Bioinformatics: from Algorithms to Applications 2018
Conference Schedule

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

MONDAY - JULY 16		
9:00AM–9:45AM	B	Registration
9:45AM–10:00AM	O	**Opening Ceremony** Alla Lapidus, *SPbU* Anton Korobeynikov, *SPbU*
10:00AM–11:00AM	I	***Tools to Link Human and Environmental Microbiomes for Health*** Rob Knight *UC San Diego*
11:00AM–11:20AM	B	**Break**
11:20AM–11:40AM	T	**Exploring the microbiome of primates using cell-free DNA** Mark Kowarsky *Stanford University*
11:40AM–12:00PM	T	**Soil microbiome as a keystone factor of soil genesis – the incite from the large-scale study of heterogeneous soil chronosequences** Elizaveta Pershina *All-Russia Research Institute for Agricultural Microbiology*
12:00PM–12:20PM	T	**QIIMP: Microbiome metadata made easy** Amanda Birmingham *University of California, San Diego*
12:20PM–12:40PM	T	**New algorithms and tools for large-scale sequence analysis of metagenomics data** Johannes Söding • Martin Steinegger *Max Planck Institute for Biophysical Chemistry*
12:40PM–2:00PM	B	**Lunch**
2:00PM–3:00PM	I	**Culture-free generation of highly contiguous microbial genomes from human and marine microbiomes** Ami Bhatt *Stanford University*
3:00PM–3:20PM	T	**Assembling a (diploid/polyploid) genome to perfection: the case of the bdelloid rotifer Adineta vaga** Jean-François Flot *Université libre de Bruxelles*

3:20PM–3:40PM	T	**Main results from the 3,000 rice genomes sequences** Nickolai Alexandrov *Inari Agriculture, Inc.*
3:40PM–4:00PM	T	**Protein storage in plant seeds is associated with the amyloid formation** Anton A. Nizhnikov *All-Russia Research Institute for Agricultural Microbiology, Saint-Petersburg*
4:00PM–4:20PM	B	**Break**
4:20PM–5:30PM	P	**Poster Session**
TUESDAY - JULY 17		
9:00AM–10:00AM	I	**Towards perfect de novo DNA assembly** Gene Myers *Max-Planck Institute for Molecular Cell Biology and Genetics*
10:00AM–10:20AM	T	**Bwise: a novel accurate, haplotype-specific genome assembler** Antoine Limasset *Université libre de Bruxelles*
10:20AM–10:40AM	B	**Break**
10:40AM–11:00AM	T	**Assembling barcoded RNA sequencing data** Andrey PrjibelskI *Center for Algorithmic Biotechnology, SPbU*
11:00AM–11:20AM	T	**BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters From Assembly Graphs** Dmitry Meleshko *Center for Algorithmic Biotechnology, SPbU*
11:20AM–11:40AM	T	**Plasmid detection and assembly in genomic and metagenomic datasets** Dmitry Antipov *Center for Algorithmic Biotechnology, SPbU*
11:40AM–12:00PM	T	**CellPi: unsupervised processing pipeline of mouse and human single-cell RNA-seq data** Alexey Samosyuk *Skoltech*
12:00PM–1:20PM	B	**Lunch**
1:20PM–2:20PM	I	**Test tubes, sequencing machines, computers: bioinformatics as a molecular biology tool** Mikhail S. Gelfand *Institute for Information Transmission Problems*
2:20PM–3:00PM	F	**Semantic-based antibody folding and structural annotation** Pavel Yakovlev *Biocad*

3:00PM–4:00PM	I	**CRISPR: fascinating biology and limitless applications** Eugene V. Koonin *National Institutes of Health*
4:00PM–4:20PM	B	**Break**
4:20PM–4:40PM	T	**ClinCNV: novel method for large-scale CNV and CNA discovery** German Demidov *Institute of Medical Genetics and Applied Genomics, Tübingen, Germany*
4:40PM–5:00PM	T	**Genome rearrangements in bacteria** Olga O Bochkareva • Pavel V Shelyakin *IITP RAS, Skoltech • VIGG RAS, Skoltech*
5:00PM–5:20PM	T	**Analysis and visualization of segmental duplications in mammalian genomes** Alla Mikheenko *Center for Algorithmic Biotechnology, SPbU*
6:00PM–9:00PM	B	**Dinner**
WEDNESDAY - JULY 18		
09:00AM–10:00AM	I	**Sequencing genome diversity in fish** Richard Durbin *Dept. of Genetics, University of Cambridge*
10:00AM–10:40AM	F	**The Genome Russia Project – 2018** Stephen J OBrien *Saint Petersburg State University*
10:40AM–11:00AM	B	**Break**
11:00AM–11:20AM	T	**A Rapid Exact Solution for the Guided Genome Aliquoting Problem** Maria Atamanova *ITMO University*
11:20AM–11:40AM	T	**Bounded-length Smith-Waterman alignment** Alexander Tiskin *University of Warwick*
11:40AM–12:00PM	T	**Reconstruction of a Set of Points from the Noise Multiset of Pairwise Distances in n^2 Steps for the Cyclic Sequencing Problem** Eduard Fomin *Institute of Cytology and Genetics SB RAS*
12:00PM–12:20PM	T	**Bayesian modelling of gene network alterations during tree-like processes: evolution or cells differentiation** Anna A. Igolkina *Peter the Great St.Petersburg Polytechnic University*
12:20PM–1:30PM	B	**Lunch**

1:30PM–2:30PM	I	**Discovering novel metabolisms via metagenomics** Ludmila Chistoserdova *Senior Scientist, University of Washington*
2:30PM–2:50PM	T	**Promoters and enhancers landscape of embryonic development and hibernation in chicken** Oleg Gusev *Kazan Federal University • RIKEN*
2:50PM–3:10PM	T	**Mathematical modeling of SNP %GC in microbial core genomes** Jon Bohlin *Norwegian Institute of Public Health*
3:10PM–3:30PM	B	**Break**
3:30PM–4:10PM	F	**Building time- and cost-effective bioinformatics pipelines in the Cloud - from bcl to visual analysis with New Genome Browser** *EPAM*
4:10PM–4:20PM	O	**Closing Remarks** Alla Lapidus
THURSDAY - JULY 19		
09:30AM–11:00AM	W	**QIIME Workshop** Rob Knight, Amanda Birmingham
11:00AM–11:30AM	B	**Break**
11:30AM–1:00PM	W	**QIIME Workshop** Rob Knight, Amanda Birmingham
1:00PM–2:00PM	B	**Lunch**

The conference was supported by the Russian Fund for Basic Research (grant 18-04-20035)

Conference Sponsors

BIOCAD
Biotechnology Company

ДИА•М
современная лаборатория

30
лет

АССОЦИАЦИЯ ВЫПУСКНИКОВ СПбГУ

25 YEARS
RFBR

Media Partners

Since 1991
Monomax
CONGRESSES & CONFERENCES

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.

MONDAY – JULY 16, DAY 1 SCHEDULE

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

9:00AM–9:45AM	B	Registration
9:45AM–10:00AM	O	**Opening Ceremony** Alla Lapidus, *SPbU* Anton Korobeynikov, *SPbU*
10:00AM–11:00AM	I	***Tools to Link Human and Environmental Microbiomes for Health*** Rob Knight *UC San Diego*
11:00AM–11:20AM	B	**Break**
11:20AM–11:40AM	T	**Exploring the microbiome of primates using cell-free DNA** Mark Kowarsky *Stanford University*
11:40AM–12:00PM	T	**Soil microbiome as a keystone factor of soil genesis – the incite from the large-scale study of heterogeneous soil chronosequences** Elizaveta Pershina *All-Russia Research Institute for Agricultural Microbiology*
12:00PM–12:20PM	T	**QIIMP: Microbiome metadata made easy** Amanda Birmingham *University of California, San Diego*
12:20PM–12:40PM	T	**New algorithms and tools for large-scale sequence analysis of metagenomics data** Johannes Söding • Martin Steinegger *Max Planck Institute for Biophysical Chemistry*
12:40PM–2:00PM	B	**Lunch**
2:00PM–3:00PM	I	**Culture-free generation of highly contiguous microbial genomes from human and marine microbiomes** Ami Bhatt *Stanford University*
3:00PM–3:20PM	T	**Assembling a (diploid/polyploid) genome to perfection: the case of the bdelloid rotifer Adineta vaga** Jean-François Flot *Université libre de Bruxelles*

3:20PM–3:40PM	T	**Main results from the 3,000 rice genomes sequences** Nickolai Alexandrov *Inari Agriculture, Inc.*
3:40PM–4:00PM	T	**Protein storage in plant seeds is associated with the amyloid formation** Anton A. Nizhnikov *All-Russia Research Institute for Agricultural Microbiology, Saint-Petersburg*
4:00PM–4:20PM	B	**Break**
4:20PM–5:30PM	P	**Poster Session**

The conference was supported by the Russian Fund for Basic Research (grant 18-04-20035)

MONDAY – JULY 16

DAY 1 TALK SUMMARIES

Tools to Link Human and Environmental Microbiomes for Health

Rob Knight (UC San Diego)

The rapid decline in cost of sequencing technology together with advances in computational techniques has led to the possibility of integrating microbial knowledge across spatial and temporal scales. In this talk, I describe approaches developed for the Human Microbiome Project that allow us to map microbes from birth to death and across the body. I also describe how these human-associated microbial communities relate to those in the environment. Finally, I show how we can integrate chemical and microbial mapping to understand systems like the cystic fibrosis lung, and, ultimately, to take control of our own gut microbiology to improve our health.

Exploring the microbiome of primates using cell-free DNA

Mark Kowarsky (Stanford University)
Iwijn De Vlaminck (Cornell University)
Jennifer Okamoto (Chan Zuckerberg Biohub)
Norma Neff (Chan Zuckerberg Biohub)
Nathan D Wolfe (Metabiota / Global Viral)
Stephen Quake (Chan Zuckerberg Biohub / Stanford)

Blood circulates throughout the body and contains molecules drawn from almost every tissue. What can we learn by studying the circulating nucleic acids in it? Using high-throughput, non-targeted shotgun sequencing of circulating cell-free DNA from plasma, besides sequences from the host, we detect those from: bacteria, archaea, eukaryotic parasites and viruses. After careful host subtraction and iterative stages of assembly and annotation there are thousands of microbial contigs over 1 kbp. The majority of these have predicted coding sequences, however most have low levels or no homology to existing sequences. The presence of the novel sequences was validated using independent sequencing experiments and direct PCR amplification. Known sequences support many prior observations of taxa present in primate microbiomes. The structure of the microbiome detected in blood is stable for a few months, and correlates with the environment more strongly than the host species, although viruses have a host-taxa association. Numerous potentially zoonotic taxa can be identified in an unbiased manner, and this together with the breadth of novel taxa, show that microbial diversity and the need to monitor environment reservoirs is higher than previously appreciated.

Soil microbiome as a keystone factor of soil genesis – the incite from the large-scale study of heterogeneous soil chronosequences

Elizaveta Pershina (All-Russia Research Institute for Agricultural Microbiology, Department of Microbiology, Saint-Petersburg State University)
Ekaterina Ivanova (All-Russia Research Institute for Agricultural Microbiology, VV Dokuchaev Soil Science Institute)
Evgeny Abakumov (Department of Applied Ecology, Saint-Petersburg State University)
Evgeny Andronov (All-Russia Research Institute for Agricultural Microbiology, Saint-Petersburg State University)

As soil is the most complex ecosystem in terms of microbial biodiversity, the models differentiating ecological factors at every stage of soil evolution, are strongly needed. The best models are soil chronosequences, where soil formation occur from the initial stage to the embryonic soil containing the horizons of zonal soil type. The study characterizes the microbiomes of the set of chronosequences in various climatic zones: mining sites, post pyrogenic soils, Antarctic moraines and soils of the coastal transgression. In all chronosequences the soil evolutionary stages were described, the replicated samples were taken from the entire soil profile. The amplicon libraries of the 16S rRNA gene were sequenced by use of ILLUMINA MiSeq platform. The quantitative approaches were performed to study the amount of bacteria, archaea and fungi in soil samples. Bioinformatics analysis included both the traditional and original methods. The profile analysis revealed a relationship between microbiome structure and soil genesis, especially the processes of decomposition and the ongoing mineralization of organic residues. It was also shown that the strength of the ecological factors determining the structure of microbiomes dependent on climatic zone. The main factors were soil pH and vegetation for temperate climate, water regime and the deglaciation for polar and semi-polar regions and the amount of mineralized organic matter for post-pyrogenic soils.

The study was funded by RSF №14-26-00094p (soil sampling) and RSF № 17-16-01030 (sequencing).

QIIMP: Microbiome metadata made easy

Amanda Birmingham (University of California, San Diego)

Drawing meaningful conclusions from even the best microbiome data is impossible without accurate and relevant metadata about the samples. However, researchers routinely struggle to record complete, consistent metadata that meet international minimum information standards. The Center for Microbiome Innovation has therefore developed QIIMP, the Quick and Intuitive Interactive Metadata Portal, which guides researchers through the generation of high-quality, standards-compliant metadata files. I will introduce QIIMP and demonstrate how it integrates with and improves upon existing metadata handling approaches.

New algorithms and tools for large-scale sequence analysis of metagenomics data

Martin Steinegger (Max Planck Institute for Biophysical Chemistry)
Johannes Söding (Max Planck Institute for Biophysical Chemistry)

Sequencing costs have dropped much faster than Moore's law in the past decade. The analysis of large metagenomic datasets and not its generation is the now the main time and cost bottleneck. We present three methods that together much alleviate the challenges posed by the exploding amount of metagenomics data and that allow us to go from an experiment-by experiment analysis to large-scale analyses of hundreds or thousands of datasets.

MMseqs2 [1] is a protein sequence and profile search method slightly more sensitive than PSI-BLAST and 400 times faster. MMseqs2 can annotate 1.1 billion sequences in 8.3 hours on 28 cores. MMseqs2 offers great potential to increase the fraction of annotatable (meta)genomic sequences. Linclust [2] is a sequence clustering method whose run time scales linearly with the input set size, not nearly quadratically as in conventional algorithms. It can cluster 1.6 billion metagenomic sequence fragments in 10 hours on a single server to 50% sequence identity, >1000 times faster than has been possible previously. PLASS (unpublished) is a metagenomic protein sequence assembler whose runtime and memory scale linearly with dataset size. It can assemble ten times more protein sequences from soil metagenomes, and faster than Megahit and other popular nucleotide-level assemblers.

