BiATA'19

BIOINFORMATICS: FROM ALGORITHMS TO APPLICATIONS

CONFERENCE PROCEEDINGS & CONFERENCE SCHEDULE

June 20–22, 2019

St. Petersburg, Russia

Bioinformatics: from Algorithms to Applications 2019
Conference Schedule

B – Break	I – Invited Talk	O – Opening or Closing Talk	F – Featured Talk
T – Talk	D – Dinner	P – Posters	

THURSDAY – JUNE 20

9:00AM–9:45AM	B	**Registration**
9:45AM–10:00AM	O	**Opening Ceremony** Alla Lapidus, *SPbU* Anton Korobeynikov, *SPbU*
10:00AM–11:00AM	I	***Direct queries of metagenome assembly graphs reveal hidden biological complexity*** C. Titus Brown *UC Davis*
11:00AM–11:20AM	T	**Synteny paths for assembly graphs comparison** Evgeny Polevikov *ITMO University*
11:20AM–11:40AM	T	**Indexing De Bruijn graphs with minimizers** Camille Marchet *Lille University*
11:40AM–12:00PM	T	**PathRacer: racing profile HMM paths on assembly graph** Alex Shlemov *Saint Petersburg State University*
12:00PM–12:30PM	B	**Break**
12:30PM–1:30PM	I	**Reproducible metagenomic at scale** Mick Watson *Roslyn Institute, University of Edinburgh*
1:30PM–1:50PM	T	**Recovery of closed genomes of member species from enrichment reactor microbial communities using long read metagenomics** Krithika Arumugam *Nanyang Technological University*
1:50PM–2:10PM	T	**Computational vision to detect chromosomal loops** Axel Cournac *Institut Pasteur*
2:10PM–2:30PM	T	**Whole-chromosome assembly and analysis of hybridogenetic lineages of the desert ant Cataglyphis hispanica with instaGRAAL, a Hi-C based scaffolder and polisher** Lyam Baudry *Institut Pasteur*

2:30PM–3:30PM	B	**Lunch**
3:30PM–4:30PM	F	**Biotech from theory to reality: the journey of creating and marketing groundbreaking technology** Andrey Perfilyev *Atlas*
4:30PM–4:50PM	T	**SPAligner: alignment of long diverged molecular sequences to assembly graphs** Dmitry Antipov *Saint Petersburg State University*
4:50PM–5:10PM	T	**Million sequences indexing** Antoine Limasset *CNRS*
5:10PM–5:30PM	T	**Hybrid RNA-Seq assembly and its application to transcriptome study of Cynara cardunculus** Andrey Prjibelski *Saint Petersburg State University*
5:30PM–5:50PM	T	**Combinatorial Scoring of Phylogenetic Trees and Networks Based on Homoplasy-Free Characters** Max Alekseyev *George Washington University*
6:10PM–9:00PM	D	**Meet & greet**
FRIDAY – JUNE 21		
10:00AM–11:00AM	I	***Metabolic adaptation in the human gut microbiota during pregnancy and infancy*** M. Pilar Francino *FISABIO*
11:00AM–11:20AM	T	**Higher-order and multidimensional epistasis in high-throughput experimental data** Dmitry Ivankov *Skolkovo Institute of Science and Technology*
11:30PM–12:00PM	B	**Break**
12:00PM–1:00PM	I	**Insights into the human gut microbiota from a (meta-)genomic perspective** Rob Finn *EMBL-EBI*
1:00PM–1:20PM	T	**DYSBIOSIS SIGNATURES FROM METAGENOMES REVEAL DIFFERENCES AND SIMILARITIES BETWEEN INFLAMMATORY BOWEL DISEASES** Vadim Dubinsky *Tel-Aviv University*

1:20PM–1:40PM	T	**On the Verge of Colistin Resistance: Genetic Determinants Mediating Intermediate Colistin Resistance in Klebsiella pneumoniae** Sima Tokajian *Lebanese American University*
1:40PM–2:00PM	T	**Antimicrobial Resistance and Clonality of Streptococcus pneumoniae Isolates in Russia** Irina Tsvetkova *Pediatric Research and Clinical Center for Infectious Diseases*
2:00PM–3:30PM	B	**Lunch**
3:30PM–4:30PM	I	**Factors Influencing Soil and Air Microbiomes** Rob Knight *UC San Diego*
4:30PM–4:50PM	T	**Computational biology and agriculture: soil microbiome** Evgeny Andronov *FSBEI of Higher Education SPSU*
4:50PM–5:10PM	T	**Revealing mechanisms of the Bacillus thuringiensis host specificity via modeling coevolution of Cry toxins and their receptors** Yury V. Malovichko *All-Russia Research Institute for Agricultural Microbiology (ARRIAM)*
5:10PM–5:30PM	T	**Genomic determinants underlying rodenticidal properties of the Salmonella enteritidis var. Issatschenko** Kirill Antonets *All-Russia Research Institute for Agricultural Microbiology (ARRIAM)*
5:30PM–5:50PM	T	**Bacteriophage recombination site helps to reveal genes potentially acquired through horizontal gene transfer** Maria A. Daugavet *Institute of Cytology, Russian Academy of Sciences*
6:00PM–7:00PM	P	**Poster Section**
SATURDAY – JUNE 22		
10:00AM–11:00AM	I	***A moving landscape of comparative genomics in mammals*** Stephen O’Brien *Saint Petersburg State University*
11:00AM–11:20AM	T	**Higher-order and multidimensional epistasis in high-throughput experimental data** Dmitry Ivankov *Skolkovo Institute of Science and Technology*
11:00AM–11:20AM	T	**A universal transcriptomic signature of age reveals the temporal scaling of Caenorhabditis elegans aging trajectories** Andrei E. Tarkhov *Skolkovo Institute of Science and Technology*

11:20AM–11:40AM	T	**DASE-AG: conditional-specific differential alternative splicing events estimation method for around-gap regions** Kouki Yonezawa *Nagahama Institute of Bio-Science and Technology*
11:40AM–12:15PM	B	**Break**
12:15PM–13:15PM	I	**Adapting bioinformatics to bacteriophage genomics and virome studies** Marie-Agnès Petit *Micalis Institute, INRA*
1:15PM–1:35PM	T	**NPS: scoring and evaluating the statistical significance of peptidic natural product–spectrum matches** Azat Tagirdzhanov *Saint Petersburg State University*
1:35PM–1:55PM	T	**Local sequence alignment using intra-processor parallelism** Alexander Tiskin *University of Warwick*
1:55PM–2:15PM	T	**HEDGE: Highly accurate GPU-powered protein-protein docking pipeline** Timofei Ermak *Biocad*
2:15PM–2:35PM	T	**Probabilistic model of V-D junction formation in human Ig heavy chain genes and its application** Evgeny A. Bakin *Saint Petersburg State University, Bioinformatics Institute*
2:45PM–4:00PM	B	**Lunch**
4:00PM–5:00PM	I	**Connecting the Microbiome and Ecology to the Gut-Brain Axis** Rob Knight *UC San Diego*
5:00PM–5:20PM	T	**Reconstructing haplotype-specific cancer genome karyotypes with multiple sequencing technologies** *Sergey Aganezov* *Johns Hopkins University*
5:20PM–5:40PM	T	**In pursuit of molecular mechanism for induced granulocytic differentiation: systems biology approach** Svetlana Novikova *Institute of Biomedical Chemistry*
5:40PM–6:00PM	T	**Gene Set Mining In Context Relevant Pubmed Corpora** Christophe Van Neste *King Abdullah University of Science and Technology (KAUST), Ghent University*

6:00PM–6:15PM	O	**Closing remarks** Alla Lapidus *Saint Petersburg State University*
7:30PM–10:00PM	D	**VOGIS Evening Reception**

The conference was supported by the Russian Fund for Basic Research (grant 19-016-20002)

Conference Sponsors

BIOCAD

Biotechnology Company

АССОЦИАЦИЯ ВЫПУСКНИКОВ СПбГУ

25 YEARS RFBR

atlasbiomed

Media Partners

BMC

Лекториум

MONOMAX PCO

Professional Conference Organizer

Monomax PCO offers full expertise in meeting management since 1991. The professionals of Monomax have a vast experience in different aspects of the MICE industry. They are always eager to manage events with their greatest personal care to guarantee the highest standards of service.

Why contact *Monomax PCO* when planning your congress, etc.?

TIME is a valuable asset. You get a remarkable **time cost reduction** by handing over technical tasks of **congress management** to our team.

COSTS SAVING - The rates for services offered by our company can be lower than the rates negotiated by you as an independent party. We have already got a large network of proven suppliers so why not benefit from our resources?

PROFESSIONAL BUDGETING AND FINANCIAL MANAGEMENT – We provide qualified assistance in draft budget planning and registration fee estimation, account management and payments handling, liaison with vendors and many other aspects of financial planning and management.

ADVANCED TECHNOLOGIES – Company's in-house integrated congress management software – Alternative Events – is the modern instrument of any size event administration. It offers mechanisms of delegate on-line registration, abstract handling and Internet payment processing. For congress secretariat it is a useful tool for event Web site support, customized reports generation and cash flow management.

QUALIFIED SECRETARIAT MANAGEMENT - Company's experienced personnel with excellent English language skills is able to accomplish all the tasks and duties of professional congress Secretariat with maximum efficiency and accuracy.

ON-SITE MANAGEMENT – Our team will provide professional on-site coordination throughout the congress to control all services and to resolve any possible emergencies. Our personnel speak good English and we supply all the necessary equipment for registration as well as the information desk.

PROFESSIONAL TRAVEL SERVICES – Being experts in logistics handling we guarantee efficient organization of social aspects of your conference – visa support for the delegates, cultural and social program, hotel accommodation, and transportation.

EXPERIENCE AND QUALITY – Our managers have experience in managing dozens of congresses, they know how to organize an event on a step-by-step basis and how to cope with underlying potential problems in the process of organization. We work as a team with constant exchange of knowledge and experience. We work only with proven and most qualified services vendors – they know our needs and are flexible to deal with.

Monomax PCO is proud to be a member of ***International Congress & Convention Association (ICCA),*** the Netherlands, in MEETINGS MANAGEMENT category.

TRANSFERS

June 20

1. Departure: **8:00** from hotel Rossiya (pl. Chernyshevskogo 11), intermediate stop at Moskovskaya metro station (Moskovskiy pr, 189)
   Arrival: High School of Management SPbU “Mikhailovskaya Dacha”
2. Departure: **18:30** from High School of Management SPbU “Mikhailovskaya Dacha”
   Arrival: Restaurant Shuvalovka (Sankt-Peterburgskoye Schosse, 111)
3. Departure: **23:00** from Restaurant Shuvalovka (Sankt-Peterburgskoye Schosse, 111), intermediate stop at Moskovskaya metro station (Moskovskiy pr, 189)
   Arrival: hotel Rossiya (pl. Chernyshevskogo 11)

June 21

1. Departure: **8:30** from hotel Rossiya (pl. Chernyshevskogo 11), intermediate stop at Moskovskaya metro station (Moskovskiy pr, 189)
   Arrival: High School of Management SPbU “Mikhailovskaya Dacha”
2. Departure: **19:00** from High School of Management SPbU “Mikhailovskaya Dacha”, intermediate stop at Moskovskaya metro station (Moskovskiy pr, 189),
   Arrival: hotel Rossiya (pl. Chernyshevskogo 11)

June 22

1. Departure: **8:30** from hotel Rossiya (pl. Chernyshevskogo 11), intermediate stop at Moskovskaya metro station (Moskovskiy pr, 189)
   Arrival: High School of Management SPbU “Mikhailovskaya Dacha”
2. Departure: **18:15** from High School of Management SPbU “Mikhailovskaya Dacha”, intermediate stop at Moskovskaya metro station (Moskovskiy pr, 189),
   Arrival: hotel Rossiya (pl. Chernyshevskogo 11)

THURSDAY – JUNE 20, DAY 1 SCHEDULE

B – Break	I – Invited Talk	O – Opening or Closing Talk	F – Featured Talk
T – Talk	D – Dinner	P – Posters	

9:00AM–9:45AM	B	Registration
9:45AM–10:00AM	O	**Opening Ceremony** Alla Lapidus, *SPbU* Anton Korobeynikov, *SPbU*
10:00AM–11:00AM	I	***Direct queries of metagenome assembly graphs reveal hidden biological complexity*** C. Titus Brown *UC Davis*
11:00AM–11:20AM	T	**Synteny paths for assembly graphs comparison** Evgeny Polevikov *ITMO University*
11:20AM–11:40AM	T	**Indexing De Bruijn graphs with minimizers** Camille Marchet *Lille University*
11:40AM–12:00PM	T	**PathRacer: racing profile HMM paths on assembly graph** Alex Shlemov *Saint Petersburg State University*
12:00PM–12:30PM	B	**Break**
12:30PM–1:30PM	I	**Reproducible metagenomic at scale** Mick Watson *Roslyn Institute, University of Edinburgh*
1:30PM–1:50PM	T	**Recovery of closed genomes of member species from enrichment reactor microbial communities using long read metagenomics** Krithika Arumugam *Nanyang Technological University*
1:50PM–2:10PM	T	**Computational vision to detect chromosomal loops** Axel Cournac *Institut Pasteur*
2:10PM–2:30PM	T	**Whole-chromosome assembly and analysis of hybridogenetic lineages of the desert ant Cataglyphis hispanica with instaGRAAL, a Hi-C based scaffolder and polisher** Lyam Baudry *Institut Pasteur*

2:30PM–3:30PM	B	**Lunch**
3:30PM–4:30PM	F	**Biotech from theory to reality: the journey of creating and marketing groundbreaking technology** Andrey Perfilyev *Atlas*
4:30PM–4:50PM	T	**SPAligner: alignment of long diverged molecular sequences to assembly graphs** Dmitry Antipov *Saint Petersburg State University*
4:50PM–5:10PM	T	**Million sequences indexing** Antoine Limasset *CNRS*
5:10PM–5:30PM	T	**Hybrid RNA-Seq assembly and its application to transcriptome study of Cynara cardunculus** Andrey Prjibelski *Saint Petersburg State University*
5:30PM–5:50PM	T	**Combinatorial Scoring of Phylogenetic Trees and Networks Based on Homoplasy-Free Characters** Max Alekseyev *George Washington University*
6:10PM–9:00PM	D	**Meet & greet**

The conference was supported by the Russian Fund for Basic Research (grant 19-016-20002)

THURSDAY — JUNE 20

DAY 1 TALK SUMMARIES

Direct queries of metagenome assembly graphs reveal hidden biological complexity

C. Titus Brown (University of California, Davis, USA)

Metagenome assembly is extremely challenging because of the biological variability and community structure present in microbial communities. I will discuss some recent advances in graph theoretic approaches to indexing and querying large metagenome assembly graphs that reveal additional genic content and biological variability within metagenome-assembled genomes from the environment, and talk about some of our plans for further inquiry.

Synteny paths for assembly graphs comparison

Evgeny Polevikov (ITMO University, Saint Petersburg, Russia)
Mikhail Kolmogorov (Department of Computer Science and Engineering,
University of California, San Diego, USA)

Despite the recent developments of long-read sequencing technologies, *de novo* assemblies of relatively large genomes are often fragmented. The resulting genome fragments (contigs) are typically complemented by assembly graphs (commonly implemented as de Bruijn or overlap graphs), which define putative links between contigs. The connections in these graphs represent uncertainties in the assembled genome structure and can benefit to gene prediction, haplotype separation, structural variations analysis and other applications. To facilitate the further development of assembly graph-based approaches, it is important to establish algorithms for comparison and evaluation of various assembly graphs produced by different approaches. However, only a few tools for visual verification are currently available.

In this work we describe a set of algorithms for analysis and comparison of various assembly graphs produced by different assemblers. We introduce synteny paths: maximal paths of homologous sequence between the compared graphs. Similarly to synteny blocks (that are commonly used in comparative genomics studies), synteny paths highlight the structural similarities between the compared genomes but are robust to assembly fragmentation.

We define the problem of finding the minimum number of synteny paths between two assembly graphs and prove that the exact solution is NP-hard. We then propose an efficient approximate algorithm for this decomposition by transforming two assembly graphs into a colored breakpoint graph.

