sofia T. Silfverswärd; Maria I. Bkarewa*
Lipidome composition of the healthy adult human brain ***Maria Osetrova*** *(Skolkovo Institute of Science and Technology); Ekaterina Khrameeva; Anna Tkachev; Elena Stekolschikova; Aleksandra Mitina; Olga Efimova; Philipp Khaitovich*
Pupae recapitulate the embryonic expression program in insects undergoing radical metamorphosis ***Aleksandra Ozerova*** *(Skolkovo institute of science and technology); Mikhail Gelfand*
Study of influence of homology modeling on the prediction of protein stability change upon mutation ***Marina A Pak*** *(Skolkovo Institute of Science and Technology); Dmitry N. Ivankov*
What can substitution biases tell us about the origin of SARS-CoV-2 ***Alexander Y Panchin*** *(Institute for Information Transmission Problems, RAS)*
Classification of families of DNA-recognizing protein domains based on structural features of DNA-protein complexes ***Vera Panova*** *(Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University); Eugene Baulin; Anna Karyagina; Andrey Alexeevski; Sergey A Spirin*

Genomic determinants of viral host diversity ***Marina Parr*** *(Technical University of Munich); Dmitrij Frishman; Tamir Tuller*
Structural descriptors-based clustering for the evaluation of the patterns in 3D-structure of vitamin D receptor ***Karina Pats*** *(ITMO University); Ferdinand Molnar*
In-silico identification of key regulators instigating the pre-meiotic phase during respiration in *Saccharomyces cerevisiae* ***Soumita Podder*** *(Raiganj University)*
Identification and structural analysis of long noncoding RNAs ***Artem Y Pronozin*** *(institute of cytology and genetics sb ras); Afonnikov D.*
Structure-based design of monoclonal antibodies against SARS-CoV-2 ***Elena Rukhlova*** *(State Research Center of Virology and Biotechnology "Vector"); Anastasia Bakulina; Dmitry Shcherbakov; Iuliia Merkuleva*
Restriction-modification systems in microbiome of Antarctic Deep Lake ***I. Rusinov*** *(Belozersky Institute); A. Ershova; A. Alexeevski; A. Karyagina; S. Spirin*
A new database for genome-wide RNA-chromatin interactome ***Grigory Ryabykh*** *(Moscow State University); Anastasiya Zharikova; Andrey Sigorskikh; Yuriy Korostelev; Andrey Mironov*

Supposed properties of enhancers associated with human-specific brain genes ***Marat Sabirov*** *(Institute of Gene Biology RAS); Ekaterina Khrameeva*
Molecular epidemiology of HIV in Oryol district, Russia ***Ksenia Safina*** *(Skolkovo Institute of Science and Technology); Georgii Bazykin; Yulia Sidorina; Natalya Efendieva; Elena Belonosova; Dmitry Kireev*
Comparative genomics of flax pathogen *Fusarium oxysporum* f.sp. lini *A. Kanapin; A. Samsonova;* ***A. Logachev****; M. Bankin; T. Rozhmina; M. Samsonova*
MutAnt: Mutation annotation machine learning algorithm for pathogenicity evaluation of single nonsynonymous nucleotide substitutions in cancer cells ***Aleksandr Sarachakov*** *(Skoltech); Viktor Svekolkin; Zoya Antysheva; Jessica Brown; Alexander Bagaev; Nathan Fowler; Mikhail Gelfand*
Xenobiotic associated CYP450 9e2 family in Formicinae: protein family evolution and population study of black garden ant *Lasius niger* CYP9E2 genes expression ***Viktoria Scobeyeva*** *(Moscow State University); Svetlana Zhukova; N. Yu. Oparina; I. Artyushin*
Genome-wide search for positive selection in *Schizophyllum commune* ***Maria Selifanova (FBB MSU)****; Galina Klink; Anastasia Stolyarova; Georgy Bazykin*