[1] Steinegger M and Soeding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, doi: 10.1038/nbt.3988 (2017)
[2] Steinegger M and Soeding J. Clustering huge protein sequence sets in linear time. biorxiv, doi: 10.1101/104034 (2018) (accepted at Nature Communications)

Culture-free generation of highly contiguous microbial genomes from human and marine microbiomes

Ami Bhatt (Stanford University)

There are more than 1,000 species of bacteria, viruses and fungi that live in the human gut. Far from being passive passengers, these organisms strongly interact with host metabolism, the immune system, and more. For all of this interaction, the dynamics between human hosts and bacteria (microbiome) has only been explored in earnest for the last fifteen to twenty years. Compelling early experiments have shown that intestinal microbiome composition is associated with obesity, cardiovascular diseases, and the effectiveness of certain cancer chemotherapies. Therefore, understanding the impact of microbiomes speciation on noncommunicable diseases such as cancer, hematological and cardiometabolic disorders is fundamental to our health care. But how does one begin to model the dynamics of >1,000, mostly un-sequenced species and strains of bacteria, viruses and fungi? Our translational laboratory develops and applies novel molecular and computational tools to study strain level dynamics of the microbiome, to understand how microbial genomes change over time and predict the functional output of microbiomes. These innovations allow us to better (1) measure the types and functions of microbes in patients with non-communicable diseases, (2) iterate interaction models between microbial genes, gene products, and host cells and (3) test the impact of microbially targeted interventions in clinical trials.

Assembling a (diploid/polyploid) genome to perfection: the case of the bdelloid rotifer Adineta vaga

Jean-François Flot (Université libre de Bruxelles, Belgium)
Jitendra Narayan (Université de Namur, Belgium)
Lyam Baudry (Institut Pasteur, France)
Alessandro Derzelle (Université de Namur, Belgium)
Antoine Limasset (Université libre de Bruxelles, Belgium)
Romain Koszul (Institut Pasteur, France)
Karine Van Doninck (Université de Namur, Belgium)

Genome scientists commonly turn a blind eye to heterozygosity by aiming to reconstruct a non-redundant haploid genome. As a result of this methodological bias, most short-read assemblers available to date are incapable of resolving diploid or polyploid genomes. Theoretically, the use of third-generation sequencing reads of great lengths (such as PacBio and Nanopore reads) and/or the long-distance information provided by chromosome conformation capture (3C) should solve the problem of diploid/polyploid genome assembly and produce chromosome-scale, haplotype-specific assemblies, but fully resolved heterozygous genomes are still extremely rare in the literature.

As a test of the potential of these approaches to deliver "perfect" assemblies, we turned to the reasonably sized genome of the bdelloid rotifer Adineta vaga (expected size: 244 Mb). Bdelloid rotifers are famous for their tens of million years of evolution in the apparent absence of meiotic sex (only females have ever been observed, and they produce eggs clonally via mitotic parthenogenesis). Such a long evolution without recombination is expected to result in a complex genome structure replete with palindrome and colinearity breakpoints, an hypothesis that can only be tested by assembling separately all haplotypes. In 2013, we published a first diploid (actually, tetraploid) draft genome of Adineta vaga, the assembly of which from 21X 454 reads took six weeks of computation on a 64-core, 256-Gb RAM server using MIRA; final N50s was 47 kb for contig and 260 kb for scaffolds. To finish this genome and produce a fully resolved diploid assembly, we have now generated 100X 2*250 bp Illumina paired-end reads, 100X PacBio, 100X Nanopore and 150X 3C data from the same clonal lineage. Despite this plethora of data and our use of a panel of state-of-the-art approaches (including custom-developed ones), generating a perfect, telomere-to-telomere assembly turned out to be more difficult than expected and required a significant amount of manual refinements. We are now developing novel tools to automatize and streamline these approaches, with the aim of making perfect assemblies the norm rather than the exception.

Main results from the 3,000 rice genomes sequences

Nickolai Alexandrov (Inari Agriculture, Inc.)

Analysis of the 3,000 rice genomes, sequenced by CAAS, BGI and IRRI, revealed unprecedented amount of genome diversity in Oryza Sativa, defined population structure with remarkable precision, identified genes and genome regions with unusual conservation and discovered haplotype structure of domestication genes. Combining sequence data with available phenotypic information we were able to find new trait-loci associations and to confirm previously known associations. Our portal SNP-Seek is commonly used for allele mining and visualization of genome variations.

Protein storage in plant seeds is associated with the amyloid formation

Anton A. Nizhnikov (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Maria E. Belousova (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
MV Belousov (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
OY Shtark (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
AO Kosolapova (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Kirill S. Antonets (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)

Amyloids are protein fibrils exhibiting ordered spatial structure known as cross-β. Such a structure gives amyloids extreme resistance to different chemical and physical influences. Historically, amyloids were generally considered as the lethal pathogens causing dozens of incurable disorders in humans and animals but recently amyloids became clear to be essential functional quaternary protein structures involved in various biological functions in archaea, bacteria and eukaryotes. Despite their social importance, plants remain only large group of multicellular organisms where amyloids were not found. We performed a large-scale bioinformatics analysis of the distribution of potentially amyloidogenic regions in the proteomes of 75 species of the lands plants that comprised about 2.9 million of proteins. Using two bioinformatics tools, Waltz and SARP, we demonstrated that potentially amyloidogenic proteins are widespread in the proteomes of plants with their number corresponded to the number of amyloidogenic proteins in the proteomes of organisms in which amyloids were previously identified, like humans and different species of fungi. Amyloidogenic proteins of plants tended to be associated with different biological processes and functions including defense from pathogens, transmembrane transport and protein storage in seeds. In-depth analysis of the association between amyloidogenic properties and seed storage function of such proteins was done. We found that seed storage proteins comprising conservative β-barrel domain Cupin-1, mainly 7S and 11S globulins, are rich in amyloidogenic regions in the most of land plant species. So, 302 storage protein with Cupin-1 domain belonging to 54 of 75 species analyzed contained amyloidogenic regions. In addition, we identified 119 seed storage proteins with Zein domain, 121 proteins with Gliadin domain, 13 with Vicilin domain and 7 proteins with high molecular weight Glutenin that were found to be amyloidogenic. Experimental analysis performed with several storage proteins or their regions confirmed amyloid properties including formation of unbranched fibrils, binding amyloid-specific dyes, and formation of detergent-resistant aggregates. Based on these data we conclude that amyloid formation by seed storage proteins represent a novel molecular mechanism for long-term stabilization of such proteins to avoid their degradation and misfolding during natural dehydration and unfavorable environmental conditions.

This work was supported by the Russian Science Foundation (Grant No 17-16-01100).

POSTER SESSION

Analysis of metagenome state after treatment of dysbiosis with autoprobiotics

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

Diseases and stress often perturb metagenome state resulting in dysbiosis. Treatment of dysbiosis can improve organism state and contribute to recovery. Probiotics are microorganisms that are claimed to provide health benefits when consumed. Autoprobiotics are personified therapeutical products containing selected individual indigenes strains. We tested effects of autoprobiotics treatment on microbiome recovery after induced dysbiosis. Examined probiotics included lactobacilli, enterococci, bifidobacteria, feces diluted 100 fold, bacteria grown on feces in anaerobic conditions and mix of lactobacilli, enterococci and bifidobacteria. Dysbiosis was induced by 3 day treatment with ampicillin and metranidazol. Experimental groups included rats treated only with antibiotics (3ab group), rats treated with antibiotics and then with probiotics for 5 days (groups lb, en, bi, fe, an and mix), rats treated with antibiotics and then with normal saline for 5 days (8ab group) and rats treated only with normal saline for all 8 days of experiment (k group). 16S rRNA variable segments V3 and V4 were sequences for all fecal samples. 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. Taxonomic abundances profiles, PCA and unweighted Unifrac distances showed that all autoprobiotics tested shifted metagenome state closer to control group. Samples from animals, treated only with antibiotics (3ab) were most distant from control group. Samples from animals, which were treated with normal saline after induced dysbiosis (8ab) were at a greater distance from control group and closer to dysbiotic group than samples from groups treated with autoprobiotics. These results show that autoprobiotics can accelerate microbiome restoration after dysbiosis and should be a promising target for future research.

Integration and automatic assembly of metabolic mathematical models

Timofei Ermak (BIOCAD / Institute of cell biophysics RAS)
Artem Ryasik (Institute of cell biophysics RAS)
Michail Orlov (Institute of cell biophysics RAS)
Evgenia Zykova (Institute of cell biophysics RAS)
Anatoly Sorokin (Institute of cell biophysics RAS)

Understanding of structure and dynamics of metabolic networks is an essential step in modern biological and medical studies, such as the design of superproducer strains for biotechnology, identification of biomarkers and therapy targets of complex diseases, etc.
In silico approaches have been part of metabolic engineering toolkit since the mid-90s, allowing the costs and the time reduction by hypothesis testing and design of experiments. The performance of such approaches is heavily relying on the quality of the reaction network reconstruction, which by itself is the complicated and time-consuming process. More than a decade of active development of high-quality whole genome scale (WGS) models for prokaryotic species open a way for automatic reconstruction of WGS models based on genomic annotation and proteins homology data. The aim of the work is the development of tools and methods for integration and headless assembly of metabolic networks flux models.

Authors recently developed the BioGraph database, a graph-oriented storage for information about prokaryotic organisms. It contains various -omics data (genomic, proteomic, taxonomic) integrated between each other with graph representation. The properly organized network of structural and semantic similarity relationships between proteins, genes, organisms and reactions makes it possible to construct models for close strains and species using data from manually curated models. Models are generated as the file in SBML format, suitable for validation with COBRApy [1] package. The developed tool was tested on the database with several strains of E. coli, the template model was assembled for E. coli str. K-12 substr. W3110 from data retrieved from two other E. coli models: iML1515 [2] for E. coli str. K-12 substr. MG1655 and iECW_1372 [3] for E. coli W. The assembled template model carry biomass reaction flux, thus, we consider the approach working properly.

We developed tools that allow assemble of metabolic flux models in headless manner for close bacterial species from different -omics data such as genomic annotation or proteins and genes similarity. The tools were tested on the assembly of the model for E. coli str. K-12 substr. W3110. Also the BioGraph database could be used as a source of biological information integrated with graph representation.
Source code of the software is available on Github: https://github.com/arc7an/scalaBiomeDB

Development of a bioinformatics pipeline for routine analysis of whole genome sequencing data of M. tuberculosis complex members

Bert Bogaerts (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium, UGhent – Department of information Technology, IDLab, imec)
Raf Winand (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)
Qiang Fu (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)
Julien van Braekel (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)
Pieter-Jan Ceyssens (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)
Lorenzo Subissi (Bacterial Diseases, Sciensano, Brussels, Belgium)
Vanessa Mathys (Bacterial Diseases, Sciensano, Brussels, Belgium)
Sophie Bertrand (Bacterial Diseases, Sciensano, Brussels, Belgium)
Sigrid C. J. De Keersmaecker (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)
Nancy Roosens (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)
Kathleen Marchal (UGhent – Department of information Technology, IDLab, imec)
Kevin Vanneste (Platform Biotechnology and Bioinformatics, Sciensano, Brussels, Belgium)

The adaptation of whole genome sequencing (WGS) and bioinformatics for routine molecular typing and pathogen characterization in a public health setting remains problematic, which is partly due to the lack of user-friendly and validated data analysis tools that can be used for routine typing in the National Reference Centers (NRCs) and peripheral laboratories. In collaboration with the Belgian NRC Mycobacteria, we developed a pipeline for the routine analysis of Mycobacterium tuberculosis complex isolates that was specifically designed to tackle the aforementioned challenges. The push-button pipeline is executed through a user-friendly interface in Galaxy and uses Illumina WGS data to characterize M. tuberculosis complex isolates. The pipeline performs automated data processing and quality control of sequencing data before several bioinformatics assays are executed: (sub-)species identification based on 16S rDNA and the detection of specific markers (both loci such as the regions of difference RD1 and RD9 and lineage-associated SNPs that delineate distinct SNP cluster groups); resistance characterization based on the detection of SNPs associated with resistance to specific antibiotics; sequence typing using currently available MLST and cgMLST schemes in PubMLST.org; spoligotype determination using the SpoTyping tool. The pipeline performance is currently being characterized by means of a set of performance metrics and definitions that were specifically adapted towards bioinformatics assays, and which evaluate precision (repeatability and reproducibility), accuracy, sensitivity, and specificity. Preliminary results on a representative set of samples show high performance, indicating the feasibility of using WGS in routine public health settings to replace classically employed pathogen typing and characterization techniques. Similar pipelines can be developed for other pathogens and case studies, making bioinformatics analyses less complex and more time-efficient for both expert and non-expert users.

Modeling of the DNA-nanodevices for inactivation of Influenza A Virus

Alexander A. Spelkov (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation)
Yaroslav V. Solovev (ITMO University, St. Petersburg, Russian Federation)
Ekaterina A. Bryushkova (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation)
Dmitry M. Kolpashchikov (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation; Chemistry Department, University of Central Florida, United States)

According to the World Health Organization (WHO), viral infections are one of the leading cause of human death. Only with Influenza virus A (IAV) annually infects 3 to 5 million people resulting in 250 to 500 thousands deaths worldwide. Antiviral drug development resulted in the development of several pharmaceuticals. However, the major limitations of antiviral drugs include a narrow spectrum of action, development of drug-resistant mutants, toxic side effects and inefficiency towards latent diseases.

Cleavage of IAV RNA by deoxyribozyme-based DNA-nanodevices (DNA nanorobots) is one possible approach for inactivation of the virus inside infected cells. Deoxyribozymes (Dz) are synthetic single-stranded DNA molecules, with a range of catalytic activities including RNA-cleaving. A number of Dz demonstrated promising results in vitro, but never have been used in medical practice. Among major obstacles on the way of converting the DZ-based technology to anti-IAV medicine is the inaccessibility of viral RNA target for Dz recognition for efficient cleavage. In addition, high viral mutagenesis causes a decrease in the efficiency of the gene silencing drugs.