We use synteny paths to compare between Flye and Canu assemblies of 21 bacterial genomes from the NCTC collection. The final graphs produced by two assemblers from the same genome might not be identical, due to the different approaches to repeat resolution. However, if a graph is free from misassemblies, it should encode the original genome sequence. Indeed, in 14 out of 21 assemblies we found that each connected component in assembly graph was covered by a single path, revealing the putative genomic walks.

Finally, we apply synteny paths to compare between assemblies of 15 *Drosophila* genomes with extensive structural variations. We show that synteny paths decomposition reveals longer homologous segments, comparing to synteny blocks reconstructed from fragmented contigs. Interestingly, N50 of synteny paths was highly correlated with the evolutionary distance between the compared genomes. We illustrate the value of our approach for comparative genomics by performing phylogenetic tree reconstruction based on the recovered structural similarities between pairs of genomes.

Indexing De Bruijn graphs with minimizers

Camille Marchet (Lille University, France)
Maël Kerbiriou (Inria Lille, France)
Antoine Limasset (CNRS Lille, France)

The need to associate information to words is shared among a plethora of applications and methods in high throughput sequence analysis, and could be marked as fundamental. A scalability problem is promptly met when indexing billions of *k*-mers, as exact associative indexes can be memory expensive. To leverage this challenge, recent works take advantage of the *k*-mer sets properties. They exploit the overlaps shared among *k*-mers by using a De Bruijn graph as a compact *k*-mer set to provide lightweight structures.
Contribution: we propose a scalable and exact index structure able to associate unique identifiers to indexed *k*-mers and to reject alien *k*-mers. The proposed index combines an extremely compact representation along with a high throughput. Moreover, it can be efficiently built from the De Bruijn graph sequences. The efficient index implementation we provide, achieved to index the *k*-mers from the human genome with 8GB within 30 minutes and was able to scale up to the huge axolotl genome with 63GB within 10 hours. Furthermore, while being memory efficient, the index allows above a million queries per second on a single CPU in our experiments and its throughput can be raised using multiple cores. Finally, we also present the index ability to practically represent metagenomic and transcriptomic sequencing data to highlight its wide applicative range.
Availability: the index is implemented as a header-only library in C++, is open source under AGPL3 license and available at https://github.com/Malfoy/Blight. It was designed as a user-friendly library and comes along with sample code usage.

PathRacer: racing profile HMM paths on assembly graph

Alexander Shlemov (Center for Algorithmic Biotechnology, St. Petersburg University, Russia)
Anton Korobeynikov (Center for Algorithmic Biotechnology &
Department of Statistical Modelling, St. Petersburg University, Russia)

Recently large databases containing profile Hidden Markov Models (pHMMs) have emerged. These pHMMs may represent the sequences of antibiotic resistance genes or allelic variations amongst highly conserved housekeeping genes used for strain typing, etc. The typical application of such a database includes the alignment of contigs to pHMM hoping that the sequence of the gene of interest is located within the single contig. Such a condition is often violated for metagenomes preventing the effective use of such databases.
We present PathRacer — a novel standalone tool that aligns a profile HMM directly to the assembly graph (performing the codon translation on fly for amino acid pHMMs). The tool provides the set of most probable paths traversed by a HMM through the whole assembly graph, regardless whether the sequence of interest is encoded on the single contig or scattered across the set of edges, therefore significantly improving the recovery of sequences of interest even from fragmented metagenome assemblies. Compared to the analogues (Xander and MegaGTA) PathRacer can perform partial gene search and search for pseudogenes and gene sequences with frameshifts. That makes PathRacer appropriate for unpolished longread assembly annotation as well.

Reproducible metagenomic at scale

Mick Watson (Roslyn Institute, University of Edinburgh, GB)

We are in the midst of a metagenomic revolution, driven by affordable, high-quality sequence data, and advances in bioinformatics and computational power. Despite the fragmented assemblies produced, even short-read Illumina data can be efficiently and accurately binned into high quality genomes — termed metagenome-assembled genomes, or MAGs. Two recent studies released 154,000 and 92,000 MAGs from the human microbiome respectively, revealing the power of metagenomics to sequence novel microbial genomes. I will describe our work in ruminants where we have assembled over 19,000 MAGs of medium quality, and 4941 MAGs of high quality, from 288 cattle. I will discuss our results and demonstrate the novelty of our dataset, and go on to describe our reproducible pipeline for metagenomic assembly which we are adapting for use in the cloud, enabling serverless large-scale metagenomic assembly using only a web-browser. I will also discuss our use of MinION nanopore sequencing data, and how we have used long reads to assemble entire microbial chromosomes from metagenomics data.

Recovery of closed genomes of member species from enrichment reactor microbial communities using long read metagenomics

Krithika Arumugam (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore)
Irina Bessarab (Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore)
Caner Bağcı (Center for Bioinformatics, University of Tubingen, Germany)
Sina Beier (Center for Bioinformatics, University of Tubingen, Germany)
Benjamin Buchfink (Max-Planck-Institute for Developmental Biology, Tubingen, Germany)
Anna Górska (Center for Bioinformatics, University of Tubingen, Germany)
Mindia A. S. Haryono (Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore)
Guanglei Qiu (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore; School of Environment and Energy, South China University of Technology, Guangzhou, China)
Rogelio E. Zuniga Montanez (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore)
Stefan Wuertz (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore;
School of Civil and Environmental Engineering, Nanyang Technological University, Singapore)
Ying Yu Law (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore)
Federico M. Lauro (Singapore Centre for Environmental Life Sciences Engineering, Nanyang Technological University, Singapore;
Asian School of the Environment, Nanyang Technological University, Singapore)
Daniel H. Huson (Center for Bioinformatics, University of Tubingen, Germany; Life Sciences Institute, National University of Singapore)
Rohan B. H. Williams (Singapore Centre for Environmental Life Sciences Engineering, National University of Singapore)

Metagenome assembly is taking an increasingly central role in the analysis and interpretation of microbiomes and complex microbial communities, due to the ability of this approach to recover draft genomes of member species, thus providing a rigorous basis for studying the community composition and function. To date most metagenome assemblies have been undertaken using data from short read technologies, but this approach has rarely been able to generate closed genomes. New long read sequencing technologies offer huge potential for effective recovery of complete, closed genomes, and while much progress has been made on cultured isolates, the ability of these methods to recover genomes of member taxa in complex microbial communities is less clear. Here we examine the ability of long read data to permit recovery of genomes from activated sludge enrichment reactor metagenomes: as they offer a moderate level of complexity compared to their inoculum sourced from full scale wastewater treatment plants. We sampled a series of enrichment reactors designed to enrich for 1) anaerobic ammonium oxidising bacterium (AnAOB); 2)

proteobacterial and 3) non-proteobacterial polyphosphate accumulating organisms (PAO), extracting genomic DNA and obtaining both short read (Illumina 251bp or 301bp PE) and long read data (MinION Mk1B) from the same DNA aliquot. We generated long read data ranging from 398 MBp to 6.6 GBp from the MinION flowcells, with read length N50 ranging from 6,502 bp to 27,068 bp. Reactor communities varied in complexity from 25 – 100 species as defined by short read MAG analysis. We demonstrate that whole bacterial chromosomes can be obtained from whole community long read data (1 – 8 genomes/community, dependent on read depth, community structure and ecogenomic complexity). We provide a straightforward pipeline for processing such data, which includes a new approach to correcting erroneous frame-shifts, as well as descriptive statistical methods for screening associations between short read MAGs and long read chromosome length assembled sequences, in order to identify cognate genomes from both analyses. We are currently examining the impact of ecogenomic complexity (strain level diversity) on genome recoverability from long read data. We conclude that long read metagenomics on medium complexity microbial communities is feasible and can recover closed, complete genomes of the most abundant community members.

Computational vision to detect chromosomal loops

Axel Cournac (Institut Pasteur, France)
Lyam Baudry (Institut Pasteur, France)
Rémi Montagne (Institut Pasteur, France)
Cyril Matthey-doret (Institut Pasteur, France)
Axel Breuer (ENGIE Global Markets SAS)
Romain Koszul (Institut Pasteur, France)

The precise architecture of chromosomes underlies or influences major biological functions such as replication, segregation or transcriptional regulation. In parallel to microscopy, the recent development of contact technologies such as 3C, Hi-C allows the access to new spatial resolutions (~kilobase). The output of such experiments are heatmaps that give the frequency of physical contact between different loci inside a genome.

A computational challenge is to detect chromosomal loops events in such data. We propose a novel algorithm called ChromoVision inspired from computational vision approaches that allows the rapid detection of loops. The algorithm is based on pattern recognition and can be extended to various generic patterns found in chromosomal contact maps like borders, bow patterns. We assess its accuracy using both simulated data and experimental data from various biological conditions. We show biological applications in bacteria, yeast and human data sets confirming the central role of cohesin and SMC proteins in the formation of long range genomic loops contributing to the spatial organisation of chromosomes.

Whole-chromosome assembly and analysis of hybridogenetic lineages of the desert ant *Cataglyphis hispanica* with instaGRAAL, a Hi-C based scaffolder and polisher

Lyam Baudry (Institut Pasteur, France)
Hugo Darras (Université Libre de Bruxelles, Belgium)
Martial Marbouty (Institut Pasteur, France)
Jean-François Flot (Université Libre de Bruxelles, Belgium)
Serge Aron (Université Libre de Bruxelles, Belgium)
Romain Koszul (Institut Pasteur, France)

Assembling complete chromosomes of large genomes is a technical challenge presenting a number of long-standing issues, such as the bridging of gaps that can be found in draft genomes or the presence of repeated sequences that conventional assemblers have trouble resolving. As such, many large arthropod genomes are in an unfinished state, comprising many more scaffolds than the expected number of chromosomes. Here, we present instaGRAAL, a fast, open-source program that uses chromosome conformation capture (Hi-C) data to scaffold contigs based on the collision frequencies between DNA sequences in the nucleus.

InstaGRAAL builds upon and improves our formerly published program GRAAL, which uses a simple polymer model to represent the expected spatial contacts between these sequences and a Markov Chain Monte Carlo (MCMC) method to maximize the likelihood of this model (Marie-Nelly et al., 2014). When applied to the genomes of two hybridogenetic lineages of the Spanish desert ant *Cataglyphis hispanica,* instaGRAAL yielded completely assembled chromosomes and revealed large-scale structural differences that may account for their unusual reproductive strategies.

Adapting bioinformatics to bacteriophage genomics and virome studies

Marie-Agnès Petit (Micalis Institute, INRA, France)

The genetics and genomics of bacteriophages differs markedly from that of their bacterial host. Therefore, the tools adapted for bacterial genomics and metagenomics are often not adapted for bacteriophage studies. Indeed, during its replication cycle, the phage produces up to hundreds of copies of its genome in a very short period of time. This results in (1) many more replication errors, compared to bacteria, so the mutation rate is high, and genomic divergence is increased, compared to bacterial genomes. (2) many more recombination events as well, whereby phages exchange genetic material, so that different phage genomes share regions with high identities, the so-called "mosaics", which comes down to a problem similar to repeats, in terms of genome assembly.

As a consequence of the first point, homology searches with tools as BLAST or psi-BLAST do not permit annotation transfer, and most phage genomes have up to 80% of genes with unknown function. I will present the PHROGs study, a successful effort to aggregate distant protein families with HHsearch (distant homology search based on pairwise alignment profile comparisons) combined with expert curation to transfer annotation, allowing progressively to overcome this first roadblock.

As a consequence of the second point, phage genome assemblies from viral metagenomes are challenging. And in most cases, the "true assembly" solution is also missing. Starting from 650 fecal viromes, we performed a parallel study of phage culture+sequencing, and shotgun sequencing+assembly with meta-SPAdes. This allowed us to observe a reasonable match between in silico-assembled genomes and their 'true solution', except for the most mosaic phages.

SPAligner: alignment of long diverged molecular sequences to assembly graphs

Tatiana Dvorkina (Center for Algorithmic Biotechnology, St. Petersburg University, Russia)
Dmitry Antipov (Center for Algorithmic Biotechnology, St. Petersburg University, Russia)
Anton Korobeynikov (Center for Algorithmic Biotechnology &
Department of Statistical Modelling, St. Petersburg University, Russia)
Sergey Nurk (Center for Algorithmic Biotechnology, St. Petersburg State University, Russia)

Graph representation of genome assemblies has been recently used in different applications — from gene finding to haplotype separation. While many of these applications are based on aligning DNA and amino acid sequences to assembly graphs, existing software tools for finding such alignments have important limitations. We present a novel SPAligner (Saint Petersburg Aligner) tool for aligning long diverged molecular sequences to assembly graphs and demonstrate that it generates accurate alignments.

Million sequences indexing

Antoine Limasset (CNRS, France)

Most methodological contributions handling sequencing data now acknowledge the need to scale up to the terrific throughput that we face nowadays. Since BLAST, a plethora of tools have been developed to handle the massive amount of available reference sequences. Recently, new structures have been proposed to link a short sequence such as a transcript or a gene to sequencing datasets or reference genomes. The challenge of such structures is to be able to index hundreds of thousands datasets with a reasonable amount of memory, while being able to perform fast query. A rich state of the art quickly emerged, SBT SSBT, HowDeSBT, BIGSI, ... based on different combinations of bloom filters in order to link a *k*-mer to its associated datasets. While extremely efficient, as an example BIGSI was able to index half a million bacterial genomes with 1.5TB, those techniques are not able to scale to all known genomes or all transcriptome collections.

We aim to propose a new data structure that could use an order of magnitude less resources than BIGSI while being able to perform similar queries in terms of accuracy and throughput.
Instead of indexing all *k*-mers of a dataset, we choose to rely on local sensitive hashing methods to index a small subset of the input *k*-mers. This choice, allow the scaling of the methods with a comparable accuracy than the know strategies on medium sized queries (1kb or larger).
Furthermore we propose a matrix structure somewhat similar to BIGSI, conserving extremely important properties for such an index:

- Constant time insertion of a new reference sequence by adding a new row to the matrix
- Queries rely on reading columns that can be compressed column for lighter structure and faster queries
- Easily parameterizable structure where memory/accuracy trade-off can be precisely chosen

In this presentation we show the design of such a structure using the Min-Hash scheme. We present preliminary results on hundred thousand bacterial genomes on a proof of concept implementation. We compare our performances to BIGSI and Mashscreen, showing that our proposed structure can achieve a comparable accuracy with a better scaling in memory or in throughput. We finally discuss the improvements that are currently developed and what can be expected from this scheme, and its potential applications in mega-scale sequences indexing, clustering or genome assembly.

Hybrid RNA-Seq assembly and its application to transcriptome study of *Cynara cardunculus*

Andrey Prjibelski (Center for Algorithmic Biotechnology, St. Petersburg University, Russia)
Guiseppe Puglia (Consiglio Nazionale delle Ricerche,
Istituto per i Sistemi Agricoli e Forestali del Mediterraneo, Italy)
Elena Bushmanova (Center for Algorithmic Biotechnology, St. Petersburg University, Russia)
Domenico Viatale (Consiglio Nazionale delle Ricerche,
Istituto per i Sistemi Agricoli e Forestali del Mediterraneo, Italy)

De novo RNA-Seq assembly is a powerful method for analysing transcriptomes when the reference genome is not available or poorly annotated. However, due to the short length of Illumina reads it is often impossible to reconstruct complete sequences of complex alternative isoforms. Recently emerged possibility to generate long RNA reads, such as PacBio and Oxford Nanopores, may dramatically improve the assembly quality and thus the consecutive analysis. While reference-based analysis was already performed using these long reads [1], no publicly-available computational method for *de novo* assembly currently exist.

In this work we present a novel algorithm that allows to perform high-quality *de novo* transcriptome assembly by combining accuracy and reliability of short reads with exon coordination information from long error-prone reads. The algorithm is designed by adapting existing hybridSPAdes [2] approach for transcriptomic data implementing it within rnaSPAdes [3] pipeline. As an additional feature, most of long-read technologies allow to derive a full-length mRNA sequences in a read based on the terminal adapters. The developed algorithm is capable of taking benefits from such king of reads to accurately determine full isoform sequences as well as their UTRs.