Evolution of Insertions Sequences in the Evolution of Intracellular Pathogens *B. mallei* ***Ariadna Semenova*** *(MSU); Olga O. Bochkareva*
How to improve real-life applicability of drug-target interaction prediction? ***Ilya Senatorov*** *(HIPS); Olga V. Kalinina*
Development of a searching algorithm for new cell-penetrating peptides using machine learning methods ***Maria Y Serebrennikova*** *(Federal State Budgetary Institution Federal Research and Clinical Center of Physical-chemical Medicine Federal Medical Biological Agency); Ekaterina Grafskaia; Vassili Lazarev*
Genomic adaptations of chironomids to extreme environments ***Nurislam Maratovich Shaikhutdinov (****Skoltech); Galina Klink; Sofya Garushyants; Olga Kozlova; Elena Shagimardanova; Guzel Gazizova; Natalia Gogoleva; Andrey Przhiboro; Oleg Gusev; Georgii Bazykin*
Somatic deletions in the human mitochondrial genome: the global secondary structure, G-quadruplexes and direct nucleotide repeats explain majority of breakpoints ***Victor Shamanskiy*** *(Center For Mitochondrial Functional Genomics, IKBFU); Alina Mikhailova; Kristina Ushakova; Alina Mikhaylova; Sergei Oreshkov; Dmitry Knorre; Evgenii Tretiakov; Ilia Mazunin; Konstantin V Gunbin; Konstantin Khrapko; Konstantin Popadin*
Deep learning for demographic inference from whole-genome sequences ***Vladimir Shchur*** *(HSE University); Evgeniy Khomutov; Kenenbek Arzymatov; Viktoriya Vasileva*

Using transcriptomic data for reconstruction of earthworm phylogeny on the example of the *Eisenia nordenskioldi* complex ***Sergei V Shekhovtsov*** *(Institute of Cytology and Genetics SB RAS); Alexandra A. Shipova; Tatiana V. Poluboyarova; Sergei E. Peltek*
Sequence analysis of human TMTC proteins reveals their enzymatic activity and ligand binding sites ***Vladimir A. Shitov*** *(Siberian State Medical University); B. Eisenhaber; S. Sinha; C. K. Jadalanki; Q. W. Tan; F. L. Sirota; F. Eisenhaber*
Deep learning model for Hi-C maps prediction ***Alexey Shkolikov*** *(Moscow State University); Aleksandra Galitsyna; Mikhail Gelfand*
A search of embryonic selection traces: transmission disequilibrium of slightly-deleterious variants from parents to the offspring with Down syndrome ***Kseniia Sholokhova*** *(Center for Mitochondrial Functional Genomics); Konstantin Gunbin; Victor Shamanskiy; Louis-Alexandre Ongaro; Konstantin Popadin*
SPAN and JBR - tools for semi-supervised peak calling ***Oleg Shpynov*** *(JetBrains Research); Roman Chernyatchik; Petr Tsurinov; Maxim Artyomov*
Regulation of splicing in Sirt6-KO mouse brain ***Artemy Shumskiy*** *(Skoltech); Dmitrii Smirnov; Ekaterina Khrameeva; Deborah Toiber*

New RNAs revealed by RNA-chromatin interactions ***Andrey Sigorskikh*** *(Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University); Grigory Ryabykh; Yuri Korostelev; Anastasiya Zharikova; Andrey Mironov*
optimalTAD: a novel algorithm for Topologically Associating Domain optimal set prediction ***Dmitrii Smirnov*** *(Skolkovo Institute of Science and Technology); Ekaterina Khrameeva*
The species-specific burden of slightly deleterious mutations in mammals ***Anastasia D Sokol*** *(Immanuel Kant Baltic Federal University); Alina Mikhailova; Konstantin Gunbin; Konstantin Popadin*
New database and software for topology-based pathway analysis of human gene expression data: overview and application example ***Maxim I Sorokin*** *(Sechenov University); Alexander Simonov; Victor Tkachev; Anton Buzdin*
Fast simulation of viral genealogies in global pandemic *Vladimir Shchur;* ***Vadim Spirin*** *(HSE University); Victor Pokrovskii; Evgeni Burovski; Russell CorbettDetig*
Predicting quality of phylogenetic inference by machine learning *Aleksei Efremov;* ***Sergey A Spirin*** *(Belozersky Institute)*
PhyloBench, a benchmark for evaluation of phylogenetic programs *Alexei Efremov; Andrey Sigorskikh; Dmitry Penzar; Anna Karyagina;* ***Sergey A Spirin*** *(Belozersky Institute)*