The goal of the study is to develop a DNA nanorobot that will efficiently recognize and cleave IAV RNA. IAV belongs to the family of Orthomyxoviridae, whose genetic information is coded by 8 single-stranded RNA segments. A segment chosen as a target encodes a PB-1 subunit of RNA-polymerase, which plays a pivotal role in replication and transcription of viral genes. We designed and tested several DNA nanorobots which can efficiently bind a conservative among IAV spices RNA fragment and cleave it under near physiological conditions.

First we used Bioinformatics to find IAV target sequence conservative among all IAV major pathogenic IAV spices and at the same time absent in human transcriptome With the help of bioinformatics, we designed a DNA nanorobot containing 2 DZ cores, which was able to cleave IAV RNA fragment with the efficiency 1.15 times greater than that of a traditionally designed DZ cleaving agent. Importantly, the cleavage enhancement effect was not observed with short linear RNA substrates. The greater efficiency of the design is attributed to the cooperative binding of the targeted RNA by the 2 DZ, which results in efficient unwinding folded RNA structure. We conclude that a solution for RNA accessibility problem and cleavage under near physiological conditions was mitigated by this study. The design strategies developed in this study can become a basis for the future development of universal antiviral therapeutic agents.

This work was supported by RFBR grant # 18-34-00898 mol_a

Semi-supervised peak calling solution

Oleg Shpynov (JetBrains Research,
Department of Pathology & Immunology Washington University in St.Louis)
Roman Chernyatchik (JetBrains Research)
Aleksei Dievskii (JetBrains Research)
Petr Tsurinov (JetBrains Research)
Evgeny Kurbatsky (JetBrains Research)
Maxim N. Artyomov (Department of Pathology & Immunology Washington University in St.Louis)

The study "Multiomics dissection of healthy human aging" aims to comprehensively characterize changes happening in the distinct human cell type and its environment during the process of healthy aging. We performed Ultra Low Input ChIP-Seq for 5 major chromatin modifications (H3k27ac, H3k27me3, H3k36me3, H3k4me1, H3k4me3) in CD14+CD16- classical monocytes purified from blood of 20 young and 20 old donors. The comparative epigenetic study of this scale has never been undertaken previously in the context of human aging. While ULI ChIP-Seq protocol generally allows for robust peak calling, it is considerably more variable than conventional approach. Accurately dissecting the situation when both background and the signal can vary is generally prohibitively complex task for the unbiased peak calling approaches.

We propose a novel semi-supervised peak calling solution to make this task feasible. Effective semi-supervised peaks analyzer (SPAN) is capable processing both conventional and ULI ChIP-Seq tracks. Dedicated visualisation tool (JBR Genome Browser) is created to overcome one of the major challenges of the semi-supervised learning - the procedural complexity of the manual annotation of the data, which often leads to the inaccuracies and mix-ups. These tools provide readily accessible integrated peak annotation and peak calling capabilities.

Selection of RNA targeted fragments for anti-cancer and anti-viral theranostics by DNA nanorobots

Ekaterina A. Bryushkova (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation)

Daria D. Nedorezova (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation)

Daria V. Nemirich (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation)

Tatiana A. Lyalina (Laboratory of Solution Chemistry of Advanced Materials and Technologies, ITMO University, St. Petersburg, Russian Federation)

Erik Gandalipov (Department of IT in the Fuel and Energy Industry, ITMO University, St. Petersburg, Russian Federation)

The 2016 Nobel Prize in Chemistry was awarded for "the development and synthesis of molecular machines". A growing number of published studies are devoted to the development of effective, selective and biocompatible nanostructures with distinct biological activities. For example, deoxyribozymes (Dz) demonstrate high affinity and can selectively bind and cleave RNA sequences. This technology can be used for inactivation of RNA viruses or for suppression cancer cells. Nevertheless, technology based on Dz are limited by their low activity under physiological Mg2+ conditions, insufficient selectivity for the targeted RNA, and the inability of Dzs to cleave fragments involved in stable secondary RNA structures.

In this study, we combined several specialized functions, each of which addresses one of the aforementioned problems on a DNA scaffold to obtain a complex DNA nanostructure for efficient and selective cleavage of folded RNA named here DNA nanorobot. DNA nanorobot consists of several DNA fragments connected to each other by complementary Watson-Crick base pairs. The RNA-cleaving activity is provided by one or more Dz connected to the DNA scaffold. RNA-binding sequences (arms) unwind RNA secondary structure, while binary sensor provide high selectivity of target recognition. This design allows achieving the maximum efficiency of Dzs, since fixing RNA near to the catalytic site reduces the Michaelis constant of the RNA cleavage reaction.

The selection of optimal sequences for DNA nanorobots is based on stereometric and thermodynamic parameters, which calculated by bioinformatics methods. At the same time, finding optimal targeted sequences is also important. Three main points should be considered:

1) The selected fragment of the target nucleic acid sequence should be placed in a conserved site, especially for targets with highly variable genome sequences (e.g., influenza A virus). This problem is solved by methods of multiple sequence alignment.

2) It is necessary to consider the secondary structure of the target regions of selected nucleic acids (loops, hairpins, pseudoknots etc.), because it affects the energy of hybridization between the target and elements of DNA nanorobot. If the ΔG of folded structure is lower than the ΔG of the DNA nanorobot – RNA target complex, there will be a weakening of the interaction between the robot and the target.

3) For gene silencing it is important to choose exons as targets for DNA nanorobots. That requires work with genomic databases to exclude the sequence of introns from the list of potential targets.

In our study we focused on three main projects: (i) theranostics (therapy and diagnostics) of the influenza A virus; (ii) theranostics of the human papillomavirus (HPV) and (iii) cancer therapy by silencing of housekeeping genes in cancer cells.

Each one of the projects has unique features of both the DNA nanorobot design and its target. Project (i) aimed at developing a DNA naorobot for simultaneous cleavage of the viral RNA and reporting its presence by producing fluorescent signal only in the presence of viral RNA. The robot for project (ii) was designed to cleave double-stranded HPV DNA. The design of DNA nanorobot (iii) allows the cleavage of the mRNA of DAD1 gene only in the presence of a cancer marker sequence, which is N-Myc gene.

Summarizing, DNA nanorobots are modular machines which could be easily programmed, designed and redesigned for different applications. For this reasons DNA nanorobots can be considered as a promising agents for therapy and theranostics.

The work was supported by RFBR grant # 18-34-00898 mol_a

Acknowledgements: we thank Professor Dmitry M. Kolpashchikov from University of Central Florida for coordination and leadership of this work.

Link between psychiatric disorders and dominant-submissive behavior: insights from genomics study

Dmity Rodin (Ariel University)
Valeria Kogan (Ariel University)
Moshe Gross (Ariel University)
Albert Pinhasov (Ariel University)

Like humans, rodents develop social hierarchy wherein each individual's standing is determined by competition for territory, food and mating partners, forming relationships of dominance and submissiveness between conspecifics. We used selective breeding to develop distinct mouse populations characterized by features of dominance (Dom) or submissiveness (Sub), representing opposite extremes of the behavioral spectrum. Dom mice display remarkable resilience to several forms of environmental stress, while their Sub counterparts are prone to develop anxiety- and depressive-like behaviors in face of hardship. The respective resilience of Dom mice, as well as the vulnerability of their Sub counterparts, have been determined to depend upon divergent regulation of the Hypothalamic-pituitary-adrenal (HPA) axis, beginning in in utero prenatal programming of limbic regions regulating the stress response into adulthood. We therefore hypothesized that: 1) selectively bred Dom and Sub mice must possess distinct genomic profiles, and 2) genetic variation between Dom and Sub populations may be shared with clinical populations suffering from psychiatric disorders. Thus, this unique mouse model may enable preclinical intervention studies for pinpointing the functional mechanisms responsible for psychiatric disorders and identification of genetic variations associated with psychiatric disorders among Dom and Sub mice.

For these goals, we made use of Whole Genome Sequencing (WGS) to interrogate genetic variations between Dom and Sub mice. gDNA samples underwent library preparation using the Illumina Nextera XT kit in 300 cycles and were sequenced on the NextSeq 500 in two duplicate runs, four lanes per run, each run as 4 samples pooled into one library.

One technical obstacle we encountered was the lack of a reference genome for the Sabra background strain from which Dom and Sub mice were derived. To solve this issue, we identified conserved portions of the mouse genome shared by Sabra mice and reference the strain C57black/6. Paired-end reads were further mapped to the conserved genome only, upon which single nucleotide polymorphisms (SNPs) were called and short Insertions-Deletions (Indels) identified. Next, we identified those SNPs located on coding genes or regulatory regions (e.g. transcription factor binding sites, promotor regions). Candidate genetic variations were further screened against human data on certain psychiatric conditions from the Psychiatric Genomics Consortium for their conservation between mouse and human psychiatric cohorts.

Results: Analysis of WGS data from representative Dom and Sub mice identified nearly one million SNPs differentiating between strains, which are distributed within approximately 3,000 distinct protein-coding sequences. 77,000 insertions and deletions (InDels) of at least 20 bp were identified in 680 genes. Subsequent gene ontology analysis identified SNPs and InDels in functionally relevant genes interacting directly with the Glucocorticoid Receptor. Cross-referencing of genetic variations between Dom and Sub mice with human psychiatric databases identified a number of candidate risk alleles.

Thus, the identified genes may serve as biomarkers for the early diagnosis of individuals vulnerable to stress, improving the clinical outcomes of treatment for stress-induced disorders.

The genome wide analysis of the large tandem repeats in the closely related genomes

Dmitrii Ostromyshenskii (Institute of Cytology of the Russian Academy of Science)
Olga Podgornaya (Institute of Cytology of the Russian Academy of Science)

Large tandemly repeated sequences (TR, or satellite DNA) are necessary part of higher eukaryotes genomes and can comprise up to tens percent of the genomes. Much of TRs' functional nature in any genome remains enigmatic because there are only few tools available for dissecting and elucidating the TR functions.

The modified pipeline of the one used previously in our Lab [1] applied to the several databases. Four mammalian genera was used: (1) mice Mus: M.musculus, M.caroli (unassembled genome); (2) guinea pigs Cavia: C. porcellus, C. apperea; (3) bats Myotis: M.brandii, M.davidii, M.lucifugus; (4) cows Bos: B.taurus, B.mutus, B.indicus.

We tried to find all the 62 M.musculus TR families [1] in raw reads of M.caroli genome (Caroli Genome Project, PRJEB2188). There are only few TR of M. musculus in M. caroli genome. M.musculus major satellite occupied nearly 0,7% of M.caroli genome, while in M. musculus genome - ~ 11 %. In M.caroli genome we found 5 other M.musculus's TR's families.

Genus Cavia. C.porcellus genome possesses 25 TR and C.apperea – only 10 TR. 9 out of 10 C. apperea TR's family exist also in C. porcellus genome except the major TR for this species – Capp-1518. In C.porcellus genome there are two major TR – Cpor-783 is absent in the 2nd genome and Cpor-123 exists in C.apperea genome as the minor one. Genus Myotis. There is no any TR of Myotis in Repbase, but 133 TR's families are found in M.brandtii genome, 105 - in M. davidii genome and 26 - in M.lucifugus genome. Only 5 TR families exist in three genome but most of TR families are species-specific. Major TR for M. davidii and M.lucifugus is common in sequence though differ in monomer length, but the same TR is minor one in M. brandtii. The major for M.brandtii is not identified in both other genomes at all.

Genus Bos. There are three TR known for Bos in Repbase and all of them are found in all Bos assemblies. Still the major TR in all Bos assemblies differ: in B.taurus genome BTSAT4/BTSAT5 is a major TR while BTSAT6 major TR family in B.indicus genome. It is visible that most of the top TR families in genus Bos exist only in two genomes or even in one, i.e. is species-specific.

The most exhausting analysis of major TR (one for each species) of ~300 animals and plants display no readily apparent conserved characteristics [2]. We compared the TR sets. Our data evidenced that there are species-specific top TR, which are absent in genome of closely related species. In all genera examined major TRs are species-specific and hardly exist in other species of genera even as a minor ones.

This work was supported by the granting program 'Molecular and cell biology' of the Russian Academy of Sciences (grant № 01.2.01457147).

1. A.S. Komissarov et al. (2011). BMC genomics, 12(1):531-552.
2. D.P. Melters et al (2013). Genome Biol, 14(1), R10.

Conversion between Red Queen and Red King strategies in a general model of mutualistic symbiosis

M.A.Babenko (All-Russia Research Institute for Agricultural Microbiology)
A.A.Igolkina (All-Russia Research Institute for Agricultural Microbiology; Peter the Great St.Petersburg Polytechnic University, Mathematical Biology and Bioinformatics Laboratory)
E.E.Andronov (All-Russia Research Institute for Agricultural Microbiology)

The Red Queen and Red King effects are essential concepts in species coevolution, and both can be observed in mutualistic interactions between two species. The Red Queen effect argues that faster-evolving species are favored in the mutualism, while the Red King suggests slower-evolving strategy as beneficial. Here we propose a new determinant which triggers between the Red Queen and King and pretends to be the most general one - the degree of similarity between population structures of the symbionts, to which we will further refer in the context of topological beta-diversity.

We have developed a model to simulate the co-evolutionary dynamics in the plant-bacteria symbiosis. For this purpose, we considered populations of two species, which individuals are represented by vectors in different Euclidean spaces and characterized by different mutation rates. In order to generalize the model, we define an arbitrary rule of an interaction between two individuals from different species.

The simulation of mutualism consists of epochs; in each epoch, a plant individual interacts with a limited number of best-matched bacteria individuals from randomly selected bacterial subpopulation. Symbiotically efficient bacteria propagate more successfully than free-living bacteria, while a plant propagates in agreement with matching to its symbionts in the current epoch. Before the next epoch, abundances of plant and bacteria populations are normalized to the fixed number.

We performed the simulation for 1000 plant and 100000 bacteria individuals, starting from the state when these species did not interact. When two species began to form a symbiosis, i.e. they started to establish an adaptive landscape to each other, the Red Queen effect was observed and remained while the population diversities of species were not well matched. When similarity between population structures increased enough, the Red King was established. Thus, the Red monarch of strategy in mutualism depends on the degree of congruence between population structures of symbionts.