Additionally, we apply the designed algorithm to sequencing data obtained from various tissues at different growth phases of *Cynara cardunculus* artichoke — an important agricultural plant. We show significant improvement of hybrid *de novo* assembly in comparison to Illumina-only based contigs. Further, we use the assembly to enhance the existing *Cynara cardunculus* gene database and perform differential expression analysis and functional annotation of various gene families across multiple samples.

References:

[1] Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., ... & Jordan, M. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods,* 15(3), 201.
[2] Antipov, D., Korobeynikov, A., McLean, J.S. and Pevzner, P.A., 2015. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics,* 32(7), pp.1009–1015.
[3] Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D., 2018. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. *bioRxiv,* p.048942.

Combinatorial scoring of phylogenetic trees and networks based on homoplasy-free characters

Nikita Alexeev (ITMO University, Saint Petersburg, Russia)
Max Alekseyev (George Washington University, Washington, DC, USA)

Construction of phylogenetic trees and networks for extant species from their characters represents one of the key problems in phylogenomics. While solution to this problem is not always uniquely defined and there exist multiple methods for tree/network construction, it becomes important to measure how well the constructed networks capture the given character relationship across the species.

We propose a novel method for measuring the specificity of a given phylogenetic network in terms of the total number of distributions of homoplasy-free character states at the leaves that the network may impose. While for binary phylogenetic trees, this number has an exact formula and depends only on the number of leaves and character states but not on the tree topology, the situation is much more complicated for non-binary trees or networks. Nevertheless, we develop an algorithm for combinatorial enumeration of such distributions, which is applicable for arbitrary trees and networks under some reasonable assumptions. We further extend our algorithm to a special class of characters that follow Dollo's law of irreversibility.

FRIDAY – JUNE 21, DAY 2 SCHEDULE

B – Break	I – Invited Talk	O – Opening or Closing Talk	F – Featured Talk
T – Talk	D – Dinner	P – Posters	

10:00AM–11:00AM	I	***Metabolic adaptation in the human gut microbiota during pregnancy and infancy*** M. Pilar Francino *FISABIO*
11:00AM–11:20AM	T	**Higher-order and multidimensional epistasis in high-throughput experimental data** Dmitry Ivankov *Skolkovo Institute of Science and Technology*
11:30PM–12:00PM	B	**Break**
12:00PM–1:00PM	I	**Insights into the human gut microbiota from a (meta-)genomic perspective** Rob Finn *EMBL-EBI*
1:00PM–1:20PM	T	**DYSBIOSIS SIGNATURES FROM METAGENOMES REVEAL DIFFERENCES AND SIMILARITIES BETWEEN INFLAMMATORY BOWEL DISEASES** Vadim Dubinsky *Tel-Aviv University*
1:20PM–1:40PM	T	**On the Verge of Colistin Resistance: Genetic Determinants Mediating Intermediate Colistin Resistance in Klebsiella pneumoniae** Sima Tokajian *Lebanese American University*
1:40PM–2:00PM	T	**Antimicrobial Resistance and Clonality of Streptococcus pneumoniae Isolates in Russia** Irina Tsvetkova *Pediatric Research and Clinical Center for Infectious Diseases*
2:00PM–3:30PM	B	**Lunch**
3:30PM–4:30PM	I	**Factors Influencing Soil and Air Microbiomes** Rob Knight *UC San Diego*
4:30PM–4:50PM	T	**Computational biology and agriculture: soil microbiome** Evgeny Andronov *FSBEI of Higher Education SPSU*

4:50PM–5:10PM	T	**Revealing mechanisms of the Bacillus thuringiensis host specificity via modeling coevolution of Cry toxins and their receptors** Yury V. Malovichko *All-Russia Research Institute for Agricultural Microbiology (ARRIAM)*
5:10PM–5:30PM	T	**Genomic determinants underlying rodenticidal properties of the Salmonella enteritidis var. Issatschenko** Kirill Antonets *All-Russia Research Institute for Agricultural Microbiology (ARRIAM)*
5:30PM–5:50PM	T	**Bacteriophage recombination site helps to reveal genes potentially acquired through horizontal gene transfer** Maria A. Daugavet *Institute of Cytology, Russian Academy of Sciences*
6:00PM–7:00PM	P	**Poster Section**

The conference was supported by the Russian Fund for Basic Research (grant 19-016-20002)

FRIDAY — JUNE 21

DAY 2 TALK SUMMARIES

Metabolic adaptation in the human gut microbiota during pregnancy and infancy

M. Pilar Francino (Public Health Valencian Region Foundation for the Promotion of Health and Biomedical Research (FISABIO), Valencia, Spain)

The human gut microbiota develops through a complex succession that takes place mostly during infancy, and has a major impact on life-long health through early interactions with metabolism and immunity. A thorough understanding of all aspects of gut microbiota development will be necessary for engineering strategies that can modulate this process, which is currently being challenged by numerous factors associated to the contemporary lifestyle. To this aim, we need to unravel, not only the changes in taxonomic composition, but also the development of functional capacities throughout the microbial succession process occurring in the gut. Metagenomic and metatranscriptomic studies have started to produce insights into the trends of functional repertoire and gene expression change within the first year after birth. These studies have shown that a directional pattern of change towards the adult state can be observed during infancy in terms of taxonomic composition, gene abundance and gene expression, although a large variability among individuals exists at each of these levels. Moreover, temporal trends of gene expression do not always parallel those of gene abundance since the relative expression of many genes changes through infancy, indicating a stage-specific adaptation of gut microbiota activity. Hallmarks of aerobic metabolism disappear from the microbial metatranscriptome as development proceeds, while the expression of functions related to carbohydrate transport and metabolism increases and diversifies, approaching that observed in adults. Butyrate synthesis enzymes are overexpressed at three months of age, even though most butyrate-producing organisms are still rare. This suggests that butyrate production may be ensured in the gut of young infants before the typical butyrate synthesizers of the adult gut become abundant. Significant differences remain at the end of infancy with respect to the taxonomic composition, gene repertoire and gene expression patterns seen in adults, indicating that more research needs to focus on further changes occurring during childhood. Metatranscriptomic approaches also reveal how the gut microbiota adapts its function to the physiological changes that occur in late pregnancy. During this period, the microbiota readjusts the expression of carbohydrate-related functions in a manner consistent with a high availability of glucose, suggesting that it may be able to access the high levels of blood glucose characteristic of this period. Moreover, late pregnancy gut bacteria may reach stationary phase, which may affect their likelihood of translocating across the intestinal epithelium.

Higher-order and multidimensional epistasis in high-throughput experimental data

Laura Avinyo Esteban (Universitat Pompeu Fabra, Barcelona, Spain)
Natalia Bogatyreva (Institute of Protein Research, Russian Academy of Sciences)
Fyodor Kondrashov (Institute of Science and Technology, Austria)
Dmitry Ivankov (Skolkovo Institute of Science and Technology, Moskow, Russia)

Epistasis, a non-additive influence of amino acid substitutions on the phenotype, is one of the crucial factors of evolution. It is epistasis that prevents us from the prediction of phenotype from genotype, a Holy Grail of evolutionary biology. That is why it is vital to study epistasis, to touch the limits of predictability.

The information about epistasis comes from high-throughput evolutionary experiments, where the phenotype is measured for hundreds of thousand genotypes. To calculate epistatic coefficients of the N-th order one has to consider the corresponding N-dimensional hypercube.

We have recently developed an effective algorithm finding all N-dimensional hypercubes in the experimental data. When applied to the results of HIS3 protein containing data for more than 700,000 genotypes, the algorithm found about 200,000,000 hypercubes.

The next step is to distinguish unidimensional epistasis from multidimensional one. It is essential because a unidimensional case converts into a non-epistatic case due to the existence of a non-linear monotonic transformation from the phenotype to a function linear on the substitution effects. The well-known examples of multidimensional epistasis are the sign and the reciprocal sign epistasis. In our research, we have found conceptually new cases of multidimensional epistasis, which are more hidden and complicated. We have discovered more than 20,000 of such cases in experimental data on green fluorescence protein.

The brute-force implementations of the given algorithms have complexities of the order from four to eight on the amount of input data. However, we took into account the peculiarities of the problem at hand and managed to reduce the time significantly.

Insights into the human gut microbiota from a (meta-)genomic perspective

Rob Finn (European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI))

Metagenomics, the analysis of the sum of genetic material from a sample, has started to shed light on the huge diversity of microorganisms that occupy environments such as the human body, soil and the world's oceans. The assembly of metagenomics shotgun sequencing reads into contigs enables the identification of functional operons and full length protein sequences, the majority of which have never been seen before. Ultimately, such sequences (DNA and protein) need to be linked to a taxonomic identifier. But how much can we trust *de novo* assembly and binning techniques to provide draft genomes? Partly to answer this question we have undertaken a large-scale analysis of public human gut microbiome datasets, performing *de novo* assembly on thousands of samples containing adequate read depth. We have compared computed genomics bins (metagenome assembled genomes, MAGs) with isolate genomes and used the latter to assess the quality of the MAGs, the tools for assessing MAG quality and provide an estimate of how many species are yet to be represented in genomic reference catalogues for studying the human microbiome. With these new MAGs, we can start truly start understanding the composition of the human gut microbiome, and the species variability. As we build a more complete picture of the human microbiome constituents, we can start developing and testing better mechanistic models.

Dysbiosis signatures from metagenomes reveal differences and similarities between inflammatory bowel diseases

Vadim Dubinsky (School of Molecular Cell Biology and Biotechnology,
George S. Wise Faculty of Life Sciences, Tel-Aviv University, Israel)
Reshef Leah (School of Molecular Cell Biology and Biotechnology,
George S. Wise Faculty of Life Sciences, Tel-Aviv University, Israel)
Keren Rabinowitz (The Division of Gastroenterology & Felsenstein Medical Research Center,
Rabin Medical Center, Petah Tikva, Israel)
Karin Yadgar (The Division of Gastroenterology & Felsenstein Medical Research Center,
Rabin Medical Center, Petah Tikva, Israel)
Lihi Godny (The Division of Gastroenterology & Felsenstein Medical Research Center,
Rabin Medical Center, Petah Tikva, Israel)
Iris Dotan (The Division of Gastroenterology, Rabin Medical Center, Petah-Tikva, Israel;
Sackler Faculty of Medicine, Tel-Aviv University, Israel)
Uri Gophna (School of Molecular Cell Biology and Biotechnology,
George S. Wise Faculty of Life Sciences, Tel-Aviv University, Israel)

Inflammatory bowel diseases (IBD) are characterized by chronic inflammation of the gastrointestinal tract, and their etiology involves an aberrant immune response to imbalanced gut microbiome in genetically susceptible individuals. The principle types of IBD are Crohn’s disease (CD), ulcerative colitis (UC) and pouchitis. CD may affect any part of the gastrointestinal tract while UC is limited to the large intestine. Approximately 25% of patients with complicated UC may undergo total large bowel resection followed by creation of a reservoir (“pouch”) from the normal small bowel in order to restore intestinal continuity. Approximately 60% of former UC patients may develop inflammation of the previously normal small bowel comprising the pouch (pouchitis). Because the inflammatory process in pouch patients develops in a previously normal small intestine of UC patients, we have previously demonstrated that pouchitis may be a model for the development of intestinal inflammation with features resembling CD based on serologic markers, mucosal microRNA expression and bacterial *16S rRNA* sequencing.

The importance of microbial factors in IBD has been suggested by multiple clinical observations and experimental models. Gut microbiome of IBD patients was shown to be less diverse, with lower levels of beneficial species and higher levels of proinflammatory species compared to healthy individuals (“dysbiosis”).

In this study we aimed to compare the microbiome and its functions across different types of IBD. We applied shotgun metagenomics to faecal samples obtained from 78 patients with pouchitis. In addition we obtained metagenomic data from two independent cohorts, PRISM and LifeLines DEEP-NLIBD (n = 88 CD, 76 UC and 56 healthy-controls).

PCoA analysis based on bacterial species and enzymes profiles showed substantial overlap between pouch patients and CD patients. The first axes of taxonomic and functional variation correlated (Spearman $r = 0.5$, $P < 0.05$) with severity of inflammation *(faecal calprotectin)* and samples were organized along that axis according to their IBD phenotype, from healthy-controls to UC, CD and pouch (highest inflammation). Gradient boosting trees classifiers were built based on taxonomic and enzymes profiles to classify patients based on their IBD phenotype. Micro-average

area under the curve of 0.9 was achieved for the three IBD phenotypes and was the highest for pouch patients. Most of the misclassified pouch patients were labelled as CD. *Escherichia coli* and several species of *Streptococcus* and *Veillonella* were significantly more abundant in all IBD types compared to healthy-controls, with pouch patients presenting the highest levels. The opposite trend was observed for beneficial bacteria such as *Eubacterium rectale, Roseburia inulinivorans, Faecalibacterium prausnitzii, Ruminococcus bromii* and Bacteroides species. Microbial enzymes related to protection against oxidative stresses in IBD inflamed gut (glutathione-disulfide reductase, peroxiredoxin and nitric-oxide dioxygenase) were enriched in pouch patients. Moreover, numerous metabolic pathways were associated with specific IBD phenotypes. Aromatic amino-acid biosynthesis and starch degradation pathways were enriched in healthy-controls and were the lowest in pouch.

Our findings indicate that although robust bacterial taxonomic and functional overlaps are found between pouchitis and CD, pouchitis harbours a distinct signature characterized by intensified dysbiosis.

On the verge of colistin resistance: genetic determinants mediating intermediate colistin resistance in *Klebsiella pneumoniae*

Sahar Alousi (Lebanese American University, Lebanon)
Tamara Salloum (Lebanese American University, Lebanon)
Balig Panoosian (Lebanese American University, Lebanon)
Harout Arabaghian (Lebanese American University, Lebanon)
Rony Khnayzer (Lebanese American University, Lebanon)
George Araj (American University of Beirut, Lebanon)
Sima Tokajian (Lebanese American University, Lebanon)

Colistin is one of the last resort antibiotics used to treat infections by carbapenemase-producing *Klebsiella pneumoniae* (CPKP). Insertional inactivation of *mgrB*, a gene encoding a negative regulator of the *PhoPQ* two-component system (TCS), and *crrAB* (a sensory TCS) have recently gained attention as mediators of colistin resistance. In this study, broth microdilution colistin susceptibility testing and whole-genome sequencing were used to resolve phenotypic and genotypic resistance profiles in 11 clinical carbapenem- and colistin- resistant *K. pneumoniae* (KP). Whole-genome sequencing (WGS) was performed using short-paired end reads technology on an Illumina Miseq. Core genome single nucleotide polymorphisms (cg-SNP) were called by the Snippy pipeline, and recombination events were highlighted using Gubbins. The pan-genome was generated using Roary. Chromosomally encoded genes were also screened for synonymous and non-synonymous mutations, in particular, *pmrAB, pmrD, pmrC,* and *phoPQ.* The genetic environment of *mgrB* was manually validated by Sanger sequencing. Lipid A was extracted using mild acetic acid hydrolysis and profiled using MALDI-TOF MS to examine noteworthy modifications linked to decreased susceptibility to colistin. The lipid A major mass ion was observed at (m/z 1840) in all KP isolates. PCR amplification of *mgrB* revealed insertional inactivation Δ*mgrB* in three of the studied isolates (designated as KP5, KP6, and KP16) showing MICs ≥16 mg/L. *ISKpn14* was associated with KP5 and KP6, while IS903 was detected in KP16. Wildtype *mgrB* gene in the remaining 8 isolates might suggest the involvement of other mechanisms underlying their nonsusceptibility to colistin. Recombination analysis highlighted genomic loci involved in both toxin-antitoxin and MFS efflux systems as favored hotspots for recombination. All 11 isolates were negative for the *crrAB* genes. Further biochemical and molecular analysis is in progress to characterize genetic determinants that play key roles in colistin resistance.

Along with the escalating prevalence of CRKP and the lack of novel antibiotics, colistin resistance has imposed a worldwide concern. With the power of WGS and lipidomic approaches, genetic alterations in pathways responsible for lipid A modification can be detected with high precision, enabling us to better understand the molecular mechanisms involved in resistance.