Reverse transfer of ideas from genetic algorithms to synthetic biology ***Alexander V Spirov*** *(IEPhB RAS); Ekaterina M. Myankova*
Modeling of active transport of bicoid mRNA in early Drosophila embryo development ***Alexander V Spirov*** *(IEPhB RAS); Ekaterina M. Myasnikova; Marat A. Sabirov*
The rich inner world of Colorado potato beetles – a metagenomic survey of viral diversity in public data ***Maria Starchevskaya*** *(State Research Center of Virology and Biotechnology "Vector"); Denis V Antonets; Yuri Vyatkin; Tatyana Tregubchak; Tatyana Bauer; Ekaterina Kamanova; Tatyana Nepomnyaschih; Ulyana Rotskaya; Elena Kosman*
Integrative approach to intermediate filament structure ***Sergei V Strelkov*** *(KU Leuven); Anastasia V. Lilina; Giel Stalmans; Pieter-Jan Vermeire; Jan Fiala; Petr Novak*
Gene Set Co-regulation Analysis ***Vladimir Sukhov*** *(ITMO University); Maxim Artyomov; Alexey Sergushichev*
Functional significance of structural plasticity of the loop 422-437 in the active site of Neuraminidase A from *Streptococcus pneumoniae* ***Dmitry Suplatov*** *(Lomonosov Moscow State University); Yana Sharapova; Vytas Švedas*
Zebra3D: a tool for bioinformatic analysis of 3D-determinants of functional diversity in protein superfamilies using machine learning *Daria Timonina; Yana Sharapova; Vytas Švedas;* ***Dmitry Suplatov*** *(Lomonosov Moscow State University)*

Image processing methods for analysis of cellular structure of epithelial tissue ***Svetlana Surkova*** *(Peter the Great St. Petersburg Polytechnic University); Maria Samsonova; Konstantin Kozlov*
On Selective Imputation Using Beagle v.5. ***Gaik Tamazian*** *(St. Petersburg State University); Alexander Kanapin; Anastasia Samsonova; Sergey V. Malov*
Identification of conservative patterns of G-quadruplexes and transcription binding sites ***Julia Terentjeva*** *(Higher School of Economics); Maria Poptsova*
Utilizing machine learning methods and cellular gene expression patterns of longevity for identification of new lifespan-extending interventions ***Stanislav Tikhonov*** *(Lomonosov Moscow State University); Alexander Tyshkovskiy*
Bioinformatic analysis of local 3D-structure patterns in protein superfamilies ***Daria Timonina*** *(Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics); Yana Sharapova; Vytas Švedas; Dmitry Suplatov*
Heat shock protein 90 as a long-term buffer of a species-specific mutational burden ***Valeria N Timonina*** *(Immanuel Kant Baltic Federal University, Center for Functional Mitochondrial Genomics); Anastasia Sokol; Evgenii Tretiakov; Konstantin V Gunbin; Konstantin Popadin*

Data-driven approach for traits definition for multivariate genetic analysis of N-glycome composition ***Anna Timoshchuk*** *(MIPT); Nadezhda A. Potapova; Gordan Lauc; Tim Spector; Sodbo Sharapov; Yurii S. Aulchenko*
Shared heredity: a method to model genetic basis of correlated traits ***Evgeny Tiys*** *(Institute of cytology and genetics); Gulnara R. Svishcheva; Sofia G. Feoktistova; Elizaveta E. Elgaeva; Sodbo Z. Sharapov; Yakov A. Tsepilov*
Assembly and annotation of the sable (*Martes zibellina*) and pine marten (*Martes martes*) genomes ***Andrey Tomarovsky*** *(Computer Technologies Laboratory, ITMO University); Azamat A. Totikov; Violetta R. Beklemisheva; Polina L. Perelman; Natalia A. Serdyokova; Tatiana Bulyonkova; Ksenia A. Koniaeva; Alexei V. Abramov; Alexander S. Graphodatsky; Klaus-Peter Koepfli; Roger A. Powell; Sergei F. Kliver*
Reconstruction of the demographic history for three populations of the least weasel *Mustela nivalis* ***Azamat Totikov*** *(Computer Technologies Laboratory, ITMO University); Andrey A. Tomarovsky; Polina Perelman; Natalia Serdyokova; Violetta R. Beklemisheva; Tatiana M. Bulyonkova; Karol Zub; Viktor V. Panov; Anna S. Mukhacheva; Alexei V. Abramov; Klaus-Peter Koepfli; Alexander S. Graphodatsky; Jose Melo-Ferreira; Sergei F. Kliver*
FARM – new Association Rule Mining method using Fishbone diagrams ***Petr Tsurinov*** *(JetBrains Research); Oleg Shpynov; Maxim Artyomov; Nina Lukashina; Daria Likholetova*