Improving metagenomic assembly using Nanopore Read-Until technology

Sergey Kazakov (ITMO University, Saint-Petersburg, Russia)
Vladimir Ulyantsev (ITMO University, Saint-Petersburg, Russia)
Sergey Nurk (SPbSU, Saint-Petersburg, Russia)

Since the emergence of high-throughput sequencing technologies, a huge volume of data has been accumulated describing complex microbial communities (microbiota). Several ways to analyze such data exist, one of them – to assemble all data together with further analyzing of assembled genomes one by one. Despite of the popularity of such techniques, the assembly itself still remains challenging.

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 in last few years. 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 reading current DNA molecule during the process itself and go to the next one. Among many incredible things, it can significantly reduce effective cost of assembly projects!

In current work we proposed several strategies how to use this ability to improve metagenome assembly. In more detail, we assume that Illumina reads are also available for studied microbiota. Using such fragmented assembly as a reference, it is possible to allow only such Nanopore reads, which connect two or more contigs of initial assembly. Selecting only such reads, we showed that we can improve metagenome assembly for up to 1.9x times better than using Nanopore sequencing without any strategy.

Also we showed 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 and 2.1x for useful length.

16s rRNA Detection by Using Neural Networks

Semyon Grigorev (Saint Petersburg State University)
Polina Lunina (Saint Petersburg State University)

Algorithms that can efficiently and accurately identify and classify bacterial taxonomic hierarchy have become a focus in computational genetics. The idea that secondary structure of genomic sequences is sufficient for solving the detection and classification problems lies at the heart of many tools. The secondary structure can be specified in terms of formal grammars.
The sequences obtained from the real bacteria usually contain a huge number of mutations and "noise" which renders precise methods impractical. Probabilistic grammars and covariance models (CMs) are a way to take the noise into account. For example, CMs are successfully used in the Infernal tool. Neural networks is another way to deal with ``noisy'' data. Some works utilize neural networks for 16s rRNA processing and demonstrate promising results.

We combine neural networks and ordinary context-free grammars to detect genomic sequences. We extract features by using the ordinary (not probabilistic) context-free grammar and use the dense neural network for features processing. Features can be extracted by any parsing algorithm and then presented as a boolean matrix such that the cell (i,j) contains 1 iff the substring from the i-th to j-th position is derivable in the input grammar.

We evaluate the proposed approach for 16s rRNA detection. We specify context-free grammars which detect stems with the hight of more than two pairs and their arbitrary compositions.
For network training we use dataset consisting of two parts: random subsequences of 16s rRNA sequences from the Green Genes database form positive examples, while the negative examples are random subsequences of full genes from the NCBI database. All sequences have the length of 512 symbols, totally up to 310000 sequences. After training, current accuracy is 90% for validation set (up to 81000 sequences), thus we conclude that our approach is applicable.

The presented is a work in progress. The ongoing experiment is finding all instances of 16s rRNA in full genomes. Also we plan to use the proposed approach for the filtration of chimeric sequences and the classification. Composition of our approach with other methods and tools as well as grammar tuning and detailed performance evaluation may improve the applicability for the real data processing.

Comparative analysis of Streptococcus genomes

Pavel V Shelyakin (VIGG RAS, IITP RAS, Skoltech)
Olga O Bochkareva (IITP RAS, Skoltech)
Anna A Karan (FBB MSU)
Mikhail S Gelfand (IITP RAS, Skoltech)

Genome sequencing of multiple strains of the same bacterial species shows that even closely related strains can significantly differ in gene content, and that genome of a single strain cannot serve as a good description of the species as a whole, because up to 25-30% of genes in one strain can lack orthologues in another. Therefore, for describing a species (or a higher taxonomic unit), the concept of a pan-genome was created, that is, the complete set of genes observed in a given taxonomic unit. A pan-genome consists of 3 fractions: core genes present in all strains of the species, periphery genes present in a subset of strains, and strain-specific or unique genes. These fractions are enriched with genes of different functions and are assumed to evolve under different modes.

In this work, we analyzed the structure and dynamics of the pan-genome of 3 species from the genus Streptococcus: S.pyogenes, S.pneumoniae and S.suis.

We showed that pan-genome fractions differ in size, functional composition, the level of nucleotide substitutions, and predisposition to horizontal gene transfer and genomic rearrangements. The density of substitutions in intergenic regions appears to be correlated with selection acting on adjacent genes, implying that more conserved genes tend to have more conserved regulatory regions. The total pan-genome of the genus is open, but only due to strain-specific genes, whereas other pan-genome fractions reach saturation. The strain-specific fraction is enriched with mobile elements and hypothetical proteins, but also contains a number of candidate virulence-related genes, so it may have a strong impact on adaptability and pathogenicity.

Members of the genus Streptococcus have a highly dynamic, open pan-genome, that potentially confers them with the ability to adapt to changing environmental conditions, i.e. antibiotic resistance or transmission between different hosts. Hence, understanding of genome evolution is important for the identification of potential pathogens and design of drugs and vaccines.

Genome characterization of a new species of microsporidia parasitizing in a cilate

Yulia Yakovleva (SPbU)
Elena Nassonova (SPbU, INC RAS)
Andrey Vyshnyakov (SPbU)
Natalia Lebedeva (Research park SPbU)
Elena Sabaneyeva (SPbU)

Microsporidia is a group of protists related to fungi. All of them are obligate intracellular parasites. They are found in nearly all groups of animals. They are pathogens of economically important species and cause opportunistic infections in humans. Microsporidia have a peculiar cell structure and a complicated life cycle. They lack mitochondria and flagella, and they are characterized by highly reduced cell machinery. The genomes of Microsporidia are very small and vary in size from 2.3 to 24 Mb (Corradi et al., 2009). Small size of genomes and low levels of splicing make Microsporidia a good model for bioinformatic analysis. Recent analysis of microsporidial genomes has shown that their genome evolution was highly dynamic and comprised reductive evolution, balanced constraint and genome expansion during adaptation to obligate intracellular lifestyle (Nakjang et al., 2013). Bioinformatic analysis of Microsporidia, besides general interest in reductive genome evolution, is primarily aimed at revealing genes involved in host-parasite interactions, providing for successful infection, and participating in regulation of the parasite life cycle. Bioinformatic analysis of microsporidial genomes and transcriptomes revealed large numbers of hypothetical proteins of unknown function.

Though Microsporidia have a wide host range, they are rare in protists, and, especially in ciliates - less than ten species have been registered so far, morphological description provided only for four of them. Since a ciliate cell is an individual organism, such system would have an advantage for investigation of host-parasite interactions compared to tissue cell cultures infected with microsporidia. However, investigation of microsporidial genomes and transcriptomes presents some difficulties caused by their obligate intracellular lifestyle.

Recently we have described a new genus of Microsporidia, parasitizing in Paramecium primaurelia. The project proposes whole genome sequencing and its structural and functional annotation. For this purpose, two approaches in sample preparation will be used. Firstly, the parasite life stages will be isolated from the host cell manually by means of micromanipulator. Secondly, of parasite life stages will be separated and purified in Percoll gradient. NGS sequencing will be performed using HiSeq 2500 Illumina system, using paired-end (PE) sequencing reads. PE libraries will be prepared according to standard protocols using Nextera DNA Sample Preparation Kit. After filtering and error-correction genome assembly will be performed both, de novo and using two reference genomes of microsporidia from the same clade, Encephalitozoon cuniculi and Nosema bombycis.

Supported by the RFFI grant 18-04-0562a.

B-vitamin dietary requirements and sharing capabilities of the human gut microbiome

Aleksandr Arzamasov (A.A. Kharkevich Institute for Information Transmission Problems)
Matvei Khoroshkin (A.A. Kharkevich Institute for Information Transmission Problems)
Stanislav Yablokov (A.A. Kharkevich Institute for Information Transmission Problems)
Andrei Osterman (Sanford Burnham Prebys Medical Discovery Institute)
Dmitry Rodionov (A.A. Kharkevich Institute for Information Transmission Problems; Sanford Burnham Prebys Medical Discovery Institute)

The human gut microbiota (HGM) is a complex microbial community that inhabits the gastrointestinal tract. Despite rapidly growing knowledge, there is still little known about the cross-feeding of specific nutrients between the community members. Among these micronutrients, vitamins of the B group are of particular interest, since they serve as precursors for many essential coenzymes. Many members of the HGM are not capable to produce some or all B vitamins (auxotrophs), while others possess complete biosynthetic pathways for these nutrients (prototrophs). Therefore, potential vitamin exchange between auxotrophs and prototrophs might be one of the key factors shaping the community structure of the HGM.

In the present study, we used the subsystem approach implemented in the SEED genomic platform to study a potential metabolite exchange in a set of 2228 strains with sequenced genomes representing 690 cultured HGM species. For all genomes, we reconstructed metabolic pathways for biosynthesis of eight B vitamins and their corresponding cofactors (thiamine/TPP, riboflavin/FMN/FAD, niacin/NAD, folate/THF, pantothenate/CoA, pyridoxine/PLP, biotin, cobalamin/coenzyme B12) as well as for queuosine, which is a modified nucleoside that is present in certain tRNAs in bacteria and eukaryotes. We also analyzed a distribution of known vitamin and vitamin precursors transporters in these bacteria. The inferred data allowed us to classify the studied organisms with respect to their biosynthetic and transport capabilities. Incomplete biosynthesis pathways for some vitamins, such as biotin, cobalamin, thiamine and pantothenate, suggest certain vitamin deficiencies can be alternatively supplemented by their metabolic precursors (e.g. dethiobiotin, cobinamide, thiazole, pantoate). Overall, auxotrophic phenotypes are much more abundant in HGM and only a small subset of microorganisms can synthesize all vitamins. We propose several candidate transporters that could be involved in a vitamin sharing by prototrophs. In summary, the obtained in silico reconstructions contribute to our understanding of metabolic cross-feeding processes in HGM.

Comparative genomic analysis of Thermofilum genus: insights in to polysaccharide degradation and carbohydrate metabolism.

Kseniya Zayulina (Research center of Biotechnology RAS)
Ulyana Piunova (MSU)
Tatiana Kochetkova (Research center of Biotechnology RAS)
Ilya Kublanov (Research center of Biotechnology RAS)

Archaea is a one of three currently accepted domains of life and consists of several deep phylum-level lineages among which Crenarchaeota and Euryarchaeota are mostly studied so far. Characterized Archaea representatives are known by unique habitats, peculiar metabolism and adaptation mechanisms to various environments. Most of known cultivated archaea are extremophiles: organisms living in extreme environmental conditions, such as high temperature, low or high pH, high salt concentration etc. Due to deep phylogenetic position and adaptations to harsh environmental conditions archaeal metabolic pathways often implicate unusual enzymes and unique or modified reactions.

This study is focused on polysaccharide decomposition and central carbohydrate metabolism of Thermofilaceae family. This crenarchaeal family so far represented by a few number of cultivated microorganisms and several tens of environmental clones. To reconstruct the mechanisms of carbohydrate utilization by Thermofilaceae we analyzed genomes of all cultivated species T. pendens strain Hrk5, T. uzonense strain 1807-2, “T. adornatus” strain 1910b and “T. adornatus” strain 1505-1.

Since the first references regard Thermofilum representatives as opportunistic heterotrophs (Zillig et al., 1983), their metabolic potential was underestimated, however, later works have shown their relatively high capabilities in at least catabolic reactions, eg. degradation of sugars (Toshchakov et al., 2015). All cultivated species excluding “T. adornatus” 1505-1 are able to grow on starch and glucose as sole carbon source. Moreover, T. uzonense strain 1807-2 and “T. adornatus” strain 1910b are capable of glucomannan and cellulose degradation, respectively. It was recently shown that T. pendens Hrk5 and T. uzonense 1807-2 genomes contain genes, encoding proteins presumably involved in cellulose degradation (Anderson et al., 2008; Toshchakov et al., 2015). Here, we continued this analysis in more depth and for all available Thermofilum genomes. The pathways of starch, mannan and glucomannan degradation were reconstructed for strains 1910b and 1807-2. Determinants of mannose metabolism were presented in all Thermofilums genomes. Glucose utilization occurs via the archaeal type Embden–Meyerhof–Parnas (Ahmed et al., 2005), for which all necessary genes were found in all studied genomes. Sucrose synthase, which is responsible for sucrose utilization in Cyanobacteria and Viridiplantae genes were identified in 1505-1 and Hrk5 genomes, as well as enzymes involved in further metabolic reactions. Both raffinose and lactose utilization pathways were detected by the analysis of genomes of Hrk5, 1807-2 and 1910b strains, while 1505-1 genome contained only a beta-galactosidase gene. No homologs of currently known cellulase genes were found in strain 1910b genome despite its capability to grow on this substrate. Hence, proteomic approach was applied to find putative candidates, involved in this process. We conducted a proteomic analysis using strain 1910b cells, grown on cellulose or on pyruvate as the control. Proteomic analysis shown significant expression of a number of hypothetical proteins in “cellulose” experiment. Many of them contained specific domains involving in polysaccharide degradation. These results will be further verified experimentally (by heterologous expression in E. coli of respective candidate genes) allowing deeper understanding.

References:

1. Zillig et al., 1983. The archaebacterium Thermofilum pendens represents, anovel genus of the thermophilic, anaerobic sulfur respiring Thermoproteales

2. Anderson et al., 2008. Genome sequence of Thermofilum pendens reveals an exceptional loss of biosynthetic pathways without genome reduction

3. Toschakov S.V et al., 2015. Complete genome sequence of and proposal of Thermofilum uzonense sp. nov. a novel hyperthermophilic crenarchaeon and emended description of the genus Thermofilum

4. Ahmed et al., 2005. The semi-phosphorylative Entner–Doudoroff pathway in hyperthermophilic archaea: a re-evaluation.

Pathracer: racing profile HMM paths on assembly graph

Alexander Shlemov (Saint Petersburg State University)
Anton Korobeynikov (Saint Petersburg State University)

One typical way to represent the variations within the protein or gene families is via profile Hidden Markov Models (pHMMs). Recently large databases emerged that contains pHMMs representing 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 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 the Pathracer tool that aligns HMM directly to the assembly graph (performing the codon translation on fly if necessary 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 interested 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.