Antimicrobial resistance and clonality of *Streptococcus pneumoniae* isolates in Russia

Irina Tsvetkova (Department of Medical Microbiology and Molecular Epidemiology, Pediatric Research and Clinical Center for Infectious Diseases, St. Petersburg, Russia)
Sergey Belanov (Institute of Biotechnology, University of Helsinki, Finland)
Vladimir Gostev (Department of Medical Microbiology and Molecular Epidemiology, Pediatric Research and Clinical Center for Infectious Diseases, St. Petersburg, Russia)
Ekaterina Nikitina (Department of Medical Microbiology and Molecular Epidemiology, Pediatric Research and Clinical Center for Infectious Diseases, St. Petersburg, Russia)
Olga Kalinogorskaya (Department of Medical Microbiology and Molecular Epidemiology, Pediatric Research and Clinical Center for Infectious Diseases, St. Petersburg, Russia)
Marina Volkova (Department of Medical Microbiology and Molecular Epidemiology, Pediatric Research and Clinical Center for Infectious Diseases, St. Petersburg, Russia)
Sergey Sidorenko (Department of Medical Microbiology and Molecular Epidemiology, Pediatric Research and Clinical Center for Infectious Diseases, St. Petersburg, Russia)

Introduction: Monitoring of the dynamics of pneumococcal population is necessary to evaluate the effectiveness of vaccination. The population response to selective vaccination pressure can be both increasing the rate of pneumococcal infection diseases with replacement serotypes and appearance of antibiotic resistance in non-vaccine strains.
Objective: To clarify the intraspecies phylogenetic relationship between *S. pneumoniae* isolates from Russia and widespread epidemically significant genetic lines of pneumococci.
Methods: Concatenated MLST alleles analysis. This study included *S. pneumoniae* strains, registered in the PubMLST database: 516 strains from different regions of Russia (collected from 1980 to 2017); 431 referent strains (both the same rare sequence types and belonging to worldwide distributed clones); 6 referent strains, collected in the earliest period (1939, 1941, 1952, 1968, 1972). This dataset also included the pneumococcal strains, collected in our research center. Single nucleotide polymorphisms (SNPs) were extracted from concatenates of sequences of MLST alleles and used to generate a phylogeny using RaxML (v8.2.12), with the GTRGAMMA model. Annotation of the tree with various types of additional data (geographical information, years of the strains isolation and antibiotic susceptibility data) was made using iTOL (v4.3.2). The same SNPs alignment was analysed with RhierBAPS (R v.3.5.2, RhierBAPS package v.1.1.0) for analysis the associations of the major genetic clusters and the substructure within them. SplitsTree (v.4.14.8) was used for generation of dendrograms by neighbor joining (NJ) method and performing of split decomposition (SD) with bootstrap analysis (100 replications). In NJ analyses, closely related *Streptococcus mitis* strain was used as an outgroup.
Whole genome data analysis: Raw reads, available for 458 pneumococci strains, were downloaded from European Nucleotide Archive (ENA) and assembled *de novo* using SPAdes (v.3.13.0). Final quality assembly check was performed using Quast (v.5.0.1). Contaminated reads were identified using MLST (v.2.11) and excluded from the assay. The draft genome sequence was annotated using Prokka (v.1.13.3) and Genome Annotation Service in PATRIC with RAST tool kit (RASTtk). Roary (v. 3.12.0) was used to construct a pan-genome from the annotated assemblies.

Results: Totally, the 1058 pneumococci strains were analyzed in association with metadata. According to the phylogenetic trees, based on the concatenates of sequences of MLST alleles and constructed with RaxML and SplitsTree, pneumococci were divided into two Global Groups. As the result of RhierBAPS clustering, 13 major sequence clusters (SCs) were defined within our dataset. Two Global Groups were composed by the distinctive SCs and representatives of the three main SCs were distributed within the top level nodes of the trees. The multidrug-resistant isolates were identified in both groups. Beta-lactam resistance exhibited a tendency to be clonal. Pan-genome analysis allowed to estimate the genomic diversity for two Global Groups of pneumococci.
Conclusion: Understanding the population genetics of pneumococci will allow detection the changes in the prevalence of circulating genotypes. Until recently, a global assessment of pneumococcus genotypes had not been conducted in Russia. We can consider the results of this analysis as a picture of the population “before the start of vaccination”.

Factors influencing soil and air microbiomes

Rob Knight (University of California, San Diego, USA)

Microbes drive crucial processes ranging from the global carbon cycle to our own nutrition, yet until recently the microbial world was largely invisible and unknown. Today DNA sequencing is literally millions of times cheaper than a few years ago: this revolution in data acquisition sets the stage for a fundamental shift in our perspective on the role of microbes in our planet. Over the past few years, new tools for interpreting and visualizing microbial community data have enabled surprising discoveries about the microbial world at many scales including factors driving microbial communities in air, soil and ocean ecosystems. Large-scale projects such as the Earth Microbiome Project, together with technologies such as cloud computing, now allow access to microbial sequencing to communities of researchers across disciplines and across the planet, integrating insights across microbial systems and scales, and offer compelling prospects for restoration ecology.

Computational biology and agriculture: soil microbiome

Evgeny Andronov (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)

Over the past one or two decades, the importance of natural microbiomes in agriculture has been radically accentuated. Genetic, genomic and metagenomic approaches have transformed and reactivated the area of agricultural microbiology which has been quietly developing in the past 100 years. For the most part, research on microbiomes in the field of agricultural applications is still far from practical implementation. However, the outlines of new approaches in agriculture, actively using the achievements of genetics, genomics and metagenomics, are already beginning to emerge. In this report, we will present some applications that we believe have a prospect of practical application in real agriculture. We begin with the relatively simple task of identifying a wide range of pathogens using metagenomic technologies, which demonstrated that this seemingly simple task opens not only spectacular perspectives, but also hides complex computational problems. Then we will discuss the prospects of metagenomic research in agriculture, where the integration of metagenomic approaches with the classical theory of soil genesis led to the formation of a new computationally demanding area of soil biology. Then there will be presented studies of the evolution of agriculturally significant genes and genomes that demonstrate “natural engineering” ensuring a high efficiency of plant-microbe interaction. As a result, we will show that the problems of agriculture not only constitute a very attractive platform for the application of the skills and competencies of specialists in the field of computational biology, but also can open for them a diverse and fascinating world of agriculture biology.

The work is supported by RSF grant 18-16-00073

Revealing mechanisms of the *Bacillus thuringiensis* host specificity via modeling coevolution of *Cry* toxins and their receptors

Yury V. Malovichko (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia; St. Petersburg University, Russia)
Anton E. Shikov (St. Petersburg University, Russia)
Anton A. Nizhnikov (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia;, St. Petersburg University, Russia)
Kirill S. Antonets (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia;, St. Petersburg University, Russia)

Proteins possessing cytotoxic properties and commonly referred to simply as toxins comprise a vast group of bacterial virulence factors. For instance, *Bacillus thuringiensis*, a spore-forming bacterial pathogen of insects and several other invertebrate taxa, produces at least four major classes of proteinaceous toxins. Of these, crystalline pore-forming toxins produced at sporulation stage, known as *Cry* toxins, pose a particular interest because of the wide range of host species they affect. *Cry* toxins consist of three domains flanked with unstructured terminal sequences, with the N-terminal domain participating in pore formation and the other two involved in binding to a respective host's receptor. To date, more than 800 *Cry* toxins affecting species of four Insecta orders as well as those of *Nematoda phylum* have been discovered, and several membrane proteins, including cadherins and aminopeptidases, have been proposed to serve as Cry receptors; however, little is known about the molecular mechanisms underlying both mode of action and specificity of these toxins. In this work, we analyze coevolutionary substitution patterns in toxins and alanyl aminopeptidase receptors to reveal molecular mechanisms of the specificity of toxin-receptor interactions.. To enlarge the dataset of the toxins we developed a novel HMM-based tool for searching *Cry* toxins and annotating their domain structure, which outperforms its analogs. Launching this tool with all Bacillus-related sequences we discovered 340 new toxins putative novel groups within *Cry* superfamily as well as define domain boundaries within all present toxin sequences. Next, amino acid substitutions distinguishing protein sequences in both toxin and receptor epitope subsets were mapped onto minimum distance transition graphs which were then aligned to expose similarities in their topology. To prove surmise based on graph alignment evidence, we used aligned protein sequences for building maximum likelihood phylogenetic trees, which revealed a stable coevolutionary pattern between toxins and respective receptors. Subsequent in silico prediction of the effects of the mutations in these positions revealed sites determining host specificity of the toxins, which might be used to model new *Cry* toxins against certain receptors.

This work was supported by the Russian Science Foundation (Grant No 18-76-00028).

Genomic determinants underlying rodenticidal properties of the *Salmonella enteritidis var. Issatschenko*

Kirill Antonets (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia; St. Petersburg University, Russia)
Galina Minina (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Elena Bologova (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Anton Nizhnikov (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia; St. Petersburg University, Russia; St. Petersburg Scientific Center of the Russian Academy of Sciences, Russia)

Rodenticidal variant of the *Salmonella enteritidis* bacterium called *var. Issatschenko* after the eminent Russian microbiologist, academician Boris Issatschenko, was discovered in the beginning of the 20th century. This group of strains of *S. enteritidis* exhibits highly specific rodenticidal properties killing mice and rats and being safe for the most of other orders of mammals as well as humans. Biologicals based on *S. enteritidis var. Issatschenko,* are commonly used for protection of plants and food products from rodents. Nevertheless, molecular basis underlying host specificity of *S. enteritidis var. Issatschenko* remains to be mysterious. In this work, we for the first time made the complete genome assembly for the bacterium Issatschenko based on both the Nanopore long-read and Illumina short-read sequencing technologies. Comparison of the *S. enteritidis var. Issatschenko* genome with other publicly available assemblies of *S. enteritidis* revealed a set of unique single nucleotide polymorphisms in protein-encoding genes of the bacterium Issatschenko. Part of these sites is presented in coding regions of genes whose protein products are known to be involved in the virulence of *Salmonella*. Taking together, our data suggest that highly specific rodenticidal properties of the *Salmonella enteritidis var. Issatschenko* and its safety for humans are likely to be associated with mutations in several genes encoding virulence factors of this bacterium.

Bacteriophage recombination site helps to reveal genes potentially acquired through horizontal gene transfer

Daugavet M.A.
Shabelnikov S.V.
Adonin L.S.
Podgornaya O.I.
(Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia)

The cellulose-synthase gene of ascidians was gained from prokaryote donor and this is the most reliable example of horizontal gene transfer (HGT). In our previous study a new protein, rusticalin, of ascidian Styela rustica was described. Its C terminal domain coding region was also shown to be inherited from prokaryotic ancestor by means of HGT. Both for rusticalin C terminal domain and for cellulose-synthase catalytic domain it was shown that there coding regions neighbored with bacteriophage recombination site *AttP*. Thus we suggested a possible mechanism of HGT by means of bacteriophage insertion. Most of the cases of HGT are described based on sequence similarity alone, but in case of rusticalin we also demonstrated strong evidence of the mechanism of transfer by identifying the recombination site. It is possible that bacteriophage recombination site can help finding yet other new cases of HGT in eukaryotic genomes. Unfortunately the length of bacteriophage recombination site *AttP* is 43 nucleotides which is too short to find it reliably in big databases. Still we know that in rusticalin related gene *AttP*-like site is situated inside the cysteine-rich repeats coding region. Based on that we performed a remote similarity search HMMER using amino acid sequence of cysteine-rich repeats. Cysteine-rich repeats appeared to be part of larger proteins. Therefore conserved domains associated with cysteine-rich repeats were classified.

In spite of the fact that cysteine-rich repeats are found almost exclusively in eukaryotic proteins, they are usually associated with domains typical for prokaryotes or bacteriophages (in 98 proteins out of 124). Among them in 20% (26 proteins) cysteine-rich repeats are associated with phage-lysozyme (PF00959), 14% (17 proteins) with amidase_2 (PF01510). In general nine different domains associated with cysteine-rich repeats can be classified as bacterial cell-wall hydrolyzing enzymes. It is worth mentioning that phage-lysozyme domain is found together with cysteine-rich repeats in proteins of different species and even of different taxa as *Fungi* and *Metazoa*. Based on that observations we can conclude that cysteine-rich repeat in Eukaryotic proteins is usually accompanied by typical prokaryotic domains. The explanation of that might be the presence of bacteriophage recombination site inside cysteine-rich repeat coding sequence, which can facilitate HGT. The 98 genes potentially acquired through HGT from prokaryotes is found as the result.

Funding: The work was supported by program “Molecular and Cell Biology” of the Russian Academy of Sciences and RSF (19-74-20102).

POSTER SESSION

Ontogeny and phylogeny analysis in chordate embryonic development by high throughput sequencing analysis

Song Guo (Skolkovo Institute of Science and Technology, Moscow, Russia)
Haiyang Hu (CAS-MPG Partner Institute for Computational Biology, Shanghai, China)
Chuan Xu (CAS-MPG Partner Institute for Computational Biology, Shanghai, China)

The relationship between development and evolution has long been discussed for many generations of biologists in evolutionary developmental embryology. Early conservation model, hourglass model, and adaptive penetrance model were three main observations, but none of it is favored using the heterochrony data due to different evolutionary scale on species chosen in studies. Therefore, the relationship between ontology and phylogeny is far more intricate than we thought. And on the other side, we lack of a direct overview on the relationship between morphology and temporal gene expression change with the scale of evolutionary history. In this study, we want to know whether embryonically developmental stage among species is comparable or functionally equivalent. We analyzed the dynamic gene expression of eight chordate species (amphioxus, ciona, zebra fish, two species of frogs, turtle, chicken and mouse) and one out-group species (oyster) to extend the comparison. To perform meaningful transcriptome comparison, we extended the Needleman-Wunsch algorithm for gene expression alignment, pair-wisely aligned developmental stages based on developmental stage-specific genes which considered as indicators of functional conserved among species. Contrary to early conservation model and hourglass model, we found parallel relationship on two species embryonic development stage is dominated for all comparisons. And based on parallel stage alignment, we found the most conserved gene expression module is the gene with highly expressed in 2cell/8 cell with the conserved Splicesome related cellular process function.

Genome-wide analysis of multidrug-resistant *Shigella spp.* isolated from patients in Lebanon

Yara Salem (Lebanese American University, Lebanon)
George F. Araj (American University of Beirut, Lebanon)
Sima Tokajian (Lebanese American University, Lebanon)

Shigella spp. are Gram-negative rod-shaped bacteria belonging to the family *Enterobacteriaceae* and are a major cause of bacillary dysentery worldwide. In this study, whole-genome sequencing was used for the molecular characterization of ESBL producing Shigella spp. isolates collected from hospitals in Lebanon. PCRs were performed to detect β-lactam resistance gene reservoirs and to identify the ones mediating virulence and host adaptation. PCR-based replicon typing (PBRT) was performed to identify patterns of plasmid distribution, and multi-locus sequence typing (MLST), whole-genome based SNP analysis, pan-genome analysis, and pulse field gel electrophoresis (PFGE) were performed to determine the phylogenic relatedness of the isolates and to trace evolutionary lineages. *S. sonnei* was the dominant serogroup (8/10 *S. sonnei,* 1/10 *S. boydii,* 1/10 *S. flexneri*). A total of 13 genes conferring resistance to aminoglycosides, β-lactams, sulfonamides, trimethoprim, tetracycline and chloramphenicol were identified, while all the isolates were susceptible to ciprofloxacin and norfloxacin. Five types of β-lactamase genes were detected *blaCTX-M-15, blaTEM-1B, blaOXA-1* and *blaCTX-M-3,* in cephalosporin-resistant isolates. *blaOXA-1* was associated with *S. flexneri,* while *blaCTX-M-15, blaTEM-1B, blaOXA-1* and *blaCTX-M-3* with *S. sonnei. blaOXA-1* was linked to class 1 integron integrated on IncFII type plasmid, while *blaCTX-M-3* was detected on an IncI1 plasmid. The genetic environments of *blaCTX-M-3, blaCTX-M-15* and *blaTEM-1B* were also determined. All isolates harbored virulence genes and tested positive for the invasion plasmid antigen H *(ipaH).* Serine Protease A *(SepA)*, responsible for critically disrupting the intestinal epithelial barrier, was associated with *S. flexneri,* whereas the invasion-associated locus (ial) with *S. boydii. S. sonnei* had a larger core genome (by approximately 78kb) compared to *S. flexneri* and *S. boydii,* both having a smaller core genome but a wider variety of accessory genes. To the best of our knowledge this is the first detailed molecular characterization of Shigella spp. isolates recovered from patients in Lebanon. Our results revealed the association between antimicrobial resistance and increased virulence-related genes and the emergence of strains with high levels of resistance to third generation cephalosporins. Although there are still some active antimicrobial agents that can be used to treat shigellosis, further emergence of antibacterial resistance by inappropriate use should be carefully followed and prevented.