Prediction of inton-exon gene structure for reads by a convolution neural network ***Leonid Uroshlev*** *(Institute of Higher Nervous Activity and Neurophysiology RAS); A.P. Beletskiy*
Fitness landscape learning, canalization and fast evolution ***Sergey Vakulenko*** *(Institute for Mechanical Engineering Problems, RAS, S. Petersburg Problems); John Reinitz; Ovidiu Radulescu; Dmitry Grigoriev*
Semi-extractability of coding and non-coding RNAs ***Anna Valyaeva*** *(Lomonosov Moscow State University); Maria A. Tikhomirova; Anastasia A. Zharikova; Anna V. Fedotova; Andrey A. Mironov; Eugene V. Sheval*
Systematic analysis of age-dependent lipidome alterations in rat tissues ***Anastasiia Velikanova*** *(Skoltech); Aleksandra Mitina; Elena Stekolschikova; Olga Efimova; Ekaterina Tyulkova; Elena Rybnikova; Pavel Mazin; Philipp Khaitovich*
De novo identification of alternative polyadenylation from RNA-seq data ***Maria Vlasenok*** *(Skoltech); Dmitri Pervouchine*
Analysis of the mutation spectrum and selective constraints of SARS-CoV-2 ***Alexandr Voronka*** *(Immanuel Kant Baltic Federal University); Bogdan Efimenko; Sergei Oreshkov; Konstantin Gunbin; Konstantin Popadin*

Small RNAs in *Paramecium* response to *Preeria caryophila* infection ***Yulia Yakovleva*** *(Saint Petersburg State University); Sivarajan Karunanithi; Marcello Pirritano; Franziska Drews; Martin Simon; Alexey Potekhin*
Modelling of the evolutionary pathway from anti-restriction to anti-CRISPR function ***Daria Yanovskaya*** *(Moscow Institute of Physics and Technology (National Research University)); M.A. Skutel; O.S. Musharova; A. B. Isaev*
Genomic analysis of skin cancers from Xeroderma Pigmentosum subgroups revealed mechanisms behind UV mutational signatures formation ***Andrey A Yurchenko*** *(Institut Gustave Roussy); Tirzah B. P. Lajus; Hiva Fassihi; Chikako Nishigori; Konstantin Gunbin; Fatemeh Rajabi; Ismael Padioleau; Carlos F. M. Menck; Alain Sarasin; Patricia Kannouche; Sergey Nikolaev*
Let's go analytic! Two exact models of the mtDNA mutagenesis and their ramifications ***Valerian A Yurov*** *(Immanuel Kant Baltic Federal University); Konstantin Popadin*
Single Cell Navigator allows cross-matching of public scRNA-seq data: study case of tumor immune microenvironments ***Konstantin Zaitsev*** *(ITMO University); Maria Firulyova; Maxim N. Artyomov*
Causal relationships between human IgG N-glycosylation traits and twelve associated diseases ***Olga O. Zaitseva*** *(Genos Ltd); Sodbo Sharapov; Gordan Lauc; Lucija Klaric; Yakov Tsepilov*

pmx.qmmm: automated protein structure and topology generation for QM/MM calculations
Arthur O Zalevsky *(Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS); Alexander Zlobin; Andrey Golovin*

Assembling genomes on chromosome-level leading by the example of two malaria vector genomes
Anton A Zamyatin *(ITMO University); Pavel Avdeyev; Jiangtao Liang; Atashi Sharma; Chujia Chen; Varvara Lukyanchikova; Nikita Alexeev; Zhijian Tu; Max A Alekseyev; Igor V. Sharakhov*

In silico design of inhibitors of cathepsin G based on β-ketophosphonate
Nikita Zernov *(Peter the Great St. Petersburg Polytechnic University); L.S. Hunanyan*

RNA-DNA integractome analysis
Anastasia A Zharikova *(FBB MSU); Andrey I. Sigorskikh; Yuriy D. Korostelev; Andrey A. Mironov*

Features of chromatin structure & gene expression during *D. discoideum* development
Irina Zhegalova *(Skoltech); A. Galitsyna; A. Luzhin; S. Ulianov; E. Khrameeva*

Primary sequence of the Japanese quail's nucleolar organizer region
Alina A Zhukova *(The Herzen State Pedagogical University of Russia); Zakharov G.A.; Kulak M.M.; Saifitdinova A.F.*

Development of method for predicting synergistic combinations of small molecules based on RNA-seq data ***Dinara Ziganshina*** *(Moscow Institute of Physics and Technology); Khalimat Murtazalieva; Yulia Medvedeva*
A platform for storage and analysis of results of genome-wide association studies of sheep ***Alexander S Zlobin*** *(Kurchatov Genomic Center of Institute of Cytology and Genetics SB RAS); Anatoliy Kirichenko; Tatyana Shashkova; Natalya Volkova; Pavel Borodin; Lennart Karssen; Yakov Tsepilov; Yurii Aulchenko*
DESMOND 2.0: Identification of differentially expressed biclusters and investigation of their network properties ***Olga Zolotareva*** *(Technical University of Munich); Olga I. Isaeva; Zoe Chervontseva; Alexey Savchik; Evgenia Zotova; Olga Lazareva; David B. Blumenthal; Nina K. Wenke; Martin Ester; Jan Baumbach*