SGTK: a toolkit for visualization and assessment of scaffold graphs

Olga Kunyavskaya (Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia)
Andrey D. Prjibelski (Center for Algorithmic Biotechnology, Institute of Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia)

Scaffolding is an important step in every genome assembly pipeline, which allows to order contigs into longer sequences using various types of linkage information, such as
libraries and long reads. In this work we operate with a notion of a scaffold graph — a graph, vertices of which correspond to the assembled contigs, and edges represent connections between them.
We present a software package called Scaffold Graph ToolKit (SGTK) that allows to construct and visualize scaffold graphs using different kinds of sequencing data. We show that the scaffold graph appears to be useful for analyzing and assessing genome assemblies, and demonstrate several use cases that can be helpful for both assembly
software developers and their users.

Near-optimal de novo genome assembly of Saccharomyces cerevisiae strain from the Peterhof genetic collection

Yury A. Barbitoff (Dpt. Of Genetics and Biotechnology, Saint-Petersburg State University, St. Petersburg, Russia; Bioinformatics Institute, St. Petersburg, Russia)
Alexander V. Predeus (Bioinformatics Institute, St. Petersburg, Russia; University of Liverpool, Liverpool, United Kingdom)
Alexandra V. Beliavskaia (Dpt. Of Invertebrate Zoology, Saint-Petersburg State University, St. Petersburg, Russia; University of Liverpool, Liverpool, United Kingdom)
Svetlana E. Moskalenko (Dpt. Of Genetics and Biotechnology, Saint-Petersburg State University, St. Petersburg, Russia; St. Petersburg Branch, Vavilov Institute of General Genetics, Russian Academy of Sciences, St. Petersburg, Russia)
Galina A. Zhouravleva (Dpt. Of Genetics and Biotechnology, Saint-Petersburg State University, St. Petersburg, Russia; Laboratory of Amyloid Biology, St. Petersburg State University, St. Petersburg, Russia)

Third generation sequencing technologies based on long single-molecule reads allow for rapid and complete assembly of prokaryotic and small eukaryotic genomes. In our study, we used Oxford Nanopore MinION sequencing to create a de novo genome assembly of 1A-D1628 strain of Saccharomyces cerevisiae from the Peterhof genetic collection. We obtained 10.015 Gb of sequencing reads (~800x coverage of the genome) from one MinION run using 1D library preparation kit. De novo genome assembly using Canu genome assembler produced near-chromosome level assembly, yielding 23 contigs, with 18 expected from strain's genotype. Genome alignment analysis of the initial assembly indicated overwhelming similarity to S288C S. cerevisiae reference, with more than 95% of the reference present in the assembly. BUSCO analysis of the raw assembly using fungi_odb9 database identified 126 complete, 7 duplicated, 123 fragmented, and 41 missing core gene groups. We further improved our assembly with 2x150 paired short reads obtained using Illumina HiSeq 4000 platform. The assembled genome will be used in further comparative genomics studies. Our results demonstrate the utility of third generation sequencing technologies to produce near optimal assemblies of small eukaryotic genomes from single sequencing run.

The study was supported by RSF grant no. 18-14-00050. Equipment of the Biobank Resource Centre of SPBU was used in the present work.

Supplier dependent metabolism of Thermofilaceae family representatives

Ulyana Piunova (MSU FBB)
Kseniya Zayulina (Research center of Biotechnology RAS)
Tatiana Kochetkova (Research center of Biotechnology RAS)
Ilya Kublanov (Research center of Biotechnology RAS)

The family Thermofilaceae is a deeply branching phylogenetic lineage, including so far one validly published genus and two species – Thermofilum pendens and Thermofilum usonenze. Both possess peculiar metabolic properties such as growth dependence on external metabolites from co-habiting microorganisms. T. pendens was isolated from a solfataric hot spring in Iceland, and was characterized as an anaerobic, hyperthermophilic, moderately acidophilic chemoorganoheretotrophic archaeon, utilizing peptides as the energy source and sulfur as the electron acceptor. Notably, its growth is obligately dependent on Thermoproteus tenax Kra1 polar lipids fraction. Another validly published representative, T. usonenze 1807-2 was isolated from a hot spring in Uzon Caldera, Kamchatka and described as an anaerobic hyperthermophilic, slightly acidophilic chemoorganoheterotroph, fermenting peptone, yeast extract, as well as glucomannan and starch. In contrast to T. pendens, T. usonenze was dependent on culture broth filtrate of another Crenarchaeon (Fervidococcus fontis, Desulfurococcus kamchatkiensis or Pyrobaculum arsenaticum), which served as a source of unknown growth factors. Two recently isolated strains of "Thermofilum adornatus" – 1505 and 1910b also required the filtrate of crenarchaeal culture broth – F. fontis and D. kamchatkiensis, respectively. Currently, genomes of all above-mentioned Thermofilum strains are sequenced, assembled and deposited in GenBank Database as well as the genomes of Crenarchaeota, used as suppliers of the growth factors.
Present study is devoted to revealing crenarchaeal metabolites served as co-factors for Thermofilum representatives, using comparative genomics and microbiological approaches. The analysis of genome sequences using the Kyoto Encyclopedia of Genes and Genomes (KEGG) and MetaCyc databases was performed using IMG/MER and BioCyc web services.
Comparative genomic analysis of Thermofilum species and other Crenarchaeota, used as donors of growth factors, revealed significant differences in several biosynthetic metabolic pathways. In all Thermofilum genomes reduction in biosynthetic pathways for purines, several amino acids, and co-factors was observed, suggesting these compounds cannot be synthesized by Thermofilum representatives. As it was shown previously, growth of T. pendens dependent on lipid fraction of T. tenax (Zillig et al., 1983), while three other Thermofilum strains were dependent on unknown intracellular metabolites. To unravel these metabolites the genome of T. pendens was compared with other Thermofilum genomes in order to find pathways, present in T. pendens and absent in other three genomes. Comparative analysis revealed several crucial anabolic pathways like L-cysteine and L-serine biosynthesis, 4-amino-2-methyl-5-diphosphomethylpyrimidine biosynthesis, coenzyme A biosynthesis, pyridoxal 5'-phosphate, pyrimidines biosynthesis were reduced or lacked in three other Thermofilum genomes, while completely present in T. pendens.   These results suggest a broad range of metabolites or co-factors, needed to be exported for growth of the three Thermofilum strains. However, this should be verified experimentally since some of these pathways

could be modified and based on another proteins, encoded in the genomes of respective microorganisms. This experimental work will be performed in near future. Zillig et al., 1983. The archaebacterium Thermofilum pendens represents, a novel genus of the thermophilic, anaerobic sulfur respiring Thermoproteales.

Identification and annotation of repetitive DNA in the genome of sterlet (Acipenser ruthenus)

Prokopov D.Y. (Institute of Molecular and Cellular Biology Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia)
Makunin A.I. (Institute of Molecular and Cellular Bionstitute of Molecular and Cellular Biology Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russialogy Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia)
Biltueva L.S. (Institute of Molecular and Cellular Biology Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia)
Komissarov A.S. (Theodosius Dobzhansky Center for Genome Bioinformatics, Saint-Petersburg State University, Saint-Petersburg 199004, Russia)
Sarachakov A.E. (Institute of Molecular and Cellular Biology Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia)
Graphodatsky A.S. (Institute of Molecular and Cellular Biology Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia)
Trifonov V.A. (Institute of Molecular and Cellular Biology Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia)

Sturgeons represent an interesting group due to their paleopolyploid states, basal position within the Actinopterygii, slow rates of genomic evolution and cryptic sex chromosomes. The genome of sterlet can be used as a reference genome of sturgeons because of the lowest ploidy level. The composition of the repeated DNA elements of sturgeons remains unexplored but such studies can shed light on the cryptic sex chromosomes, the structure of the subgenomes of the paleopolyploids, and expand our knowledge of horizontal transfer among distant taxa of vertebrates.

In this project we sequenced the sterlet male and female genomic DNA (on the Illumina HiSeq platform) and the sterlet female DNA (on the PacBio platform) and then assembled the sterlet genome. In addition, a transcriptome of testis and ovaries was sequenced and assembled. Identification of tandem repeats was carried out using the k-mer analysis directly on raw reads. Furthermore, de novo identification of repeated elements was carried out on raw short Illumina and long PacBio reads, and genomic assemblies using specific tools. As a result, we obtained a comprehensive library of repeated elements.

Then, we identified tandem repeats, including those differing in content between male and female, but the FISH-analysis showed that these differences resulted from an intraspecific polymorphism. However, we were able to identify chromosome-specific satellites, including those that allow us to distinguish between paralogous chromosomes, resulted from ancestral polyploidization.

The analysis of mobile DNA demonstrated the predominance of DNA transposons over retroelements, which is similar to teleost genomes. In addition, we further confirmed many cases of horizontal transfer between cyclostomes, bony fishes, chondrostei, amphibians, and reptiles. New potential cases of horizontal gene transfer among DNA transposons were found. Among them, we identified an interesting case of the insertion of a transposon from the Mariner family into the transcribed part of the ribosomal RNA gene cluster, which could facilitate its wide distribution among sturgeons. The analysis of transcripts showed the activity of the investigated DNA transposons.

Our results highlight the importance of studying the repeated elements of non-model organisms and may help to identify new ways of genome evolution.

The work was supported by the Russian Science Foundation (RSF) grant number 18-44-04007

Genome evolution in Felidae family

Ksenia Krasheninnikova (Dobzhansky Center for Genome Bioinformatics, St.Petersburg State University)
Anna Zhuk (Dobzhansky Center for Genome Bioinformatics, St.Petersburg State University)
Sergey Kliver (Dobzhansky Center for Genome Bioinformatics, St.Petersburg State University)
Klaus-Peter Koepfli (Dobzhansky Center for Genome Bioinformatics, St.Petersburg State University)
Stephen J. O'Brien (Dobzhansky Center for Genome Bioinformatics, St.Petersburg State University)

Felidae family is characterised with a relatively stable karyotype while it's also known for diverse morphological adaptations. In this perspective we performed analysis of genes under positive selection in 11 Felidae species. We identified 57 unique genes involved in morphological, behavioural, metabolic, and fertility processes. Based on homologous mapping of gene markers we identified evolutionary breakpoints in different lineages of Felidae. 15 breakpoints in leopard cat lineage, 5 - in puma lineage, 3 - in Panthera lineage. The present analysis provides an insight into interspecific connections and diversity in a broad range of Felidae species and computational assessment of inter-species homology.

Tracing the ancient human history of Russia through genomics

Daria V. Zhernakova (Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation)
Valentin Shimansky (Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation)
Ivan Dmitrievsky (Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation)
Stephen J. O'Brien (Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russian Federation)

The Russian Federation spans 11 time zones and is the home of ~146,000,000 people: 80% are the ethnic Russians and the remainder identify themselves as one of ~200 indigenous ethnic minorities. The territory of modern Russia has a deep history: starting from out-of-Africa migration and interaction with Neanderthals and Denisovans, humans moved across this land, shaping the populations that we see now. Technological advances make it possible to trace back human history by analyzing gene flow and admixture of modern and ancient samples, adding new information to traditional archeological studies and providing insight into sample relatedness and migration routes.

In this study we analyzed whole genome sequencing samples of modern populations generated within the Genome Russia project and several recently published studies (Mallick et al. 2016, Pagani et al. 2016) together with previously published ancient DNA samples in order to get more insight into population history and migration events which shaped the modern populations from the territory of Russia. Our results show that Pskov and Novgorod samples show a pattern of ancient admixture similar to other populations from Eastern Europe, while Siberian populations show highly divergent patterns. Finno-Permic samples show a European-like pattern; Ugric samples contain strong Ancient North Eurasian and Siberian ancestry. Altaic groups are similar to Ugric samples with additional gene flow from Iron Age nomads detected.

Overall, this study allows to estimate ancient ancestry components for modern populations from the territory of Russia, providing insight into their history.

Versatile genome assembly evaluation with QUAST-LG

Alla Mikheenko (St. Petersburg State University, St. Petersburg, Russia)
Andrey Prjibelski (St. Petersburg State University, St. Petersburg, Russia)
Vladislav Saveliev (St. Petersburg State University, St. Petersburg, Russia)
Dmitry Antipov (St. Petersburg State University, St. Petersburg, Russia)
Alexey Gurevich (St. Petersburg State University, St. Petersburg, Russia)

Motivation: The emergence of high-throughput sequencing technologies revolutionized genomics in the early 2000s. The next revolution came with the era of long-read sequencing. These technological advances along with novel computational approaches became the next step towards the automatic pipelines capable to assemble nearly complete mammalian-size genomes.

Results: In this work, we demonstrate the performance of the state-of-the-art genome assembly software on six eukaryotic datasets sequenced using different technologies. To evaluate the results, we developed QUAST-LG --- a tool that compares large genomic de novo assemblies against reference sequences and computes relevant quality metrics. Since genomes generally cannot be reconstructed completely due to complex repeat patterns and low coverage regions, we introduce a concept of upper bound assembly for a given genome and set of reads, and compute theoretical limits on assembly correctness and completeness. Using QUAST-LG, we show how close the assemblies are to the theoretical optimum, and how far this optimum is from the finished reference.

Availability and implementation: http://cab.spbu.ru/software/quast-lg
Contact: aleksey.gurevich@spbu.ru

Hybrid transcriptome assembly and improving artichoke genome annotation

Bushmanova Elena (Center for Algorithmic Biotechnology, St. Petersburg State University)
Andrey D. Prjibelski (Center for Algorithmic Biotechnology, St. Petersburg State University)
Giuseppe Puglia (Institute for Agricultural and Forest Systems in the Mediterranean)
Domenico Vitale (Institute for Agricultural and Forest Systems in the Mediterranean)

Possibility to generate long RNA-seq reads such as Oxford Nanopore Reads can greatly improve transcriptome assembly and consequently complement the existing annotations of unfinished and poorly annotated genomes. De novo transcriptome reconstruction from short reads remains an open challenging problem, which is complicated by the varying expression levels across different genes, alternative splicing and paralogous genes. Indeed, long reads may be useful in resolving the later two problems and allow to generate more full-length isoforms. In this work we present computational methods for incorporating hybrid assembly into rnaSPAdes and provide a real life usage of this new pipeline by improving Artichoke gene annotation. We provide quality assessment reports for rnaSPAdes assemblies with and without nanopore reads, and also show impact to gene annotation caused by long reads.

cycloScore: Scoring and evaluating statistical significance of non-linear peptide-spectrum matches

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

Development of new high-throughput algorithms for identification of natural products from their mass spectra draws the attention to scoring procedures used in the quality assessment of peptide-spectrum matches. CycloScore is a new approach to such a scoring based on a probabilistic model originally developed in the scope of traditional proteomics. We benchmark cycloScore within the Dereplicator pipeline using both linear and non-linear high resolution MS/MS datasets. We show that for all the datasets, cycloScore significantly increases the number of identified compounds.