Network analysis of soil fungal communities in strong environmental gradients based on NGS-derived data

Mikryukov Vladimir (Institute of Plant and Animal Ecology, Ural Branch,
Russian Academy of Sciences, Ekaterinburg, Russia)
Lihodeevskiy Georgiy (Institute of Natural Sciences and Mathematics,
Ural Federal University, Ekaterinburg, Russia)

Environmental sequencing became a standard tool for identification of microorganisms and their abundance assessment in natural substrates. Up to now, most NGS-based studies are devoted to microbial community inventorying and diversity estimating. Meanwhile, any ecological community is not just a set of coexisting species. Like any emergent system its inherent properties and functioning is determined by the interactions of its components. Exploration of biotic interactions in ecological communities is a foreground theme in ecology. Till recently, it was achievable only for large organisms characterized by a small number of species per study unit, while microbial communities remain substantially understudied.

In this work, based on ecological network analysis, we characterized between-species interactions in soil fungal communities from forests stretching through two gradients of pollution in the Middle and Southern Urals. The analysis was based on fungal abundance profiles obtained with high-throughput sequencing of *rRNA* gene internal transcribed spacer (ITS2). Ecological networks were inferred with SPIEC-EASI pipeline (SParse InversE Covariance Estimation for Ecological Association Inference).

Obtained results illustrate pollution-induced loss of more than one-third of soil fungal species. Fungal networks in unpolluted areas are characterized by a high number of between-species links and relatively low modularity, indicating that most revealed species cooperate as an integrated community. Pollution caused decrease of network links and increase of network modularity pointing that chemically stressed soil fungi are represented by separate “sub-communities”, most likely due to high environmental heterogeneity of polluted soil.

The work was funded by Russian Foundation for Basic Research with grant no. 18-29-05042.

Metagenome assembled genomes of novel microbial lineages from Lake Baikal in summer and winter seasons

Paul Wilburn (NASA Ames Research Center, Moffett Field, USA)
Justin Podowski (University of Chicago, IL, USA)
Kirill Shchapov (Lake Lakes Observatory, University of Minnesota, Duluth, USA)
Ted Ozersky (Lake Lakes Observatory, University of Minnesota, Duluth, USA)
Maureen Coleman (University of Chicago, IL, USA)
Elena Litchman (W.K. Kellogg Biological Station, Michigan State University, USA)

Microorganisms are the essential agents of biogeochemical cycling in aquatic systems. Recent 'omic advances uncovered their immense taxonomic diversity and functional repertoire. However, the connection between taxonomy and function remains elusive. This is in large part due to the narrow phylogenetic breadth of sequenced microbial genomes, brought on by difficulties in cultivating microorganisms from natural, in particular oligotrophic, environments. Here we present 369 high quality draft metagenome assembled genomes (MAGs) from Lake Baikal, Siberia. They were binned from assemblies comprising 22 sites with wide spatial and depth coverage of the lake in summer, and two samples collected from shallow and deep waters in the winter season. Baikal is the world's most ancient, deep, and voluminous freshwater body. It is a biodiversity hotspot that we hope will contribute important evolutionary insights to genome collections. Our MAGs are culture-independent and span the archaea domain and 15 bacterial phyla, four of which have no previously sequenced representatives from the lake. Most genomes are small but with large variation. At the same time, the most stable, aseasonal, and resource poor sites in the Lake Baikal hypolimnion harbored the smallest genomes with remarkably little size variation. These results could reflect Baikal's overall oligotrophic environment, where millions of years allowed microorganisms to maximize occupancy of available resource niches. We hope this report will set the stage for future model-based work on relationships between phylogenetic diversity and metabolic function in this and other natural systems.

Hybrid assembly and comparative analysis of two *Acanthamoeba castellanii* strains

Cyril Matthey-Doret
Pedro Escoll Guerrero
Agnès Thierry
Pierrick Moreau
Charlotte Cockram
Martial Marbouty
Carmen Buchrieser
Romain Koszul
(Pasteur Institute, France)

Acanthamoeba castellanii is a free living amoeba and one of the most widespread protists. It is the natural host of intracellular bacteria as well as many giant viruses. It also features extensive horizontal gene transfer with those microorganisms, making it ideal for the study evolutionary relationships between intracellular pathogens and their hosts. However, the current genome assembly of A. castellanii is fragmented and largely incomplete. Here, we assemble the genomes of two A. castellanii strains *de novo* using a hybrid approach combining Oxford Nanopore long reads with 3C (chromosome conformation capture) based reassembly using our homemade program instaGRAAL. We achieve near chromosome-level assembly and characterize the spatial organisation of the *A. castellanii* genome, highlighting a strong clustering of centromeres. We are now performing a comparative analysis of viral and bacterial gene content between the two strains to gain insight into their evolutionary histories in terms of pathogen association.

Probiotics treatment of IBS
in Russian and Vietnamese patients

Yulia Kondratenko (Saint Petersburg University, Russia)
Alexander Suvorov (Institute of Experimental Medicine, St. Petersburg, Russia)
Alena Karaseva (Institute of Experimental Medicine, St. Petersburg, Russia)
Marina Kotyleva (Institute of Experimental Medicine, St. Petersburg, Russia)
Nadezhda Lavrenova (Institute of Experimental Medicine, St. Petersburg, Russia)
Anton Korobeynikov (Saint Petersburg University, Russia)
Galina Leontieva (Institute of Experimental Medicine, St. Petersburg, Russia)
Tatiana Kramskaya (Institute of Experimental Medicine, St. Petersburg, Russia)
Igor Kudryavtsev (Institute of Experimental Medicine, St. Petersburg, Russia)
Alla Lapidus (Saint Petersburg University, Russia)
Elena Ermolenko (Institute of Experimental Medicine, St. Petersburg, Russia)

Irritable bowel syndrome is a group of diseases with shared symptoms — abdominal pain and changes in bowel movement. These symptoms are accompanied by reduction of gut microbiome diversity and some changes in microbiome composition, such as decrease in bacteria from phylum *Bacteroidetes*. In our work we investigated the effects of probiotics treatment of IBS in patients of Russian and Vietnamese origin. Investigated probiotics included *autoenterococci, bifidobacteria* and *Enterococcus faecium L3 strain*. Microbiome composition was examined by sequencing of *16S rRNA* variable segments V3 and V4. CD-HIT-OTU-MiSeq was used for clustering of resulting reads and operational taxonomic units (OTUs) retrieval. Greengenes database v.13.5 was used to annotate OTUs. Microbiome composition was different in IBS patients of Russian and Vietnamese origin. Vietnamese patients with IBS had higher fractions of phyla *Proteobacteria* and *Actinobacteria* than patients of Russian origin. Vietnamese patients with IBS had higher alpha diversity of microbiome than patients of Russian origin. Microbiome composition also varied in healthy individuals of Russian and Vietnamese origin with higher fraction of *Bacteroidetes* in Russian individuals and higher fraction of *Actinobacteria* in Vietnamese. Difference in microbiome composition between healthy individuals and patients with IBS was more prominent in Vietnamese individuals than in Russian. Effect of probiotics treatment on microbiome composition was also more prominent in Vietnamese patients with IBS than in Russian patients. These data provide one more example of the influence of region of origin on microbiome composition, on microbial features associated with disease and on the effectiveness of treatment of the same disease.

This work has been supported by the Russian Science Foundation (grant 16-15-10085)

Analysis of genomic dataset from the host-parasite system *Paramecium — Microsporidia*

Yulia Yakovleva (Saint Petersburg University, Russia)
Natalya Bondarenko (Saint Petersburg University, Russia)
Natalia Lebedeva (Research park, Saint Petersburg University, Russia)
Andrey Vishnyakov (Saint Petersburg University, Russia)
Elena Sabaneyeva (Saint Petersburg University, Russia)
Elena Nassonova (Saint Petersburg University, Russia;
Institute of Cytology, Russian Academy of Sciences, St. Petersburg, Russia)

Microsporidia is a group of obligate intracellular parasites related to fungi.These protists are found in nearly all groups of animals. Though *Microsporidia* have a wide host range, they are rare in protists, and, especially in ciliates. Therefore investigation of this host-parasite system is of special interest. At the present time almost 30 microsporidian genomes are available in NCBI database, however full multigenic phylogenetic revision of Microsporidia has never been performed so far. *Paramecium primaurelia strain SpM5-3* infected with a new species of microsporidia was used for sample preparation. Parasite life stages were isolated from the host cell manually by means of micromanipulator. The material obtained from a few parasitised paramecium cells was stored and used later for total DNA extraction performed using the Arcturus PicoPure DNA kit (Thermo Fisher Scientific). Whole genome amplification of the DNA was carried out using multiple displacement amplification with the REPLI-g kit (QIAGEN). After assessing the DNA quantity we prepared two TruSeq Paired-end libraries and sequenced them on HiSeq 2500 Illumina (2 x 125 bp) with v4 chemistry. 32,104,294 and 31,990,912 reads were obtained for each library.
The quality of paired-reads was assessed with FastQC before and after quality trimming and trimming adapters with Trimmomatic v.038. Trimmed reads were corrected using SPAdes v.3.13.0 and were aligned against *Paramecium* reference genomes from NCBI database using BWA v.0.7.17. Mapped reads were eliminated, unmapped reads were assembled using SPAdes with five k-mer values (21, 33, 55, 77, 99) and --careful option. Final assembly was aligned against 56 opisthosporidian marker genes with BLAST+ (blastn v.2.5.0). Though we did not obtain the full microsporidian genome, 36 marker genes were found for further multigenic phylogenetic analysis.

Supported by RFBR grant 18-04-00562 and RSF grant 19-74-20136.

Genomics-based quantitative analysis of metabolic potential in microbial communities

Stanislav Iablokov (IITP Russian Academy of Sciences, Moscow, Russia;
P.G. Demidov Yaroslavl State University, Russia)
Dmitry Rodionov (IITP Russian Academy of Sciences, Moscow, Russia;
Sanford Burnham Prebys Institute, San Diego, CA, USA)

High-throughput metagenomic sequencing allows to study microbial communities, such as human or soil microbiota, based on *16S rRNA* amplicon data. Taxonomic profiling of a given community is usually done based on the amplicon's relative abundance values. We've developed a new approach, complementary to the phylogenetic description, that allows to infer metabolic capabilities of community, which may include but not limited to vitamins production, carbohydrates utilization, quorum sensing activity etc. Each capability is described in terms of Community Phenotype Index (CPI) which represents the expected fraction of bacterial cells with predicted metabolic capability. These CPIs are calculated using the abundance weighted average taxonomy-based mapping to the collection of binary phenotypes for 2662 reference bacterial organisms. Binary phenotypes, where 1 stands for the presence of a particular metabolic pathway and 0 stands for its absence, were obtained by mcSEED-based metabolic reconstruction of pathways using complete genomes.

To predict more accurate values for CPIs we use two additional heuristics. First, each OTU's representative sequence is given a multitaxonomic assignment with higher phylogenetic resolution (usually on the species level) instead of a single taxonomy with low resolution consensus (i.e., on the family level). This approach allows to narrow the range of reference organisms whose phenotypes are being averaged, therefore, reducing the uncertainty in CPI prediction. Secondly, relative abundance values of OTUs are renormalized using *16s rRNA* copy count values derived from rrnDB project. This procedure provides us with presumably original proportions of individual organisms' abundances of a given community.

We've investigated the effect of these heuristics for three popular human gut microbiome studies (HMP, AGP, UK Twins) and discuss the results and potential applications.

Genome heterogeneity affecting binning of complex fungal communities

Gulnara Tagirdzhanova
Toby Spribille
(University of Alberta, Canada)

The vast majority of fungi are yet to be described and cultured. Since in nature these species mostly occur mixed with other organisms, accessing genomic information from these fungi is a serious challenge. Shotgun sequencing techniques do not offer a reliable way to extract the genome of a target fungus from a mixed dataset, which might include other eukaryotic genomes. Previously, some standard database-independent binning approaches were applied to metagenomes of complex eukaryotic communities. These methods are based on oligonucleotide frequency distribution and rely on the assumption of homogeneity of sequence composition across any given genome. This assumption, however, might not hold true for some fungi. Genomes of these species show strong intragenomic difference in base composition, a phenomenon thought to be caused by repeat-induced point mutation (RIP). RIP is a mechanism used by fungi against transposable elements, silencing multicopy DNA elements by directed mutational processes. Lichens are complex symbiotic communities including multiple species of fungi, algae, and bacteria, and represent a case where two phenomena, unculturable fungi and heterogeneous fungal genomes, overlap. In our study, we aim to assess the extent to which genome heterogeneity might affect metagenomic binning and propose a strategy to improve the binning of complex fungal communities.

Evaluation of assemblers and development of analysis pipeline for gut virome classification

Yasumasa Kimura (Division of Systems Immunology,
The Institute of Medical Science, The University of Tokyo, Japan)
Kosuke Fujimoto (Department of Immunology and Genomics,
Osaka City University Graduate School of Medicine, Japan)
Seiya Imoto (Division of Health Medical Data Science, Health Intelligence Center,
The Institute of Medical Science, The University of Tokyo, Japan)
Satoshi Uematsu (Department of Immunology and Genomics,
Osaka City University Graduate School of Medicine, Japan)

Metagenomic research on intestinal viral microbiome (virome) is still challenging compared with intestinal bacterial microbiome analysis. There is no universally conserved gene in viruses equivalent to the *16S rRNA* gene in bacteria or the internal transcribed spacer in fungi, which are used for taxonomic classification. Therefore, whole metagenome shotgun sequencing is required to estimate the diversity and taxonomy of viruses. Since public databases currently lack a sufficient accumulation of viral genomes, the vast majority of sequence reads do not align to known viral sequences; these are termed “viral dark matter” and present a major obstacle in comprehensively defining viromes.

Recently, metagenomic assembly approach that reconstructs viral genomes (contigs) that are used as a starting material of the analysis is often taken. After constructing contigs, genes on the contigs are predicted and homologous viral proteins are analyzed. Although de novo assembly is the first and important analysis step, assembly results are known to be varied by employed assemblers. Therefore, using spike-in viruses, we compared four assemblers: SPAdes, MetaSPAdes, IDBA-UD, and MEGAHIT. We added four spike phages comprising two single-stranded DNA phages and two double-stranded DNA phages to mice feces and performed sequencing analysis to obtain viral sequencing reads. The assemblers were evaluated by comparing the assembled contigs with reference genomes of the four spike phages in terms of the size of assemblies and the number of misassemblies. The result indicated that MetaSPAdes was the most effective assembler of the gut viral metagenome.

Taking the result of assembly comparison into account, we developed virome classification pipeline. From the raw reads, the pipeline performs a series of analyses, quality filtering and error correction, assembly by MetaSPAdes, selection of viral contigs by VirSorter and VirFinder, open reading frame (ORF) prediction, taxonomic assignment to ORFs by homologous viral proteins using GHOST-MP, and output classification of contigs based on viral protein taxonomy and viral structural proteins annotated by Pfam.

In this presentation, we will show the details of the assembly comparison and discuss the effects of analysis tools and their parameters on viral contig classification.