Analysis of structural variations in Human IGH Locus

Andrey Bzikadze

The formation of an antibody is a complex process involving a number of stochastic processes. The number of pathogens that humans face during their lifetime is enormous and so is the diversity of antibodies in each individual. This variability is achieved by several probabilistic processes including VDJ recombination and somatic hypermutagenesis (SHM). The immunoglobulin heavy (IGH) chain locus contains three families of germline antibody segments named V, D, and J segments. VDJ recombination generates a new antibody by random selection of a single segment from each family followed by their concatenation and somatic hypermutations that introduces mutations into this concatenate. Since each antibody is unique and since human immune system is constantly evolving, the information content of the immune system is orders of magnitude larger than the the information content of the entire human genome.

Population-wide variability analysis of the immune locus in humans is critically important for understanding the formation of antibodies. The IGH locus represents one of the most complex, repeat-rich and variable regions of the human genome that has evaded many previous attempts to assemble it across human population. Although the reference sequence of this locus was assembled twenty years ago, the loci in individual humans widely differ from this reference genome. Even though recent studies improved the assembly of this locus, its population-wide analysis and the analysis of its structural variations between individuals remains an open problem. We will present recent results in analysis of IGH locus variability showing that structural rearrangements are one of the main driving forces for achieving IGH locus diversity.

GAligner. A tool for efficient and accurate alignment of long reads onto assembly graph

Tatiana Dvorkina (Center for Algorithmic Biotechnology, Saint Petersburg State University)
Dmitry Antipov (Center for Algorithmic Biotechnology, Saint Petersburg State University)
Anton Korobeynikov (Center for Algorithmic Biotechnology, Saint Petersburg State University)
Sergey Nurk (Center for Algorithmic Biotechnology, Saint Petersburg State University)
Pavel A. Pevzner (Center for Algorithmic Biotechnology, Saint Petersburg State University)

Draft genome assemblies are commonly represented by assembly graphs. Tools for accurately aligning nucleotide and amino-acid sequences onto such graphs can facilitate hybrid assembly, read error correction, analysis of hypervariable genes, reconstruction of complex multi-domain genes and haplotype separation. At the same time, currently available solutions, such as vg, GraphAligner, TAG, have significant limitations. Here, we present GAligner (GA) — a general purpose tool for local alignment of DNA sequences onto assembly graphs. In particular, GAligner is able to accurately and efficiently map long erroneous sequences (e.g. single-molecule sequencing reads). We comprehensively benchmark GA using different datasets and sequencing technologies and show that it produces accurate alignments on assembly graphs of varying complexity.

This work has been supported by the Russian Science Foundation (grant 14-50-00069)

Reconstructing Individual Bacterial Genomes from Multiple Metagenomic Samples

Sergey Nurk (Center for Algorithmic Biotechnology, Saint-Petersburg State University, Russia)
Yury Gorshkov (Computer Technologies Laboratory, ITMO University, Russia)
Margarita Akseshina (Saint-Petersburg Academic University, Russia)
Pavel A. Pevzner (University of California San Diego, USA)

Availability of sequencing data for multiple related environmental samples provides an unprecedented opportunity for reconstruction of individual microbial community members. Incorporation of the differential abundance signal allowed to significantly improve the accuracy of unsupervised approaches for binning assembled contigs.
Yet, state-of-the-art strategies for reconstructing MAGs (metagenome assembled genomes) demonstrate some major deficiencies and limitations forcing leading groups to complement them with laborious and cryptic curation protocols.

Here we present our novel MTS (Metagenomic Time Series) pipeline for improved reconstruction of individual genomes from metagenomic series. MTS complements modern differential binning tools with automated analysis of assembly graphs and SNV-based analysis of population heterogeneity to deliver high quality MAGs.

This work has been supported by the Russian Science Foundation grant 14-50-00069.

cloudSPAdes: Metagenome Assembly from Synthetic Long Reads using de Bruijn graphs

Ivan Tolstoganov (Center for Algorithmic Biotechnology, Saint Petersburg State University)
Anton Bankevich (Center for Algorithmic Biotechnology, Saint Petersburg State University)

The emerging Synthetic Long Read (SLR) technologies pioneered by Illumina (TruSeq Synthetic Long Read technology) and 10x Genomics (Chromium technology) found many applications in genome assembly and analysis. However SLR applications in metagenomics remain limited. We demonstrate that properties of 10x Genomics SLRs generated from metagenomics data differ from that of mammalian data, which calls for new metagenomic assembly strategies. We present new results of cloudSPAdes algorithm for SLR metagenome assembly based on our analysis of the de Bruijn graph. Our benchmarking demonstrated that cloudSPAdes improves on the state-of-the-art metagenomics SLR assemblers on various metagenomic datasets.

This work was supported by the Russian Science Foundation (grant #14-50-00069)

AntEvolo: a novel approach for joint recombination and clonal analysis of antibody repertoires

Andrey Slabodkin (Center for Algorithmic Biotechnology, St. Petersburg State University)
Maria Chernigovskaya (Center for Algorithmic Biotechnology, St. Petersburg State University)
Andrey Bzikadze (Center for Algorithmic Biotechnology, St. Petersburg State University)
Yana Safonova (Center for Information Theory and Applications, University of California, San Diego)

The diversity of an antibody repertoire is achieved by various complex processes: V(D)J recombination, chain pairing, and somatic hypermutagenesis. As a result of multiple cycles of such diversification, the antibody repertoire represents a set of antibody “families” with various abundances. Each such family can be viewed as a clonal tree that represents the evolutionary development of the family. Understanding of evolutionary processes in an antibody repertoire using clonal trees gives an insight into the antibody generation and helps to understand the nature of immune response.

In this work, we propose AntEvolo, an algorithm for joint reconstruction of clonal lineages and VDJ labeling. In contrast to previous approaches, AntEvolo constructs preliminary clonal families and iteratively improves VDJ labeling, SHM assignment and clonal lineages at the same time. Proposed approach significantly improves accuracy of the computed clonal lineages. Our analysis revealed that an ideal representation of the evolutionary development of antibody should allow cycles and alternative direct paths, so the canonical representation of a repertoire as a tree can be insufficient and one should consider clonal directed networks instead of that.

The work was supported by Russian Science Foundation (grant 14-50-00069)

Revealing molecular mechanisms of specificity of the insecticidal Bacillus thuringiensis strains using whole-genome sequencing and isobaric proteomics

Yury V. Malovichko (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Valentina P. Ermolova (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Svetlana D. Grishechkina (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Maria E. Belousova (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Anton A. Nizhnikov (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)
Kirill S. Antonets (All-Russia Research Institute for Agricultural Microbiology, Pushkin, St. Petersburg, Russia)

Gram-positive, toxin-producing bacterium Bacillus thuringiensis (Bt) belonging to the phylum Firmicutes represents a widely used source of industrial insecticides, antibiotics, and enzymes. To date, biological insecticides based on Bt including transgenic crops producing Bt toxins occupy more than 60 % of the global biopecticide market. The key of commercial success of Bt-based insecticides is in the extreme diversity of toxins coupled with their high specificity and safety for humans. Various Bt strains exhibit specific insecticidal action against different orders and even families of Insecta, Gastropoda, Nematoda, and, in several cases, malignant mammalian cells. At least five major groups of exo- and endotoxins of protein and non-protein nature encoded by more than 500 open reading frames were found to control pathogenic properties of Bt strains. In this work, we plan to perform the comparative genomic and proteomic analysis of three different industrial Bt strains from the collection of All-Russia Research Institute for Agricultural Microbiology (ARRIAM) as well as their avirulent derivatives. These strains exhibit differential insecticidal properties against Coleoptera, Lepidoptera and Diptera, and some of them also demonstrate polyfunctional (fungicidal and plant growth stimulating) activities. To perform a comparative whole-genome analysis of these strains, two different sequencing platforms (Illumina and Oxford Nanopore) are used. For proteomic analysis, the bottom-up approach including high-performance liquid chromatography coupled with time-of-flight mass-spectrometry (HPLC-MS) is applied. In addition, we assume to employ isobaric labelling of the tryptic peptides that provides quantification of the levels of production of the corresponding proteins. Overall, this comparative genomic and proteomic study of Bt strains will highlight the molecular mechanisms underlying specificity and pleiotropic activities of this bacterium against different insects and fungi and contribute to the knowledge of mechanisms of the virulence of prokaryotes.

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

TUESDAY – JULY 17, DAY 2 SCHEDULE

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

9:00AM–10:00AM	I	**Towards perfect de novo DNA assembly** Gene Myers *Max-Planck Institute for Molecular Cell Biology and Genetics*
10:00AM–10:20AM	T	**Bwise: a novel accurate, haplotype-specific genome assembler** Antoine Limasset *Université libre de Bruxelles*
10:20AM–10:40AM	B	**Break**
10:40AM–11:00AM	T	**Assembling barcoded RNA sequencing data** Andrey PrjibelskI *Center for Algorithmic Biotechnology, SPbU*
11:00AM–11:20AM	T	**BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters From Assembly Graphs** Dmitry Meleshko *Center for Algorithmic Biotechnology, SPbU*
11:20AM–11:40AM	T	**Plasmid detection and assembly in genomic and metagenomic datasets** Dmitry Antipov *Center for Algorithmic Biotechnology, SPbU*
11:40AM–12:00PM	T	**CellPi: unsupervised processing pipeline of mouse and human single-cell RNA-seq data** Alexey Samosyuk *Skoltech*
12:00PM–1:20PM	B	**Lunch**
1:20PM–2:20PM	I	**Test tubes, sequencing machines, computers: bioinformatics as a molecular biology tool** Mikhail S. Gelfand *Institute for Information Transmission Problems*
2:20PM–3:00PM	F	**Semantic-based antibody folding and structural annotation** Pavel Yakovlev *Biocad*
3:00PM–4:00PM	I	**CRISPR: fascinating biology and limitless applications** Eugene V. Koonin *National Institutes of Health*

4:00PM–4:20PM	B	**Break**
4:20PM–4:40PM	T	**ClinCNV: novel method for large-scale CNV and CNA discovery** German Demidov *Institute of Medical Genetics and Applied Genomics, Tübingen, Germany*
4:40PM–5:00PM	T	**Genome rearrangements in bacteria** Olga O Bochkareva • Pavel V Shelyakin *IITP RAS, Skoltech • VIGG RAS, Skoltech*
5:00PM–5:20PM	T	**Analysis and visualization of segmental duplications in mammalian genomes** Alla Mikheenko *Center for Algorithmic Biotechnology, SPbU*
6:00PM–9:00PM	B	**Dinner**

The conference was supported by the Russian Fund for Basic Research (grant 18-04-20035)

TUESDAY – JULY 17

DAY 2 TALK SUMMARIES

Towards perfect de novo DNA assembly

Gene Myers (Max-Planck Institute for Molecular Cell Biology and Genetics)

We are about to enter an era of DNA sequencing where one can in the near future produce, *de novo*, a reference-quality genome of any living species for 1,000 EU. This ability will revolutionize ecology, evolution, and conservation science and effectively mark the beginning of a new exploration of the natural world.

The technological driver is the advent of long read sequencers such as the PacBio Sequel and Oxford Promethion. The long reads in effect make assembly easier, and one sees corresponding improvements in the continuity of the results, but the underlying algorithms are effectively the same as those first developed 20 years ago, and repetitions at the scale of read length are still an issue. Indeed, truly better assembly requires finding all artifacts in the reads and the resolution of repeat families, topics that I don't think have received sufficient attention and that are particularly critical issues for long reads.

Therefore we are developing algorithms that carefully analyze a long read shotgun data set before assembly. By efficiently comparing all the data against itself we have developed a computational approach to accurately determine the quality of any stretch of a PacBio read based only on the sequence data itself. These *intrinsic* QVs allow us to accurately identify low quality regions, chimers, and missed adaptamers. Removing these artifacts with a process we call *scrubbing* leaves one with reads that assemble without the need for base-level error correction. We have further developed a heuristic consensus algorithm that is far more efficient and accurate than previous methods and further identifies potential sites of variation due to haplotypes or repeats. Using this algorithm we further correct reads, typically to Q40 (99.99% accurate). In effect, we have developed a process that takes Q7 reads full of artifacts, and produces Q40 artifact-free reads solving all aspects of the assembly problem save the separation of nearly identical, ubiquitous, and large repeats.

Bwise: a novel accurate, haplotype-specific genome assembler

Antoine Limasset (Université libre de Bruxelles, Belgium)
Camille Marchet (Université de Rennes, INRIA, CNRS, IRISA, France)
Pierre Peterlongo (Université de Rennes, INRIA, CNRS, IRISA, France)
Jean-François Flot (Université libre de Bruxelles, Belgium)

Assemblers based on the de Bruijn graph (DBG) paradigm usually discard lots of useful information from short paired-end reads, resulting in fragmented assembly (particularly in the case of heterozygous genomes).
String graphs based assemblers may be able to use the whole read information but suffer from low scalability on large datasets.
To combine those two approaches, we efficiently align reads (paired or not) on DBG in a new assembler dubbed Bwise (short for “de Bruijn workflow using integrally the information of short paired-end reads”).
Bwise maps reads (or read pairs) on the (cleaned, compacted) DBG generated from the same set of reads.
A previous work, a short read corrector BCOOL[1], showed that such mapping provided very accurate sequences.
Here we use them as so called super-reads (i.e. linear paths of unitigs) that are subsequently filtered and assembled into contigs.
To improve the initial set of contigs, a new DBG can be constructed with a higher kmer size to reiterate the assembly process on a simpler and less fragmented graph.
As k increases up to read length (or even beyond it), the contig graph outputted by Bwise becomes progressively simpler and the statistics of the contig set improve dramatically. Bwise were originally designed for assembling complex diploid or polyploid genomes and showed great results in that way.
In the case of the rotifer A.vaga the MIRA assembler obtained an assembly with a 47kb N50 in 9 months on a cluster presenting more than one terabyte of RAM where
Bwise was able to propose a N50 of 150kb in 2 hours on a 20 core cluster.
But Bwise also performs very well on haploid or meta-genomic data-set, often delivering assemblies more continuous and accurate than other state-of-the-art approaches.
Bwise is also scalable and is able to assemble a 100X human data-set using less than 100GB of RAM in two days on a 20 cores cluster.