Identification of small RNAs derived from commensal microbiota or infections

Pawel Zayakin (Latvian Biomedical Research and Study Centre, Riga, Latvia)

The unpredictable mixture of sRNAs of human and non-human origin in RNA-seq data is the most complex problem to be resolved for analysis of the data obtained from a wide range of biofluids. The other species sources (bacteria, fungi and viruses) should be separated for more accurate analysis of the sRNA reads of human origin. A new algorithm, which allows reducing false assigning of reads to improper species of origin by two-pass analysis of Blast output on "nr" database for a representable random subset of reads, has been developed for this purpose. The second pass will assign the hit to the species, which were most frequently encountered in the first pass, in case of a similar score. At the same time, valuable research information on accompanying species will be also obtained. Only the genomes of the species, most represented in successful Blast hits will be used for aligning step. Contrary to the case with full-length mRNA, it is common for sRNA that the reads are aligned in multiple sites of the genome. Therefore, the usage of the commonly used alignment methods, when multi-aligned reads are ignored or randomly chosen as one from the equally scored alignments, can be inefficient. Reads are aligned allowing multiple alignments per read and then they are reassigned taking into account the local coverage by ShortStack algorithm. In addition, the pipeline currently includes the creation of a catalogue of expressed RNA types using human genome annotations and differential expression analysis using DESeq2 for all of the classifiable RNA types, generation of metagenomic profile and diversity statistics. Human genome annotations have been expanded and include Ensembl database, as well as miRBase, lncipedia, piRBase, piRNAdb, piRNAbank, GtRNAdb and GtRNAdb derived tRFs databases. The prioritization algorithm for building a catalogue of expressed RNA types, which allows solving the problem with an annotation overlap has been developed. Presented algorithms will be included in a future release of sRNAflow — a software tool for the analysis of small RNAs in biofluids.

Identification of B/TCR sequences created by additional diversification mechanisms in the Rep-Seq data

Andrey Slabodkin (Independent researcher)
Maria Chernigovskaya (ITMO University, Saint Petersburg, Russia)
Anastasia Vinogradova (ITMO University, Saint Petersburg, Russia)

V(D)J-recombination is one of the main mechanisms responsible for diversity of adaptive immune repertoires. Most tools for analysis of repertoire sequencing data rely on the common knowledge about the recombination process: there is exactly one V, one D and one J segment in the rearranged immunoglobulin gene, along with the random insertions between them. However, there are additional mechanisms that cannot be described in this standard model: VH-replacement, gene conversion, V/D gene duplication, etc.

Here we present a tool for fast and accurate identification and analysis of B/TCR sequences that were created with the help of such mechanisms in the Rep-Seq data. We believe that analysis of these sequences can help understanding the additional processes responsible for B/TCR diversity.

Bioinformatic analysis of the structure and taxonomic features of *Azospirillum brasilense Sp245* flagellum proteins

Sergei Shchyogolev
Angelina Budanova
Larisa Matora
(Institute of Biochemistry and Physiology of Plants and Microorganisms,
Russian Academy of Sciences, Saratov, Russia)

We analyzed for the first time the 3D structures and taxonomic characteristics of *Azospirillum brasilense Sp245* flagellins, involved in the formation of the bacterial H antigen. These included *FlgE* (hook material), *FlgK* and *FlgL* (hook-filament linking elements), *FliC* (filament material), and *FliD* (cap material), all responsible for the immunochemical properties and various effects of bacteria–environment interactions. Amino acid sequences of the flagellins were selected from databases by their annotations and by using SmartBLAST technology. 2D and 3D structures were determined by homologous modeling with I-TASSER software. The 3D structures of the proteins were compared and the taxonomic relationships between them were determined at the 3D level with the mTM-align program. Phylogenetic analysis of the molecular sequences by routine methods was supplemented with results based on whole-genome DNA sequencing by using the ANI and AAI tests. By comparing the amino acid sequences and 3D structures, we identified varieties of the flagellins whose coding sequences were localized in the genome of *A. brasilense Sp245* on different replicons. Despite the high homology (E-value=5e–09) and the similarity of the 3D structures of the *FliC* from azospirilla and salmonella, the amino acid sequences of these proteins were only 20% identical and 53% similar. The overall 3D structure of the *FliC* from *A. brasilense Sp245* corresponded to that of its distantly related homologues from members of *Salmonella, Pseudomonas, Burkholderia,* and *Sphingomonas,* with a presence in these proteins of a common evolutionarily developed core region, which was located in their highly conservative D0–D2 domains (alpha helices). Comparison of the 3D structures of the *FliC* from members of the more closely related genera *Azospirillum, Niveispirillum,* and *Nitrospirillum* showed that this core region also extended to the variable domain D3 (loops and beta sheets).

Under standard conditions of SmartBLAST, all sequences of the flagellins closely related to the *FliC* from *A. brasilense Sp245* were among the proteins of bacteria within the family *Azospirillaceae* of the order *Azospirillales*, a class of alpha-proteobacteria in accordance with the nomenclature of the GTDB project of global phylogenetic studies of prokaryotes. Thus, without going in fact beyond the family *Azospirillaceae* according to this trait, the azospirilla acquired a composition of the amino acid sequences of the polar flagellum filament protein *FliC* that is apparently unique to this family. Nevertheless, it ensures the formation of a 3D flagellin structure that overall corresponds to the typical 3D structures of functionally similar proteins from fairly distantly related bacteria.

When the *FliC* amino acid sequences were used as phylogenetic markers for members of the Azospirillaceae, a high level of correspondence was found between the topology of the obtained phylogenetic constructs and that of the trees resulting from the whole-genome ANI and AAI tests. This fact makes it necessary to correct the views regarding the level of conservatism of the *FliC* sequences and their possible usefulness in taxonomic studies of prokaryotes. This opportunity can be ensured by the molecular and spatial variability of their D3 intermediate domain.

Target selection protocol for DNA-machines development

K. P. Chalenko (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russia)
M. S. Rotkevich (Theodosius Dobzhansky Center for Genome Bioinformatics, Saint Petersburg University, Russia)
D. M. Kolpashchikov (Chemistry Department, University of Central Florida, Orlando, FL, USA)
E. I. Koshel (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russia)

DNA-machines are constructions based on deoxyribozymes that can cleave mRNA of target gene. This technique was successfully used against cancer cells and Influenza A Virus. Deoxyribozyme based DNA-machines are universal approach that can be applied to prokaryotic and eukaryotic organisms. Choosing the right target gene is still a fundamental stage for using DNA-machines. In case of eukaryotic organism number of housekeeping genes can reach of thousands, what makes cumbersome subsequent gene analysis by hand. Faced with such a large amount of data existing protocols were not effective enough. We developed Python script to manage this challenge.

To achieve the goals of our research we developed Python script utilizing Entrez library, BLAST software and The NCBI SRA Toolkit to access mRNA sequences and estimate the level of gene expression. Firstly, it downloads and creates local indexed databases via 'sra-toolkit' and 'makeblastdb' applications respectively. Secondly, it queries genes sequences in prepared databases to retrieve summary statistics for their occurrences using 'blastn' software. This software speed up the detection of over-expressed genes, moreover it could deal with both eukaryotic and prokaryotic organisms. For this purpose, in case of prokaryotic organism it analyzes DNA sequence of gene, whereas for eukaryotic organism it operates with mRNA sequence of gene due to occurrence of introns.

DNA-machine is very sensitive to mismatches in sequences therefore rapid evolution of genes can disrupt the process of mRNA cleaving. According to this, MEGA program was used to identify the most conservative genes.

Furthermore, to extend the time of work DNA-machines we should choose genes with stable mRNA. This characteristic is determined by the half-life of the mRNA. In addition, many genes are connected with replication process, but replication cycle may take a long time. We should avoid these targets to reduce the time of in vivo experiments.

To evaluate the essentiality of target genes we verified it in BioCyc Database Collection. This database can demonstrate result of selected genes knockout. The absence of vulnerable housekeeping gene will lead to cell death.

CDSnake: Snakemake pipeline for retrieval of annotated OTUs from paired-end reads using CD-HIT utilities

Yulia Kondratenko
Anton Korobeynikov
Alla Lapidus
(Saint Petersburg University, Russia)

Sequencing of *16S rRNA* is a commonly used method for cost-efficient characterizing of microbial communities. Illumina paired-end reads are often used as sequencing method. Since even short variable regions of *16S* provide sufficient information for microbe identification, sequenced fragment length is often smaller than the sum of lengths of paired reads. Thus reads of pairs can be merged for downstream analysis, commonly including clustering of sequences and matching the resulting clusters' representative sequences with annotated database. In spite of development of several tools for assembly of paired-end reads into contigs, poor quality at the 3' ends of both paired-end reads in the overlapping region prevents the correct assembly of significant portion of read pairs. Wrongly or uncertainly merged reads either have to be excluded from downstream analysis or retained with high risk of spurious sequences creation.

Recently CD-HIT-OTU-Miseq was presented as a new approach, avoiding reads merging due to separate clustering of paired reads and discarding of reads voting for non-matching clusters as chimeric. CD-HIT-OTU-Miseq utilities are command line tools written in C++ and Perl. Here we assembled CD-HIT-OTU-Miseq utilities into pipeline using Snakemake workflow. We benchmarked our pipeline with two commonly used pipelines for OTU retrieval, incorporated into popular workflow for microbiome analysis, QIIME2 – DADA2 and deblur. Benchmarking was made of 3 mock datasets, Balanced, HMP, and Extreme, each sequenced at a depth of over 500,000 highly overlapping paired-end Illumina MiSeq 2 × 250 reads. The Balanced community contained 57 bacteria and archaea at nominally equal frequencies, the HMP community contained 21 bacteria at nominally equal frequencies, and the Extreme community contained 27 bacterial strains at frequencies spanning five orders of magnitude and differing over the sequenced region by as little as 1 nucleotide (nt). Balanced dataset had higher sequence quality (Mean Q = 35.9 forward/33.5 reverse); Extreme had moderate quality (33.0/29.3); and HMP had lower quality (32.3/28.7). CDSnake outputted less OTUs than DADA2 and deblur, since last two tools aim to output sub-OTUs by error processing, and OTU-MiSeq doesn't process errors and tries to output most correct OTUs using clustering. However, on Balanced and HMP datasets number of OTUs outputted by CDSnake was closer to real number of strains which were used for mock community generation, than those outputted by DADA2 and deblur. On Extreme dataset CDSnake, as expected, performed worse than DADA2 and deblur, since clustering algorithm cannot separate sequencing errors from actual 1-nt differences, present between strains in this community.

CD-HIT-OTU-MiSeq provides one more approach for amplicon analysis capable to outperform popular tools in certain conditions. We developed Snakemake pipeline for OTU-MiSeq utilities, which can be helpful for easier automated runs.

This work has been supported by the Russian Science Foundation (grant 19-16-00049)

Analysis of plasmid CRISPR-Cas systems and their potential for plasmid host range prediction

Iana Fedorova (Skolkovo Institute of Science and Technology, Moskow, Russia)
Mikhail Kongoev (ITMO University, St. Petersburg, Russia)
Mikhail Raiko (Center for Algorithmic Biotechnology, Saint Petersburg University, Russia)

Horizontal gene transfer plays a highly important role in evolution of bacteria. Presumably, gene exchange between bacteria occurs by genetic mobile elements such as plasmids and bacteriophages. But nowadays there is no reliable way to check if the certain plasmid can “travel” between bacteria of different origin, and how wide the plasmid host range could be. Also it is important to be able to predict plasmid host in case of metagenomic data, where we usually have dozens of novel plasmids without any information of host species.

To answer this question, we analyzed CRISPR cassettes in bacterial genomes — repetitive sequences in bacterial DNA, interspaced with unique “spacer” sequences, which were extracted from genetic mobile elements infected the bacteria or its ancestors. Spacers in CRISPR cassette can be considered as a link between the plasmid and its host. We used CRISPR Finder spacers database and the RefSeq database of all plasmids known to date. Blasting spacers over plasmids sequences allowed us to determine plasmid host ranges: variety of bacterial organisms where the plasmid can exist. By taxonomy analysis we found some plasmids which can live in different families of organisms, they can be useful in genetic engineering as a natural shuttle vectors.

Taxonomy analysis showed that a bunch of plasmids have additional hosts except of host they were related to according to RefSeq database: 543 hits — additional hosts of different genus, 29 — different family, 19 — different order, 12 — different class and even 2 additional hosts of different phylum. Thus, such broad host range plasmids can be important players in the process of evolution. We also found that a lot of plasmids carry their own defense CRISPR systems (10% of RefSeq plasmids). Part of these systems (10%) seems to be active — there are Cas1 gene homologs near CRISPR cassettes. Role of these systems in plasmid propagation, host fitness and evolutionary relationship with the known chromosomal CRISPR-Cas systems is the subject of future research.

Identification differentially expressed genes in cadmium-tolerant mutant pea *(Pisum sativum L.)* line *SGECdt*

Olga A. Kulaeva (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Mikhail L. Gordon (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Alexey M. Afonin (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Evgeny A. Zorin (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Igor A. Tikhonovich (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia; Saint Petersburg University, Russia)
Vladimir A. Zhukov (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)
Viktor E. Tsyganov (All-Russia Research Institute for Agricultural Microbiology (ARRIAM), Pushkin, St. Petersburg, Russia)

Cadmium is a widespread pollutant often can be found in biologically available form in soil and water. It is toxic for the majority of organisms, due to its ability to disrupt metabolism at the cellular level. Some plant species are able to accumulate cadmium in their biomass in toxic concentrations for another organisms. Garden pea mutant *SGECdt* shows tolerance to cadmium and accumulates significant concentrations of this metal in roots and shoots. The mutation in the *cdt* locus leading to this phenotype is characterized by a monogenic recessive mode of inheritance and was localized in the VI linkage group (Kulaeva, Tsyganov, 2013). But the biological mechanisms underlying the cdt phenotype remain unclear.

This work aims to identify differences in gene expression levels in plants of the initial line *SGE* and mutant *SGECdt* under control conditions and after cadmium exposure. In order to achieve this, plants were grown hydroponically with addition of 3 μM CdCl2 and without cadmium (control), harvested after 1 and 3 days of growth. Assessing transcript abundances was performed using 3'MACE technology, which is a modification of RNA-seq developed by GenXPro GmbH. MACE-Seq is based on sequencing the 3’ end of each cDNA molecule and provides PCR-unbiased transcript quantification.

While Pisum sativum genome assembly still unavailable, a transcriptome assembly with 94360 transcribed pea sequences was used as a reference for aligning 3'MACE reads. Alignment and quantification were performed using the BBmap package. Different samples were sent for sequencing at different times, shifting representation of some transcripts in the obtained data (batch effect). The inclusion of a technical predictor in the regression model eliminated undesirable variance from experimentally significant factors.

Differential expression analysis was carried out using DESeq2 package. As a result, the most significant differences in the expression of certain genes were identified depending on each of the experimental factors: genotype, exposure to cadmium, and the time of growth. Transcriptome assembly was annotated using both Trinnotate and MapMan, allowing to assign functional annotation to differentially expressed genes and perform Gene Ontology enrichment analysis. Analysis in terms of the ontologies of biological processes and molecular functions revealed several functional groups among differentially expressed genes, such as bivalent cation transporters, regulators of the response to oxidative stress, as well as the transport of lipids and some others.

Currently, work is underway to refine the annotation and compare the results with published data, focusing on the involvement of the identified genes in metabolic pathways and oxidative stress regulation. This will give a more complete and systematic explanation of the biological mechanisms underlying the *cdt* phenotype.

Transcriptomic landscape of follicular lymphoma

Anna Gorbunova (Saint Petersburg University, Russia;
Mechnikov North-West State Medical University, St. Petersburg, Russia)
Yuri Krivolapov (Mechnikov North-West State Medical University, St. Petersburg, Russia)
Ekaterina Bozhokina (Institute of Cytology RAS, St. Petersburg, Russia;
Mechnikov North-West State Medical University, St. Petersburg, Russia)
Maria Firuleva (Saint Petersburg University, Russia)
Igor Evsukov (Saint Petersburg University, Russia)

Follicular lymphoma is a tumor that arises from a germinal-center B cells. In the majority of cases, tumor cells acquired a t(14;18) translocation that upregulates BCL2, a key gene in the apoptosis. Microenvironment with non-neoplastic T cells, macrophages, and dendritic cells plays a critical role in follicular lymphoma progression. Understanding the relationship between the tumor cells and components of the microenvironment will be crucial for developing new targeted treatment options. A major aim of approaches that target the microenvironments is to improve the function of immune effector cells. These approaches include immune checkpoint inhibition with antibodies against PD1. For many cancer patients, anti-PD1 immunotherapy is effective, however, resistance remains a challenge. Further investigations needed for increasing the efficacy of anti-PD1 therapies.

In this study whole transcriptome RNA-sequencing profiling was performed on biopsy specimens obtained from patients with untreated follicular lymphoma and non-tumor lymph nodes. Combined gene expression and fusion transcript analyses revealed the presence of known oncogenes and novel rearrangements. We identified a total of 997 differentially expressed genes, including 430 upregulated and 567 downregulated genes in tumor samples with the criteria of log2FC ≥ 1 and FDR < 0.05. Upregulated genes were enriched in the canonical pathway gene sets associated with B cell receptor signaling pathway, DNA repair, cellular response to stress, chromosome organization, cell cycle and RNA processing. Downregulated genes were enriched in gene sets related to the immune response, especially T-cell activation and cytokine production. The ‘T-cell activation’ group included CD28, CCR2, CCR7, IL7R, IL6R, IL15 and other genes. The protein encoded by CD28 gene belongs to the B7-CD28 stimulatory checkpoint molecules superfamily and is essential for T-cell proliferation, survival, cytokine production, and T-helper type-2 development. CD28 co-stimulation is also essential for the efficacy of anti-PD1 therapy and in the absence of CD28/B7 interactions, anti-PD1 therapy failed to rescue T cells from their exhausted state.

The work was carried out with the support of the Russian Foundation for Basic Research, project No.17-54-33051.

Analysis of microbiomes of the ecogenetic series of podzolic soils using an artificial neural network

Ekaterina Ivanova (Saint Petersburg University, Russia)
Elizaveta Pershina (Saint Petersburg University, Russia)
Vasilieva Nadezda (V.V. Dokuchaev Soil Science Institute, Moskow, Russia)
Evgeny Andronov (Saint Petersburg University, Russia)
Evgeny Abakumov (Saint Petersburg University, Russia)

The application of molecular-genetic methods is currently one of the necessary steps in the natural microbiomes analysis. The use of these approaches in the study of microbial complexes of soil chronoseries is promising in identifying of taxonomic markers and microbiological drivers of pedogenesis. At the same time, the search for optimal ways of high-throughput sequencing data processing remains relevant today.

The aim was the analysis of the prokaryotic community of the genetic horizons of the podzolic soils chronosequence formed on the surface of the open pits for the sandy quarry. Sample set included: 1–2 years, fresh sand lithostrat with no signs of pedogenesis; three stages of self-overgrowing (15–20 years with pine-tree and sod-podbur occurrence, 30–35 years old with embryopodzol, 70 years with the distinct podzol profile) and the background soil under the pine forest near the quarry.

Soil samples were analyzed by classical methods (physic-chemical), DNA isolation was performed using the PowerSoil® DNA Isolation Kit (MO BIO, USA). To amplify a fragment of the 16S rRNA gene, universal primers (F515 and R806) to its V4 variable region were used (Bates et al. 2010). The sequencing of the libraries obtained was performed by ILLUMINA MiSeq. Processing of the sequence data was carried out using “Trimmomatic” (Bolger et al. 2014) and “QIIME” (Caporaso et al. 2010) software. Samples clustering and concurrent microbial taxonomic composition analysis was carried out using an artificial neural network that is trained using unsupervised learning - Kohonen's self-organizing map (R package ‘kohonen’; Wehrens and Kruisselbrink, 2018; Chitwood et al, 2013).

Characteristic patterns of the relative contents of prokaryotic taxa were obtained in the early and late stages of podzolic soils succession. Specific patterns have been identified for different soil horizons. Based on the analysis of the relationship of the characteristic patterns obtained, the leading factors and mechanisms of changes in the composition of the microbiome caused by the time and soil profile differentiation were revealed.

Taking into account the complex nature of high-throughput sequencing data, the use of neural networks with training is a promising approach to the analysis of soil microbiomes, which in turn helps get the biologically relevant information (the adaptive and evolutionary strategies of microorganisms during soil-formation) and practically valuable results.

This work was supported by Russian Scientific Foundation, project № 17-16-01030

Improved Architecture of Artificial Neural Network for Secondary Structure Analysis

Semyon Grigorev
Polina Lunina
(Saint Petersburg University, Russia)

Algorithms that can efficiently perform sequences classification and subsequences detection have recently become a focus in bioinformatics and many of them utilize the idea about considering these sequences secondary structure. One of the classical ways of describing secondary structure is formal grammars.

An approach for biological sequences processing by using a composition of formal grammars and neural networks is proposed. While classical way is to model secondary structure of the full sequence by using grammar, the proposed approach utilizes it only for primitive secondary structure features description. These features can be extracted by parsing algorithm and processed by using an artificial neural network. It is shown that this approach is applicable for real-world data processing and some questions are formulated for future research. In this work, we provide answers to some of them.

The first question is whether it is possible to use convolutional neural networks for parsing result processing. The result of the matrix-based parsing algorithm for an input string and fixed nonterminal is an upper triangular boolean matrix. Presently, we came up with two possible ways of these matrices representation. The first one is to drop out the bottom left triangle and vectorize the rest of matrix row by row. It requires the equal length of the input sequences, therefore we propose to either cut sequences or add some special symbol till the definite length. The second way is to represent this matrix as an image: the false bits of the matrix as white pixels and the true bits as black ones. This approach makes it possible to process sequences with different length since the images may be easily transformed to the same size. To handle these images we use the network with a small number of convolutional layers, linearization and then the same structure as for vectorized data (dense and dropout layers with batch normalization).

The second question is whether it is possible to move parsing to network training step. This question is important because parsing is the most time-consuming operation of the proposed solution. We solve this problem by using two-staged learning. At the first step, we prepare a neural network (vector- or image-based) for our task which takes parsed data as an input. After that, we extend the trained network with a number of input layers that should convert the original nucleotide sequence into parsing result. This way we create a network which can handle sequences, not parsing result. So, parsing is required only for training the first network.

We use the proposed improvements to create neural networks for tRNA sequences analysis problems: classification of tRNA into 2 classes: eukaryotes and prokaryotes and 4 classes: archaea, bacteria, plants, and fungi. We train networks on sequences from GtRNAdb and tRNADB-CE. Accuracy for both image- and vector-based classifiers is up to 90% on the test set. Accuracy for networks which handle sequences is better than the accuracy of networks which handle parsing results.

Closing gaps in draft genome assemblies using Oxford Nanopore sequencing and Read-Until technology

Sergey Kazakov (ITMO University, St. Petersburg, Russia)
Vladimir Ulyantsev (ITMO University, St. Petersburg, Russia)
Sergey Nurk (Saint Petersburg University, Russia)

For many biological and medical researches it is very important to have a complete genome sequence of organisms included in the study. However, for many of them there are no such sequences, even for organisms thoroughly studied for a long time.

While the most reliable sequencing technology both for genome and metagenome projects still remains Illumina, the third generation sequencing technologies (such as Oxford Nanopore and PacBio) become more accessible and wide spread nowadays. They can produce ultra-long reads (hundreds of kbp) that can be used to solve “complicated places” in assembly, originating because of repeats and identical genomes’ parts in different species. Moreover, Oxford Nanopore sequencer provides Read-Until technology that brings the ability to skip current DNA molecule while reading process is going on! This technology can significantly reduce the effective cost of assembly projects.

In current work we proposed several strategies how to use this technology to close gaps in draft genome assembly; in which cases it is reasonable to use it and what benefits one can get using it. In more detail, we assume that a draft assembly is available for studied organism. Using such fragmented assembly as a reference, it is possible to select only such Nanopore reads, which very likely will connect two or more contigs of assembly.

For experiments we use two datasets with R9 and R9.4 Nanopore reads for *Escherichia coli str. K-12* bacteria. The initial assembly consists of 52 long contigs with 52 gaps between them. Selecting only such reads, we showed that we can close 83% of gaps with 1.9 times more useful reads covering the gaps compared to the baseline, for the first dataset with R9 reads, and 94% of gaps and 2.0x times more useful reads for second dataset. It is known that the main problem with such strategies is “short reads” appearing during Nanopore sequencing. Including minimal read length threshold to 5 kbp, enrichment increases up to 2.5x for useful reads count with small change in number of covered gaps. Results for other organisms will also be presented.

imputeqc: An R Package for assession and optimization of genotype imputation parameters

Gennady V Khvorykh
Andrey V Khrunin
(Department of Molecular Bases of Human Genetics,
Institute of Molecular Genetics of Russian Academy of Sciences, Moskow, Russia)

Genotype imputation has a potential for harmonization of genotype datasets, thus increasing the power of studies. However there are difficulties with imputation of data from different platforms or data having low frequency polymorphisms. Therefore, the imputation quality should be assessed in each particular case. Nevertheless, not all imputation software control the error of output, e.g., last release of fastPHASE program (1.4.8) lacks such an option. There is also an uncertainty in choosing the parameters for imputation models. fastPHASE is based on haplotype clusters, where the number of clusters should be set a priori. The parameter influences the results of imputation and downstream analysis.

We present an approach and software toolkit imputeqc to assess the imputation quality and/or to choose the model parameters for imputation. We demonstrate the toolkit for estimating the genotype imputation quality, when imputations of genotypes from HapMap and 1000 Genomes Project are made with fastPHASE and compared to BEAGLE software. It is also shown how to choose the optimal numbers of haplotype clusters and expectation-maximization cycles to be applied with fastPHASE program. The number of haplotype clusters thus estimated is further applied for hapFLK testing that revealed signatures of selection in multiple population dataset from North European and Western Siberia parts of Russia.

The toolkit is implemented as an R package imputeqc and command line scripts. The code is freely available at https://github.com/inzilico/imputeqc under the MIT license.

The research was supported by the Russian Foundation for Basic Research (grant No 19-29-01151).

SATURDAY – JUNE 22, DAY 3 SCHEDULE

B – Break	I – Invited Talk	O – Opening or Closing Talk	F – Featured Talk
T – Talk	D – Dinner	P – Posters	

10:00AM–11:00AM	I	***A moving landscape of comparative genomics in mammals*** Stephen O'Brien *Saint Petersburg State University*
11:00AM–11:20AM	T	**A universal transcriptomic signature of age reveals the temporal scaling of Caenorhabditis elegans aging trajectories** Andrei E. Tarkhov *Skolkovo Institute of Science and Technology*
11:20AM–11:40AM	T	**DASE-AG: conditional-specific differential alternative splicing events estimation method for around-gap regions** Kouki Yonezawa *Nagahama Institute of Bio-Science and Technology*
11:40AM–12:15PM	B	**Break**
12:15PM–13:15PM	I	**Adapting bioinformatics to bacteriophage genomics and virome studies** Marie-Agnès Petit *Micalis Institute, INRA*
1:15PM–1:35PM	T	**NPS: scoring and evaluating the statistical significance of peptidic natural product–spectrum matches** Azat Tagirdzhanov *Saint Petersburg State University*
1:35PM–1:55PM	T	**Local sequence alignment using intra-processor parallelism** Alexander Tiskin *University of Warwick*
1:55PM–2:15PM	T	**HEDGE: Highly accurate GPU-powered protein-protein docking pipeline** Timofei Ermak *Biocad*
2:15PM–2:35PM	T	**Probabilistic model of V-D junction formation in human Ig heavy chain genes and its application** Evgeny A. Bakin *Saint Petersburg State University, Bioinformatics Institute*
2:45PM–4:00PM	B	**Lunch**
4:00PM–5:00PM	I	**Connecting the Microbiome and Ecology to the Gut-Brain Axis** Rob Knight *UC San Diego*

5:00PM–5:20PM	T	**Reconstructing haplotype-specific cancer genome karyotypes with multiple sequencing technologies** *Sergey Aganezov* *Johns Hopkins University*
5:20PM–5:40PM	T	**In pursuit of molecular mechanism for induced granulocytic differentiation: systems biology approach** Svetlana Novikova *Institute of Biomedical Chemistry*
5:40PM–6:00PM	T	**Gene Set Mining In Context Relevant Pubmed Corpora** Christophe Van Neste *King Abdullah University of Science and Technology (KAUST), Ghent University*
6:00PM–6:15PM	O	**Closing remarks** Alla Lapidus *Saint Petersburg State University*
7:30PM–10:00PM	D	**VOGIS Evening Reception**

The conference was supported by the Russian Fund for Basic Research (grant 19-016-20002)

SATURDAY — JUNE 22

DAY 3 TALK SUMMARIES

A universal transcriptomic signature of age reveals the temporal scaling of *Caenorhabditis elegans* aging trajectories

*Andrei E. Tarkhov (Gero LLC, Moscow, Russia;
Skolkovo Institute of Science and Technology, Moskow, Russia)*
Ramani Alla (Central Arkansas Veterans Healthcare System, Research Service, Little Rock, AR, USA; Department of Geriatrics, Reynolds Institute on Aging, University of Arkansas for Medical Sciences, Little Rock, AR, USA)
Srinivas Ayyadevara (Central Arkansas Veterans Healthcare System, Research Service, Little Rock, AR, USA; Department of Geriatrics, Reynolds Institute on Aging, University of Arkansas for Medical Sciences, Little Rock, AR, USA)
*Mikhail Pyatnitskiy (Gero LLC, Moscow, Russia;
Institute of Biomedical Chemistry, Moscow, Russia)*
*Leonid I. Menshikov (Gero LLC, Moscow, Russia;
National Research Center "Kurchatov Institute", Moscow, Russia)*
Robert Shmookler Reis (Central Arkansas Veterans Healthcare System, Research Service, Little Rock, AR, USA; Department of Geriatrics, Reynolds Institute on Aging, University of Arkansas for Medical Sciences, Little Rock, AR, USA; Bioinformatics Program, University of Arkansas for Medical Sciences, and University of Arkansas at Little Rock, Little Rock, AR, USA)
*Peter O. Fedichev (Gero LLC, Moscow, Russia;
Moscow Institute of Physics and Technology, Russia)*

We collected 60 age-dependent transcriptomes for *C. elegans* strains including four exceptionally long-lived mutants (mean adult lifespan extended 2.2- to 9.4-fold) and three examples of lifespan-increasing RNAi treatments. Principal Component Analysis (PCA) reveals aging as a transcriptomic drift along a single direction, consistent across the vastly diverse biological conditions and coinciding with the first principal component, a hallmark of the criticality of the underlying gene regulatory network. We therefore expected that the organism's aging state could be characterized by a single number closely related to vitality deficit or biological age. The "aging trajectory", i.e., the dependence of the biological age on chronological age, is then a universal stochastic function modulated by the network stiffness; a macroscopic parameter reflecting the network topology and associated with the rate of aging. To corroborate this view, we used publicly available datasets to define a transcriptomic biomarker of age and observed that the rescaling of age by lifespan simultaneously brings together aging trajectories of transcription and survival curves. In accordance with the theoretical prediction, the limiting mortality value at the plateau agrees closely with the mortality rate doubling exponent estimated at the cross-over age near the average lifespan. Finally, we used the transcriptomic signature of age to identify possible life-extending drug compounds and successfully tested a handful of the top-ranking molecules in *C. elegans* survival assays and achieved up to a +30% extension of mean lifespan.

DASE-AG: conditional-specific differential alternative splicing events estimation method for around-gap regions

Kouki Yonezawa
Ryuhei Minei
Atsushi Ogura
(Nagahama Institute of Bio-Science and Technology, Japan)

Alternative splicing is a mechanism to generate more than one *mRNA* isoforms from a single locus, and it increases the genetic diversity during post-transcriptional gene regulation. Furthermore, alternative splicing is often differentially regulated across tissues and during development. It suggests that each splicing isoform may have specific spatial and temporal roles in life system.