[1] Limasset, Antoine, Jean-Francois Flot, and Pierre Peterlongo. "Toward perfect reads: self-correction of short reads via mapping on de Bruijn graphs" arXiv preprint arXiv:1711.03336 (2017).

Assembling barcoded RNA sequencing data

Andrey Prjibelski (Center for Algorithmic Biotechnology, SPbU)
Elena Bushmanova (Center for Algorithmic Biotechnology, SPbU)
Dmitrii Meleshko (Weill Cornell Medicine)
Hagen Tilgner (Brain and Mind Research Institute, Weill Cornell Medicine)

De novo transcriptome assembly is a valuable alternative to the classic reference-based methods for RNA-Seq analysis. Although, multiple approaches and algorithms were developed, the problem of restoring all complete full-length isoforms remains a challenging problem that, in some cases, cannot be possibly resolved just by using short paired-end reads or coverage depth. We propose to utilize a recently developed barcoded RNA sequencing protocol that allows to generate reads with each barcode corresponding to a separate RNA molecule. To enable assembly of this protocol we developed algorithms on top of existing RNA-Seq assembler rnaSPAdes. In this manuscript we demonstrate that using barcoded data leads to dramatic improvements in de novo transcriptome assembly quality, such as generation of complete transcript sequences even for the complex eukaryotic genes with multiple expressing isoforms.

BiosyntheticSPAdes: Reconstructing Biosynthetic Gene Clusters From Assembly Graphs

Dmitrii Meleshko (Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia, Tri-Institutional PhD Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, United States)
Hosein Mohimani (Department of Computer Science and Engineering, University of California, San Diego, Computer Biology Department, School of Computer Sciences, Carnegie Mellon University)
Iman Hajirasouliha (Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Cornell Medicine of Cornell University, New York, United States, Englander Institute for Precision Medicine, Meyer Cancer Center, Weill Cornell Medicine, New York, United States)
Vittorio Tracanna (Bioinformatics Group, Wageningen University, Wageningen, The Netherlands)
Marnix H. Medema (Bioinformatics Group, Wageningen University, Wageningen, The Netherlands)
Anton Korobeynikov (Center for Algorithmic Biotechnology, St. Petersburg State University; Department of Statistical Modelling, St. Petersburg State University, St. Petersburg, Russia)
Pavel Pevzner (Center for Algorithmic Biotechnology, Institute for Translational Biomedicine, St. Petersburg State University, St. Petersburg, Russia)

Predicting Biosynthetic Gene Clusters (BGCs) is critically important for discovery of antibiotics and other natural products. While BGC prediction from complete genomes is a well-studied problem, predicting BGC in fragmented genomic assemblies remains an open problem. The existing BGC prediction tools often assume that each BGC is encoded within a single contig in the genome assembly, a condition that is violated for many sequenced microbial genomes where BGCs are often scattered through several contigs, making it difficult to reconstruct them. The situation is even more severe in shotgun metagenomics, where the contigs are often short, and the existing tools fail to predict a large fraction of long BGCs. While it is difficult to predict BGCs spanning multiple contigs, the structure of the genome assembly graph often provides clues on how to combine multiple contigs into segments encoding long BGCs. We describe biosyntheticSPAdes, a novel tool for predicting BGCs in assembly graphs and demonstrate that it greatly improves the reconstruction of BGCs from genomic and metagenomics datasets.

Plasmid detection and assembly in genomic and metagenomic datasets

Dmitry Antipov (SPbU)
Mikhail Raiko (SPbU)
Alla Lapidus (SPbU)
Pavel A. Pevzner (SPbU, UCSD)

Although plasmids are important for bacterial survival and adaptation, plasmid detection and assembly from genomic, let alone metagenomic, samples remains challenging. The recently developed plasmidSPAdes assembler addressed some of these challenges in the case of isolate genomes but stopped short of detecting plasmids in metagenomic assemblies.
We present the metaplasmidSPAdes tool that enabled plasmid assembly in metagenomics datasets and reduced the false positive rate of plasmid detection as compared to the state-of-the-art approaches. Applications of plasmidSPAdes and metaplasmidSPAdes to diverse isolate and metagenomics datasets revealed a surprisingly high yield of novel plasmids without significant similarities with known plasmids and plasmids carrying antibiotic-resistance genes.

CellPi: unsupervised processing pipeline of mouse and human single-cell RNA-seq data

Alexey Samosyuk (Skoltech)
Ilia Kurochkin (Skoltech)
Dmitri Pervouchine (Skoltech)

Single-cell RNA-seq becomes a standard for cell types characterization. Different single cell-specific tools been developed in the last four years. We present a pipeline that combines well known tools and ideas into a user-friendly R package that provides consistently good clustering results on a wide range of single cell experiments. We demonstrate that CellPi is capable to separate highly homogeneous cell populations of the developing embryo as small as 6-10 cells. At the same time CelPi comes along with an ability to detect inseparable sub-populations up to 2000 cells without using any additional annotations in a fully unsupervised way.

Test tubes, sequencing machines, computers: bioinformatics as a molecular biology tool

Mikhail S. Gelfand (Institute for Information Transmission Problems)

Combination of comparative genomics approaches and large-scale data analyses allows one to make specific predictions about the function and regulation of concrete genes that can then be validated using standard experimental techniques. Notably, these prediction go way beyond simple similarity-based annotations and often describe novel biological phenomena.

I shall present some recent examples of such studies, including discovery of the second lactose catabolism pathway and a global regulator of motility in *Escherichia coli*, and characterization of the desiccation-rehydration cycle in a midge *Polypedilum vanderplanki*.

Semantic-based antibody folding and structural annotation

Pavel Yakovlev (Biocad)

CRISPR: fascinating biology and limitless applications

Eugene V. Koonin (National Institutes of Health)

CRISPR is the new generation of genome editing and regulation tools that have rapidly revolutionized the practice of genome engineering. However, CRISPR is much more than that. It is a system of microbial adaptive immunity the existence of which has not been suspected until recently and that embodies the Lamarckian principle of evolution by inheritance of acquired characters. The evolutionary history of CRISPR itself is also remarkable, revealing surprising connections between parasitic genetic elements and host defense. I will discuss the biology and evolution of CRISPR-Cas and the molecular features that make it uniquely efficient as a genome editing tool.

ClinCNV: novel method for large-scale CNV and CNA discovery

German Demidov (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Spain; Institut für Medizinische Genetik und angewandte Genomik, Tübingen),
Stephan Ossowski (Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; Institut für Medizinische Genetik und angewandte Genomik, Tübingen)

Germline copy number variants (CNV) and somatic copy number alterations (CNA) are a common source of genomic variation involved in many genomic disorders, such as schizophrenia or cancer. Genomic microarrays, FISH, MLPA, as well as many other technologies, are widely used for detection of CNVs. Whole-genome sequencing (WGS) and whole-exome sequencing (WES) are well established, highly accurate tools for the detection of point mutations and small indels. CNV/CNA detection using WGS/WES data has been emerging as a competitive alternative for interrogating such type of variation, but remains challenging.

We have developed a new method for multi-sample CNV/CNA detection. The ClinCNV method can integrate multiple data types, including signatures derived from various WES, WGS, and microarray protocols. At first, the reference genome is divided into non-overlapping windows and different sources of information such as read depth, hybridization intensity or B-allele frequency ratio are quantified and normalized, taking into account both window- and sample-specific variability. In case if the same sample was analyzed several times with different experimental techniques (e.g., WES and shallow WGS), which is a common approach for diagnostics, evidence of copy number changes from different sources is summed up into a single matrix of likelihoods of size [number of windows] by [number of states], where states denotes distinct copy numbers. Next, ClinCNV recursively identifies segments with the strongest evidence of CNV presence in a two-step manner: common CNVs that have >5% frequency within the studied cohort are identified first, while less frequent variants are detected in the second step. Finally, we use strict filtering to remove spurious results. The algorithm’s computational complexity is linearly dependent on the number of states, allowing simultaneous detection of hundreds of distinct non-discrete copy numbers, which is especially useful for CNA detection in heterogeneous samples with complex clonal structure as frequently found in cancer.

Using the ClinCNV method we analyzed a cohort of 2834 WGS samples from the Pan Cancer Analysis of Whole Genomes (PCAWG) study, 2651 of which passed QC control. We detected 16,907, 6,156 and 568 bi-allelic deletions, duplications and mCNV events, respectively, of size greater than 3KB. FDRs for the three variant types were estimated using the IRS [1] method and available microarray intensity data and were equal to 0.0229, 0.029 and 0.049. We also investigated segments that show non-diploid coverage patterns across the majority of samples, which potentially represent reference assembly errors or highly homologous regions such as segmental duplications. We will furthermore report results for 436 chronic lymphocytic leukemia (CLL) tumor and normal pairs analyzed by WES,and 67 shallow WGS samples with coverage depth from 0.5x to 9x (average 3.8x). Comparisons between different platforms and their power to detect CN changes will be provided.

Genome rearrangements in bacteria

Olga O Bochkareva (IITP RAS, Skoltech)
Pavel V Shelyakin (VIGG RAS, Skoltech)
Mikhail S Gelfand (IITP RAS, Skoltech)

Bacterial chromosomes are complex fast-evolving systems. Genome rearrangements and horizontal gene transfer lead to the genome plasticity that is necessary for adaptation for changes in life style. Genome rearrangements play the important role in bacterial evolution as they can destroy genes, create new genes and change the copy number of gene transcripts.
Accumulation of large amount of whole-sequenced bacterial genomes from closely-related species allows to study genome rearrangements in context of evolution.
We reconstructed the evolutionary history of genome rearrangements for bacterial species from diverse ecological niches and with different genome organization. Our results show that rearrangement rates differ dramatically in different bacterial species, that is likely to be related to the adaptation driven by changes in life style. Meanwhile, for newly formed pathogens Yersinia pestis and Burkholderia mallei we revealed the correlation between mutations rates and inversions rates.
Analysis of contradictions between the obtained evolutionary trees based on the alignments of common genes and the gene order yielded numerous parallel rearrangements. Numerous gene losses and inversions likely have been caused by a high rate of intragenomic recombination between limited number of repeated elements such as transposases and 16S-23S rRNA clusters.
In Streptococcus pneumoniae and Burkholderia pseudomallei we revealed parallel inversions that may result in phase (antigenic) variation.
The reconstructed inter-chromosome translocations in bacterial genomes with multi-chromosome genome organization indicate strong selection against transfer of large fractions of genes between the leading and the lagging strands.

Supported by the Russian Science Foundation under grant 18-14-00358.

Analysis and visualization of segmental duplications in mammalian genomes

Alla Mikheenko (Center for Algorithmic Biotechnology, Saint Petersburg State University, Russia)
Lianrong Pu (Department of Computer Science and Technology, Shandong University, Jinan, China)
Yu Lin (Research School of Computer Science, Australian National University, Canberra, Australia)
Pavel Pevzner (University of California at San Diego, San Diego, USA)

Segmental duplications (SDs) play key roles in gene evolution and genomic diseases. Nevertheless, the real extent of SDs in the genomes remains unknown because SDs represent a significant impediment to accurate human genome assembly and existing assembly tools often collapse highly similar SDs. Thus, the tools capable of accurate finding and thorough analysis of SDs are extremely important. The recently developed SDquest algorithm for SD detection has shown that SDs account for at least 6% of the human genome. The novel genome assembler Flye possesses a unique possibility of reconstructing the mosaic structure of SDs using the assembly graph. At the same time, a huge number of identified SDs makes their analysis a challenging problem due to the lack of suitable visualization tools. To counter this gap, we developed a novel genome visualizer for accurate assessment and analysis of SDs. It allows to explore intra- and interchromosomal duplications and analyze their complex mosaic structure. The visualization can ease the process of SD analysis and help to reveal new SD patterns. The tool is available online.

WEDNESDAY – JULY 18, DAY 3 SCHEDULE

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

09:00AM–10:00AM	I	**Sequencing genome diversity in fish** Richard Durbin *Dept. of Genetics, University of Cambridge*
10:00AM–10:40AM	F	**The Genome Russia Project – 2018** Stephen J OBrien *Saint Petersburg State University*
10:40AM–11:00AM	B	**Break**
11:00AM–11:20AM	T	**A Rapid Exact Solution for the Guided Genome Aliquoting Problem** Maria Atamanova *ITMO University*
11:20AM–11:40AM	T	**Bounded-length Smith-Waterman alignment** Alexander Tiskin *University of Warwick*
11:40AM–12:00PM	T	**Reconstruction of a Set of Points from the Noise Multiset of Pairwise Distances in n^2 Steps for the Cyclic Sequencing Problem** Eduard Fomin *Institute of Cytology and Genetics SB RAS*
12:00PM–12:20PM	T	**Bayesian modelling of gene network alterations during tree-like processes: evolution or cells differentiation** Anna A. Igolkina *Peter the Great St.Petersburg Polytechnic University*
12:20PM–1:30PM	B	**Lunch**
1:30PM–2:30PM	I	**Discovering novel metabolisms via metagenomics** Ludmila Chistoserdova *Senior Scientist, University of Washington*
2:30PM–2:50PM	T	**Promoters and enhancers landscape of embryonic development and hibernation in chicken** Oleg Gusev *Kazan Federal University • RIKEN*

2:50PM–3:10PM	T	**Mathematical modeling of SNP %GC in microbial core genomes** Jon Bohlin *Norwegian Institute of Public Health*
3:10PM–3:30PM	B	**Break**
3:30PM–4:10PM	F	**Building time- and cost-effective bioinformatics pipelines in the Cloud - from bcl to visual analysis with New Genome Browser** *EPAM*
4:10PM–4:20PM	O	**Closing Remarks** Alla Lapidus

The conference was supported by the Russian Fund for Basic Research (grant 18-04-20035)

WEDNESDAY – JULY 18

DAY 3 TALK SUMMARIES

Sequencing genome diversity in fish

Richard Durbin (University of Cambridge)

Nearly half of vertebrate species are fish, and within them there is enormous genetic and evolutionary diversity. We have recently been involved in two large scale fish genome sequencing projects. The first focuses on the hundreds of cichlid fish species in Lake Malawi, which constitute the most extensive recent vertebrate adaptive radiation. We have mapped its genomic diversity by sequencing 134 individuals covering 73 species across all major lineages. Phylogenetic analyses suggest that no single species tree adequately represents all species relationships, with evidence for substantial gene flow at multiple times. Sequencing of related species from East African rivers indicates that the Malawi radiation arose from a hybridisation between at least two previously separated lineages, and that differentially fixed variants contributed from the ancestral lineages have been under adaptive selection within the Malawi radiation. In parallel, we have been generating high quality de novo genome reference sequences for representatives of fish orders, in the context of the international Vertebrate Genomes Project. Using long sequencing reads from single molecule technologies, and related data, we can now generate near chromosomal sequences, and there are exciting prospects for scaling these approaches towards sequencing all accessible species in the coming decade.