We have developed the differential alternative splicing variants estimation method, DASE and DASE2. DASE2 uses FPKMs or TPMs as expression quantities. FPKMs and TPMs are read counts normalized with the lengths of transcripts. However, DASE2 had three problems in finding splicing events. First, splicing events involve gaps in some of the transcripts but DASE2 also considered a series of mismatched nucleotides as splicing events. Second, DASE2 tended to give consecutive gaps at 5’- and 3’-ends higher ranks than those at internal positions. Third, expression quantities of regions around gaps at internal positions are important for detecting splicing events but DASE2 treated expression quantities of whole transcripts. To find alternative splicing (AS) events, for example, intron retention, exon skipping and alternative splice sites, expression quantities of regions including gaps in some of variants and nucleotides in the others are required. We therefore developed DASE-AG for finding series of gaps with their flanking regions with different trends of expressions under the different condition as candidates of AS events. Alternative 5’- and 3’-splice sites found in *de novo* assembly tend to be more false-positive than skipped exons (SE), retained introns (RI) and mutually exclusive exons (MXE). Therefore, DASE-AG focuses only on series of gaps and their flanking nucleotides, called around-gap regions, and aims to comprehensively detect candidates of SE, RI and MXE.
We performed estimation of alternative splicing in different conditions of mice. DASE-AG considered one of the genes discussed to possibly have conditional-specific splicing variants, which was validated to have different expression tendencies of the variants under the two conditions using quantitative RT-PCR and semiquantitative RT-PCR. From this result we conclude that DASE has an ability to effectively detect conditional-specific AS variants.

Biotech from theory to reality: the journey of creating and marketing groundbreaking technology

Andrey Perfilyev (Atlas Biomed Group, Moskow, Russia)

- Atlas Biomed Group have been bringing genetic technologies to consumers and medical practices since 2014
- Connecting IT, bioinformatics, medical and health technologies is the main area of expertise of Atlas Biomed team
- More than 20,000 people in Europe and CIS countries became clients of Atlas Biomed Group using DNA and microbiome tests
- Atlas Biomed will share its experience of implementing bioinformatics for B2B and B2C markets

NPS: scoring and evaluating the statistical significance of peptidic natural product – spectrum matches

Azat Tagirdzhanov
Alexander Shlemov
Alexey Gurevich
(Center for Algorithmic Biotechnology, St. Petersburg University, Russia)

Motivation: Peptidic Natural Products (PNPs) are considered a promising compound class that has many applications in medicine. Recently developed mass spectrometry-based pipelines are transforming PNP discovery into a high-throughput technology. However, the current computational methods for PNP identification via database search of mass spectra are still in their infancy and could be substantially improved.

Results: Here we present NPS, a statistical learning-based approach for scoring PNP–spectrum matches. We incorporated NPS into two leading PNP discovery tools and benchmarked them on millions of natural product mass spectra. The results demonstrate more than 45% increase in the number of identified spectra and 20% more found PNPs at a false discovery rate of 1%.

Availability: NPS is available as a command line tool and as a web application at http://cab.spbu.ru/software/NPS

Local sequence alignment using intra-processor parallelism

Dmitry Orekhov (St. Petersburg University, Russia;
Bioinformatics Institute, St. Petersburg, Russia)
Alexander Tiskin (University of Warwick, GB)

Local alignment of DNA sequences is a fundamental problem of bioinformatics. Standard solutions include fast heuristic methods such as BLAST, as well as the more time-consuming exact methods. An efficient exact local alignment technique, based on a “sliding window” approach, was developed by a University of Warwick team, resulting in a number of biologically significant results. The efficiency of that implementation was achieved, in particular, by utilising low-level intra-processor parallelism. In recent years, commodity processor architecture has been developing rapidly, culminating with Intel’s AVX-512, an instruction set taking intra-processor parallelism to a new level of efficiency and sophistication, while also being surprisingly well-suited for speeding up the “seaweed combing”’ sequence alignment technique developed by the second author. In this talk, we present a prototype software tool that is, to our knowledge, the first sequence alignment software taking advantage of AVX-512 parallelism. Our tool allows one to produce semi-local alignments between short DNA fragments and long DNA strings, using seaweed combing and intra-processor parallelism to achieve competitive performance. In future, we plan to extend our implementation to a very fast exact local sequence aligner with “sliding window” functionality.

HEDGE: Highly accurate GPU-powered protein-protein docking pipeline

Timofei Ermak (BIOCAD, Russia)
Artem Shehovtsov (BIOCAD, Russia)
Pavel Yakovlev (BIOCAD, Russia)

Protein-protein interactions play key roles in living systems: cell signaling, immune system reactions, microelements transport and many other processes are based on protein-protein complexes functions. Thus, protein-protein complexes prediction is very important task especially in terms of drug discovery. For example, in silico optimization stages of antibody-based drug development process requires to solve the problem hundreds of times. To perform in silico optimization accurately the docking problem must be solved with high accuracy in short time ranges. But it is one of the hardest structural bioinformatics problems due to large solution space (possible molecules orientations), relatively big sizes of protein systems and infinite space of molecules conformations.

Recently we developed a high-performant tool for protein-protein docking called HEDGE. The pipeline of algorithms implemented can be briefly described as follows:

1) scanning translational solution space using FFT correlation theorem with energy-like correlation function;
2) clustering of solutions by RMSD as a distance metric using method based on spanning tree construction algorithm;
3) refinement of full complex structures with minimization of potential energy, Polak-Ribière-Polyak conjugate gradient method [1] is used to solve optimization problem. Optimization target is OPLS [2] force field, which was implemented and highly optimized on GPU as well as its analytical gradient.
4) Finally we rank solutions by change of Gibbs free energy (ΔG), which can be considered as the most accurate for ranking of predicted molecular complexes.

Each step of the pipeline above is well-parallelizable, so, we utilize the full power of GPUs (graphics processing units), that allows to scan huge solution space and select best with solid ΔG metric. Moreover, different rotations of molecules can be processed independently, therefore, multi-GPU mode is supported to scale linearly and achieve maximal performance on multi-GPU supercomputers.

Accuracy was tested on a subset of CAPRI [3] dataset showing about 50% of correct predictions. Time required for prediction of one complex in rigid mode (without minimization) is about 7 minutes on Tesla V100 GPU, while other well-known tools (e.g. PIPER [4]) require about 90 minutes on 8 CPUs. Flexible mode requires much more calculations and takes about 1.5 hours on Tesla V100.

References:

[1] Polak, Elijah, and Gerard Ribiere. “Note sur la convergence de méthodes de directions conjuguées” *Revue française d'informatique et de recherche opérationnelle*. Série rouge 3.16 (1969): 35–43.

[2] Robertson, Michael J., Julian Tirado-Rives, and William L. Jorgensen. “Improved peptide and protein torsional energetics with the OPLS-AA force field” *Journal of chemical theory and computation* 11.7 (2015): 3499–3509.
[3] Janin, Joel. “Welcome to CAPRI: a critical assessment of predicted interactions” *Proteins: Structure, Function, and Bioinformatics* 47.3 (2002): 257–257.
[4] Kozakov, Dima, et al. “PIPER: an FFT–based protein docking program with pairwise potentials” *Proteins: Structure, Function, and Bioinformatics* 65.2 (2006): 392–406.

Probabilistic model of V-D junction formation in human *Ig* heavy chain genes and its application

Evgeny A. Bakin (St. Petersburg University, Russia;
Bioinformatics Institute, St. Petersburg, Russia)
Elena A. Pazhenkova (St. Petersburg University)
Oksana V. Stanevich (First I. Pavlov State Medical University, St. Petersburg, Russia)

Immunoglobulins (*Ig*s) play a crucial role in the adaptive immune system. Igs are composed of polypeptide subunits: light and heavy chains. The latter contains a variable domain that is important for an antigen binding. The coding sequences for IG heavy chain are produced through a complex process, including VDJ recombination and somatic hypermutation (SHM). The latter masks initial segments, which complicates a precise sequence analysis of *Ig* genes in B-cells. Thus, in this research we focus on such a robust parameter of Ig genes sequence as the length of a V-D junction, which strongly influences antibodies affinity. As is known, this junction may be subject to an abnormal recombination, sometimes leading to autoreactivity and a subsequent lymphomagenesis (e.g. due to VH-replacement). Initially, a V-D junction consists of palindromic (p)-nucleotides (produced by a protein complex of *Ku70/Ku80* and Artemis) and non-templated (n)-nucleotides (added by a *TdT* protein), which further undergoes an impact of exo- and endonucleases.

For all the three stages of V-D junction maturation, we propose a simple, yet tractable probabilistic models resulting in a general model describing a distribution of V-D junction lengths in normal immunoglobulins. The parameters for the developed model were fitted by means of datasets obtained from healthy individuals, which are available in open databases such as GenBank and ENA. For this purpose, we have developed a pipeline containing the following steps:

1. *Ig* genes repertoire assembly (pRESTO);
2. clonal families detection and data decorrelation (Partis);
3. sequences demarcation and V-D junction extraction (IMGT HighV-QUEST);
4. fitting model parameters via maximum likelihood estimation (custom Python scripts).

The application of the chi-square test showed a consistency of the model with the processed sample. The trained model was further applied to datasets describing Ig genes sequences for individuals with various diseases. For the individuals with oncohematological diseases, or for those who were at risk to have such diseases, a statistically significant divergence with the model was detected. At the same time, no divergence was detected for all the other kinds of diseases. This experiment has shown that a V-D junction length distribution in Ig repertoire may be used as an indicator of the presence of pathological clones in a B-cell population. The possibility of the model application as an early predictor of various diseases presents a significant interest for further research.

Reconstructing haplotype-specific cancer genome karyotypes with multiple sequencing technologies

Sergey Aganezov (Johns Hopkins University, Baltimore, MD, USA)
Fritz J. Sedlazeck (Baylor College of Medicine, Houston, TX, USA)
Sara Goodwin (Cold Spring Harbor Laboratory, NY, USA)
Gayatri Arun (Cold Spring Harbor Laboratory, NY, USA)
Isac Lee (Johns Hopkins University, Baltimore, MD, USA)
Sam Kovaka (Johns Hopkins University, Baltimore, MD, USA)
Michael Kirsche (Johns Hopkins University, Baltimore, MD, USA)
Rachel Sherman (Johns Hopkins University, Baltimore, MD, USA)
Robert Wappel (Cold Spring Harbor Laboratory, NY, USA)
Melissa Kramer (Cold Spring Harbor Laboratory, NY, USA)
Karen Kostroff (Northwell Health, NY, USA)
David L. Spector (Cold Spring Harbor Laboratory, NY, USA)
Winston Timp (Johns Hopkins University, Baltimore, MD, USA)
Michael C. Schatz (Johns Hopkins University, Baltimore, MD, USA)
W. Richard McCombie (Cold Spring Harbor Laboratory, NY, USA)

Genomic instability is a hallmark of cancer, where somatic mutations accumulate from a single cell and result in a collection of cells containing distinct derived genomes that constitute a tumor. We focus on large genome rearrangements (>50bp) that change the chromosomal structure, as understanding these can lead to more targeted and effective treatment options as well as advance our understanding of cancer progression.

We recently performed whole genome sequencing of organoid-grown tumor and matching normal cells from two breast cancer patients using 10X/Illumina, PacBio, and Oxford Nanopore sequencing. For short reads we obtained an ~18x coverage and for long reads we obtained a ~45x coverage. We then compared the structural variations (SVs) identified using an ensemble of methods including: NAIBR, LongRanger, and GROCSV for barcoded 10X reads; Lumpy, SvABA, and Manta for paired-end short reads; and PBSV and Sniffles for both PacBio and Nanopore long reads. We demonstrate that long read SV inference is superior to short read SV inference across all SV sizes, both in quality and quantity. We also note, that our long read SV inference reveals hundreds of SVs affecting COSMIC11 gene regions which would have been missed with a traditional short read based analysis. For long read based SV inference we conducted a downsampling experiment to identify coverage levels at which a robust SV inference is achievable. Our results demonstrate that at ~30x coverage a precision of 0.9 and a recall of 0.8 is achieved, suggesting an economically attainable sequencing coverage threshold for future long read based SV analysis in cancer genomes.

With the individual SV calls resolved, we then utilize our new method RCK for reconstructing clone- and haplotype-specific cancer genomes karyotypes. RCK leverages segmented allele-specific segment copy number profiles and a consensus set of SVs to reconstruct haplotype-specific karyotypes for the observed cancer genome. We also leveraged long reads that span multiple SVs to introduce constraints on haplotypes of origin on segments, that are joined together by respective

SVs in observed cancer samples. RCK produces haplotype-specific cancer karyotype structures that are constrained by a reasonable somatic evolutionary model and admit a tenable structure of the rearranged linear chromosomes and circular double-minutes in observed cancer samples. RCK's inference allowed us to identify many more CNAs affecting COSMIC gene regions than we found by either of the state-of-the-art methods like TitanCNA or HATCHet, providing a more comprehensive view of the rearranged cancer genomes. Our analysis highlights that long-read sequencing allows for a more accurate and sensitive SV detection, and our inferred karyotypes present a more accurate representation of the observed mutated genomes. Together, this allows for a more precise analysis of genetic instability in cancer.

In pursuit of molecular mechanism for induced granulocytic differentiation: systems biology approach

Svetlana Novikova
Olga Tikhonova
Leonid Kurbatov
Igor Vakhrushev
Alexey Lupatov
Alisa Gisina
Tatiana Farafonova
Victor Zgoda
Alexander Archakov
(Institute of Biomedical Chemistry (IBMC), Moskow. Russia)

Differentiation is one of the pivotal biological processes occurring in a multicellular organism and its disturbance is often associated with the development of oncological diseases, including leukemia. HL-60 cell line is a convenient model for the study of induced granulocytic differentiation under treatment of various chemicals including all-trans-retinoic acid (ATRA), which underlies therapy of acute promyelocytic leukemia. To unravel of molecular mechanism of induced cell maturation we applied whole-genome transcriptome profiling and high-performance mass spectrometry combined with bioinformatic search for transcription factors and key regulators.

We obtained transcriptome- and proteome-based modeling pathways of ATRA-induced granulocytic differentiation. Applying selected reaction monitoring, we measured number of predicted molecules and revealed up-regulation of transcription factors HIC1 and CEBPB, and LYN kinase and down-regulation of key molecule PARP1. Regulatory networks reflect the activation of differentiation processes, down-regulation of proliferation, altered balance between the processes of survival and apoptosis in a *p53*-independent manner. The differentially expressed transcripts and proteins, predicted transcriptional factors, and key molecules may be considered as potential targets for differentiation therapy of acute myeloid leukemia.

Research of Russian National Science Academies for 2013–2020 years using the equipment of “Human Proteome” Core Facility (IBMC).

Gene set mining In context relevant Pubmed corpora

Christophe Van Neste (KAUST, Thuwal, Saudi Arabia; Ghent University, Belgium)
Adil Salhi (KAUST, Thuwal, Saudi Arabia)
Frank Speleman (Ghent University, Belgium)
Vladimir Bajic (KAUST, Thuwal, Saudi Arabia)

With gene set enrichment analysis, researchers aim to reduce the complexity of their gene-based biological datasets and get more easily interpretable findings as to the functionally relevant differences between experimental conditions. Many methods exist to assess the enrichment of gene sets and make ranked lists out of a collection of gene sets, but they all depend on the coherency of those gene sets in the first place. In general, gene sets are synthesized knowledge from different biological or experimental conditions (tissues, diseases, phenotypes). Only a subset of genes within a gene set might be of relevance for one specific experimental condition or research question. We have developed a literature gene set mining tool, that allows composing a gene set out of genes that are relevant to specific conditions and the research question at hand, by selecting a specific corpus of documents with which to establish the gene set through text mining. After this, the gene set enrichment for that specific set can be analyzed. Furthermore, we include analysis for historic auditing of the gene set. Historic auditing of a gene set allows researchers to see when a gene set became enriched — at a predefined threshold — throughout time in the research niche of their interest, showing the novelty strength of their latest experimental results. We present a specific example: metastasis-related genes for neuroblastoma. Neuroblastoma is a pediatric cancer with a heavy metastasis burden for high-risk patients. However, the type of metastasis is very specific for neuroblastoma and cannot be directly compared to adult metastasized cancers. We show the workflow of mining for the neuroblastoma related gene set of metastasis-relevant genes and analyze its enrichment in neuroblastoma experimental data. As a comparison, we then run a similar analysis on metastatic samples from breast cancer to illustrate the added value of research-specific gene set enrichment analysis. The gene set analysis tool is part of a broader text mining tool “sina” (search indexed nomenclature associations) that we are developing and is available at https://github.com/dicaso/sina.