The Genome Russia Project – 2018

Stephen J O’Brien (Saint Petersburg State University)

The Russian Federation spans 11 time zones and is the home of ~146,000,000 people: 80% are the ethnic Russians and the remainder identify themselves as one of ~200 indigenous ethnic minorities. Despite the large population size and high ethnic diversity, no centralized reference database of functional and endemic genetic variation has been established to date. The national Genome Russia Project aims to perform high coverage whole genome sequencing and analysis of peoples of the Russian Federation. We shall describe our progress based upon resolving genome-wide variation (SNPs, indels, and copy number variation) from 264 healthy adults, including 60 newly sequenced samples consisting of family trios from three geographic regions: Pskov, Novgorod and Yakutia,. People of Russia are shown to carry known and novel genetic variants of adaptive, clinical and functional consequence that in many cases show appreciable occurrence or allele frequency divergence from the neighboring Eurasian populations. Population genetic phylogenetic analyses revealed strong geographic partitions among indigenous ethnicities corresponding to the geographic locales where they have lived. Allele frequency spectra identified strong constraints to gene flow corresponding to the geological barriers (e.g. the Ural Mountains and Verkhoyansk mountain range). These first conclusion of the Genome Russia Project include results important for medical genetics as well as for population natural history studies.

A Rapid Exact Solution for the Guided Genome Aliquoting Problem

Maria Atamanova (ITMO University)
Anton Nekhai (The George Washington University)
Pavel Avdeyev (The George Washington University)
Max A. Alekseyev (The George Washington University)

Genome rearrangements are large-scale evolutionary events that shuffle genomic architectures. Since such events are rare, the maximum parsimony assumption implies that the evolutionary distance between genomes can be estimated as the minimum number of genome rearrangements, which further enables reconstruction of ancestral genomes by minimizing the total evolutionary distance along the branches of the evolutionary tree. The basic case of this problem for three given genomes is known as the genome median problem (GMP), which asks for a single ancestral genome (median genome) at the minimum total distance from the given ones. A median genome corresponds to an optimal perfect matching in the breakpoint graph of the given genomes that maximizes the total number of 2-colored alternating cycles. While the GMP is NP-hard (Tannier and et. al, 2009), one of the prominent exact and practical solutions to the GMP is based on decomposition of the breakpoint graph into adequate subgraphs, i.e., induced subgraphs where any optimal matching can be extended to an optimal matching in the breakpoint graph (Xu and et. al, 2008).

Whole genome duplications (WGDs) represent yet another type of dramatic evolutionary events, which simultaneously duplicate each chromosome of a genome. In particular, WGDs are known to happen in the evolution of plants and yeasts. A WGD can be viewed as a partial case of a whole genome multiplication (WGM), which simultaneously creates $m \geq 2$ copies of each chromosome. An analog of the GMP in presence of a WGM is known as the guided genome aliquoting problem (GGAP). The GGAP for given genomes A and B, where all genes in B are present in a single copy (ordinary genome), while all genes in A are present in m copies, asks for an ordinary ancestral genome R that minimizes the total distance between genomes A and mR (genome resulted from the WGM of R) and between B and R.

In the present study, we propose an exact fast algorithm for solving the GGAP for $m = 2$ and $m = 3$, which is based on extension of the adequate subgraphs approach. Namely, we identify all simple adequate subgraphs of small size for the GGAP. Our algorithm searches for such subgraphs in the given breakpoint graph, finds optimal matchings in them, which are further combined and extended to an optimal matching (representing a solution R) in the breakpoint graph.

Bounded-length Smith–Waterman alignment

Alexander Tiskin (University of Warwick)

Given a fixed alignment scoring scheme, the bounded-sum Smith--Waterman alignment problem on a pair of strings of lengths m, n, asks for the maximum alignment score across all substring pairs, such that the sum of the substring lengths is above the given threshold w. This problem was introduced by Arslan and E{\u g}ecio{\u g}lu under the name ``local alignment with length threshold''. They describe a dynamic programming algorithm solving the problem in time $O(mn^2)$, and also an approximation algorithm running in time $O(rmn)$, where r is a parameter controlling the accuracy of approximation. We introduce the bounded-length Smith--Waterman alignment problem, which is closed related to the bounded-sum problem. We then show that both these problems can be solved exactly in time $O(mn)$, assuming a rational scoring scheme. Our algorithms rely on the techniques of fast window-substring alignment and implicit unit-Monge matrix searching, developed previously by the author.

Reconstruction of a Set of Points from the Noise Multiset of Pairwise Distances in n^2 Steps for the Cyclic Sequencing Problem

Eduard Fomin (Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia)

Motivation and Aim: An important fraction of the peptidoma of bacteria is non-ribosomal peptides (NRP), representing a class of secondary peptide metabolites, usually produced by bacteria and fungi, and having an extremely wide range of biological activity and pharmacological properties. In the overwhelming majority of cases (64%), NRPs have a cyclic structure [1]. In connection with their biosynthesis from the non-rybosomal path, the identification of NRPs by classical methods of bioinformatics and genomics is impossible, and is carried out only on the basis of mass spectrometry.

Mathematically, the sequencing of a cyclic chain from mass spectra is reduced to the problem known to mathematicians for long: the recovery of the coordinates of a set of points X from the multiset of pairwise distances between them ΔX (so-called the beltway problem, which having no polynomial-time algorithm in the general case). The computational complexity of the best algorithm developed by now is O(n^n log n) [2]. Despite the many approaches used (the brute force method, graph models, dynamic programming, the divide and conquer method, hidden Markov models, spectral convolution, etc.), attempts to design a polynomial algorithm for the beltway problem failed. So at present, the possibilities of de novo reconstruction of the structure of cyclic NRPs are limited. Thus, the development of new bioinformatic methods for the reconstruction of bacterial non-ribosomal peptides is very relevant.

Methods and Algorithms: We proposed a new method to solve the problem [3,4]. It is based on sequential removal of redundancy from the inputs. For the error-free inputs that simulate mass spectra with high accuracy (~10^-3 Da), the size of inputs decreases from O(n^2) to O(n). In this way, exhaustive search can be almost completely removed from the algorithms, and the number of steps to reconstruct a sequence is n^2, where n is is the cardinality of the set X, n=|X|.

Results: Now in [5] we generalized this method through the use of integral transforms. It is shown that the generalized approach can be successfully used for reconstructing the set X not only from a complete and error-free set of pairwise distances ΔX, but also for a set ΔX + f containing a large number of redundant and missing data f (noise), |f| > |ΔX|. The high efficiency of the proposed method was shown. The computational complexity of the our algorithm is O(n^2), where n is the cardinality of the input set ΔX + f, n=|ΔX + f|.

The work was carried out with the support of the Russian Foundation for Basic Research, project No.17-00-00462.

References

1. Caboche, S., M. Pupin, V. Leclère, A. (2008) Fontaine, P. Jacques, and G. Kucherov. Norine: a database of nonribosomal peptides. Nucleic Acids Res. 36: D326–D331.

2. Lemke P, Skiena SS, Smith WD, Reconstructing sets from interpoint distances, Discrete Comput Geometry Algorithms Combinatorics 25:597–631, 2003.

3. Fomin E. (2016) A Simple Approach to the Reconstruction of a Set of Points from the Multiset of n^2 Pairwise Distances in n^2 Steps for the Sequencing Problem: I. Theory. J. Comput. Biol, 23(9): 769-75;

4. Fomin E. (2016) A Simple Approach to the Reconstruction of a Set of Points from the Multiset of n^2 Pairwise Distances in n^2 Steps for the Sequencing Problem: II. Algorithm J. Comput. Biol, 23(12): 934-942.

5. Fomin E. (2018) A simple approach to the reconstruction of a set of points from the multiset of pairwise distances in n^2 steps for the sequencing problem: III. Spectra with noise. J. Comput. Biol, accepted for publication.

Bayesian modelling of gene network alterations during tree-like processes: evolution or cells differentiation

Anna A. Igolkina (Peter the Great St.Petersburg Polytechnic University)
Maria G. Samsonova (Peter the Great St.Petersburg Polytechnic University)

Tree is a typical diagrammatic representation of relationships among objects in various biological processes, for instance, phylogenetic trees or trees of cellular differentiation. The problem to predict some characteristics of objects within inner tree nodes is well studied when these characteristics are independent. Here we represent the case when these characteristics are non-independent. To be specific, we considered each object characterised by a network of interacting genes together with their expression levels and built the model to predict the configurations of the gene network within inner tree nodes: ancestral states in phylogenies or progenitor cells in the differentiation.

In our model we assumed that a tree-like process is continuous, i.e. the gene expression covariance matrix together with coefficients of gene-gene interactions change from the root state to leaves in agreement with a continuous-states time-homogeneous Markov Process, specifically the Wiener Process. We also assumed that the gene network topology should be maintained during this process so that within each inner node and outer leaf of the tree, the gene network satisfies the Structural Equation Model (SEM). We utilised the Bayesian inference to construct the scheme for MCMC parameter optimisation method.

We applied the developed model to the tree-like process of blood differentiation from hematopoietic stem cells through different progenitor states to mature states (monocytes, lymphocytes, neutrophils, etc.). We used gene expression data within leaves of the tree (microarray Human Map dataset) and optimised all parameters of both SEM model and the Wiener Process by MCMC. We modelled RAS signalling network as it involves in Leukemia development. We predicted the states of this gene network in inner nodes and, using parameters of the Wiener Process, predicted the point on the tree where the cancer cells (T-cells or B-cells) have its own branch. The knowledge of this point can potentially help in leukaemia treatment. We consider, the developed methodology can be readily applied to other cell development and also phylogenetic studies.

Discovering novel metabolisms via metagenomics

Ludmila Chistoserdova (University of Washington)

I will demonstrate the power of metagenomics in uncovering the details and the nuances of some major microbially-driven biogeochemical processes, focusing on the bacterial methane cycling as an important part of the global carbon turnover on Earth. First, by combining metagenomics with stable isotope probing, we uncover that the major species involved in the methane cycle in lake sediments (the methanotrophs) are not the ones easily cultivated in the laboratory. Second, we uncover specific satellite organisms, associated with the methanotrophs, that appear to also feed on carbon originating from methane. Additionally, we uncover the denitrification capabilities for both functional groups, suggesting that methane cycling may be linked to nitrogen cycling in oxic/unoxic interface environments. We further uncover dependence of methanotrophy on lanthanides, the rare Earth elements previously assumed to be biologically inert, and uncover a complex interplay between alternative enzymes relying on common (calcium) versus exotic (lanthanides) metals for both activity and expression. I will further highlight more recent discoveries from combining synthetic ecology approaches with meta-omics, which include further insights into communal metabolism of methane and into novel genes and enzymes, as well as into additional actors in global methane turnover that appear to function in concert with bona fide methanotrophs. Finally, I conclude that metagenomics has had a revolutionary impact on the field of methanotrophy over the past decade, and that the momentum is still going strong.

Promoters and enhancers landscape of embryonic development and hibernation in chicken

Oleg Gusev (Kazan Federal University; RIKEN)

Mathematical modeling of SNP %GC in microbial core genomes

Jon Bohlin (Norwegian Institute of Public Health)
Vegard Eldholm (Norwegian Institute of Public Health)
Ola Brynildsrud (Norwegian Institute of Public Health)
John H.O. Pettersson (Norwegian Institute of Public Health)
Kristian Alfsnes (Norwegian Institute of Public Health)

The present talk will address whether the GC content of non-recombinant substituted bases in microbial core genomes (sbGC) exhibits any association with the GC content of the corresponding core genomes (cgGC).

The GC content of the substituted bases of the strains comprising each core genome, 36 in total each representing a separate microbial species consisting of at least 10 strains, was compared with the GC content of the corresponding core genomes.

We found that sbGC within each core genome showed a non-linear association with cgGC with a bias towards higher GC content for most core genomes, assuming as a null-hypothesis that sbGC should be approximately equal to cgGC. The most GC rich core genomes (i.e. approximately %GC>60), on the other hand, exhibited slightly less GC-biased sbGC than expected. We present a simple mathematical model that estimates sbGC from cgGC. The model assumes only that the estimated sbGC is a function of cgGC proportional to fixed AT->GC (α) and GC->AT (β) mutation rates. Using non-linear regression to estimate α and β from the empirical data described above, we find that the best fitted model indicates that GC->AT mutation rates β=(1.91±0.13) p<0.001 are approximately (1.91/0.79)=2.42 times higher, on average, than AT->GC α=(-0.79±0.25) p<0.001 mutation rates. Whether the observed sbGC GC-bias for all but the most GC-rich prokaryotic species is due to selection, compensating for the GC->AT mutation bias, and/or selective neutral processes is currently debated. Residual standard error was found to be σ=0.076 indicating estimated errors of sbGC to be approximately within ±15.2% GC (95% confidence interval) for the strains of all species in the study.

Not only did our model give reasonable estimates of sbGC it also provides further support to previous observations that mutation rates in prokaryotes exhibit an universal GC->AT bias that appears to be remarkably consistent between taxa.

Building time- and cost-effective bioinformatics pipelines in the Cloud - from bcl to visual analysis with New Genome Browser

EPAM

The talk covers contemporary approaches and practical experience of building bioinformatics pipelines based on cloud technologies using data processing parallelization and flexible resources orchestration with an eye to the most efficient use of time and money.